HOW TO BUILD CONSCIOUS MACHINES

BY MICHAEL TIMOTHY BENNETT

THE AUSTRALIAN NATIONAL UNIVERSITY
DOCTORAL THESIS IN COMPUTER SCIENCE

HOW TO BUILD CONSCIOUS MACHINES

DOCTORAL THESIS ABSTRACT BY MICHAEL TIMOTHY BENNETT
THE AUSTRALIAN NATIONAL UNIVERSITY, MAY 13TH 2025

HOW TO BUILD A CONSCIOUS MACHINE? For that matter, what is consciousness? Why is my world made of qualia like the colour red or the smell of coffee? Are these fundamental building blocks of reality, or can I break them down into something more basic? If so, that suggests qualia are like an abstraction layer in a computer. A simplification. Some say simplicity is the key to intelligence. Systems which prefer simpler models need fewer resources to adapt. They "generalise" better. Yet simplicity is a property of form. Generalisation is of function. Any correlation between them depends on interpretation. In theory there could be no correlation and yet in practice, there is. Why? Software depends on the hardware that interprets it. It is made of abstraction layers, each interpreted by the layer below. I argue hardware is just another layer. As software is interpreted by hardware, hardware is by physics. There is no way to know where the stack ends. Hence I formalise an infinite stack of layers to describe all possible worlds.

EACH LAYER EMBODIES POLICIES THAT CONSTRAIN POSSIBLE WORLDS. A task is the worlds in which it is completed. Adaptive systems are abstraction layers are polycomputers, and a policy simultaneously completes more than one task. When the environment changes state, a subset of tasks are completed. This is the *cosmic ought* from which goal-directed behaviour emerges (e.g. natural selection). "Simp-maxing" systems prefer simpler policies, and "w-maxing" systems choose weaker constraints on possible worlds. I show w-maxing maximises generalisation, proving an upper bound on intelligence. I show all policies can take equally simple forms. Simp-maxing shouldn't work. To explain why it does, I invoke the Bekenstein bound. It means layers can use only finite subsets of all possible forms. Processes that favour generalisation (e.g. natural selection) will then make weak constraints take simple forms.

I PERFORM EXPERIMENTS. W-maxing generalises at $110-500\%$ the rate of simp-maxing. I formalise how systems delegate adaptation down their stacks. I show w-maxing will simp-max if control is infinitely delegated. Biological systems are more adaptable than artificial because they delegate adaptation further down. They are bioelectric polycomputers. As they scale from cells to organs, they go from simple attraction and repulsion to rich tapestries of valence. These tapestries classify objects and properties that *cause* valence, which I call causal-identities. I propose the psychophysical principle of causality arguing qualia are tapestries of valence. A vast orchestra of cells play a symphony of valence, classifying and judging. A system can learn 1ST, 2ND and higher order tapestries for itself. Phenomenal "what it is like" consciousness begins at 1ST-order-self. Conscious access for communication begins at 2ND-order-selves, making philosophical zombies impossible. This links intelligence and consciousness. So why do we have the qualia we do? A stable environment is a layer where systems can w-max without simp-maxing. Stacks can then grow tall and complex. This may shed light on the origins of life and the Fermi paradox. Diverse intelligences could be everywhere, but we cannot perceive them because they do not meet preconditions for a causal-identity afforded by our stack. I conclude by integrating all this to explain how to build a conscious machine, and a problem I call The Temporal Gap.

AUSTRALIAN NATIONAL UNIVERSITY
DOCTORAL THESIS IN COMPUTER SCIENCE

MICHAELTIMOTHYBENNETT.COM

Please send questions, hate and fan mail via the contact form or social media, all of which are linked via the above website.

This dissertation is an account of research that began March 2019. It is comprised of 13 chapters, based on 13 of my papers, written under 13 advisors and completed on the 13th of May, 2025. It was subsequently examined and accepted without any changes.

The work presented in this thesis is that of the candidate alone, except where indicated by due literature reference and acknowledgements in the text. It has not been submitted in whole or in part for any other degree at this or any other university.

Michael Timothy Bennett, September 2025

ACKNOWLEDGEMENTS

THIS WORK WAS MOSTLY FUNDED BY MY PERSONAL SAVINGS ACCOUNT. RIP. This work was partly funded by a Fundaçao para a Ciência e a Tecnologia (FCT) grant under the reference PTDC/FER-FIL/4802/2020, JST (JPMJMS2033), and an Australian Government Research Training Program (RTP) Scholarship.

I'D LIKE TO THANK the 13 people who have advised me during my various attempts at research at the ANU, both during my masters and during my PhD[1]: Sean Welsh, Anna Ciaunica, Yoshihiro Maruyama, Colin Klein, Sylvie Thiebaux, Marcus Hutter, Marcus Hegland, Michael Barnsley, Elizabeth Williams, Ehsan Nabavi, Uwe R. Zimmer, Badri Vellambi and Samuel Allen Alexander. I'd particularly like to thank Yoshi, Sean and Anna. I would not be here without you. To Yoshi, who became my primary supervisor two years into my PhD: You saw something in me and my half-complete project, which I can only imagine must have sounded mad. If you hadn't co-authored that first journal article with me, my academic career might have been over before it began. To Sean and Anna who have worked so tirelessly to get me over the finish line, without your support I might never have finished! My thesis has benefited immensely from your inputs, and your support. You have made a huge difference! On top of that I must note that Sean has done all this in his spare time as an independent researcher. Sean you have been incredibly generous with your time and feedback, and this thesis would be much less polished without your oversight. I will never forget it!

[1] These ended up as one big research project, starting with fractal compression and ending with consciousness. It has been a rather tumultuous ride due to contretemps like a global plague, university restructuring, and my stubborn refusal to heed most advice... anyway at the time I am writing this, my supervisory panel is officially listed in the university system as Sean Welsh, Anna Ciaunica, Yoshihiro Maruyama, Colin Klein and Samuel Allen Alexander.

I ALSO WANT TO THANK THE AGI SOCIETY and its members. The 2023 and 2024 AGI conferences were the highlights of my PhD. The encouragement, awards and sense of belonging I felt there quite profoundly changed my life! I'd also like to acknowledge the many others who have helped me, but there are too many. To Ricard Solé, Lenore and Manuel Blum, Karl Friston, Peter Watts, Noel Hinton, Vincent Abbott, Lucas Scott, Simon Strauss, Elija Perrier, Paul McMahon, Tim Wicks, Seth Lazar and the many others who were so generous with either encouragement or feedback: Thank you!

FINALLY TO ASHITHA GANAPATHY: You have been with me through the highs and lows of all of this. My maniacal obsessions with new topics that extended the length of this thesis by years. My moments of despair, forgetfulness and stubbornness. We even wrote our first paper together. You have listened to every part of this thesis more times than I can count. I don't know what I would have done without you. More than anyone, credit for getting me this far goes to you. From the bottom of my heart, thank you. This thesis is your achievement as well.

PRIOR WORK

To validate my progress I have continuously published throughout my PhD. Key published results include optimal learning[2] *(chapters 6 and 7)*, and my arguments regarding meaning[3] *(chapter 10)*, causality[4] *(chapters 9 and 12)* which links consciousness to intelligence, the fermi paradox[5] *(chapter 11)*, complexity[6] *(chapter 7)*, the artificial scientist[7] *(chapter 13)* and abstraction layers[8] *(chapters 4, 5, 6 and 8)*. I published The Mirror Symbol hypothesis, which informs many of the results in meaning, in an IEEE journal[9] *(chapter 10)*. My argument regarding the hard problem[10] *(chapters 12 and 13)* and my more recent survey of AGI[11] *(chapter 3)* are currently under review, but the former was accepted to and presented at both ASSC27 and MoC5. My paper on systems as a stack is of central importance to this thesis, and is forthcoming[12] *(chapters 4, 5, 8, 10 and 11)*. I co-authored and published a precursor to that paper at an IEEE cybernetics conference[13]. My paper on Computable Artificial General Intelligence[14] was important but it has been under review with IEEE Transactions on Emerging Topics in Computational Intelligence for 3 years. Fortunately, I was able to publish the key result of that paper at AGI-23 and 24 instead. I've written 21 papers in total. I expect 19 of those will have passed peer review by the time this thesis is out. My other papers are also cited but are not particularly important for this thesis. So this thesis is comprised of 13 chapters, based mostly on 13 of my papers, written under 13 advisors and completed on the 13th of May, 2025. Though many of these results were published in stand alone papers, they were all written in service of the vision I present here.

[2] Michael Timothy Bennett. The optimal choice of hypothesis is the weakest, not the shortest. In *Artificial General Intelligence*. Springer Nature, 2023a; and Michael Timothy Bennett. A formal theory of optimal learning with experimental results. *IJCAI*, 2025f

[3] Michael Timothy Bennett. Symbol emergence and the solutions to any task. In *Artificial General Intelligence*. Springer Nature, 2022a; and Michael Timothy Bennett. On the computation of meaning, language models and incomprehensible horrors. In *Artificial General Intelligence*. Springer Nature, 2023c

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

[5] Michael Timothy Bennett. Compression, the fermi paradox and artificial super-intelligence. In *Artificial General Intelligence*. Springer Nature, 2022b

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

[7] Michael Timothy Bennett and Yoshihiro Maruyama. The artificial scientist: Logicist, emergentist, and universalist approaches to artificial general intelligence. In *Artificial General Intelligence*. Springer Nature, 2022b

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

[9] Michael Timothy Bennett and Yoshihiro Maruyama. Philosophical specification of empathetic ethical artificial intelligence. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2): 292–300, 2022a

[10] Michael Timothy Bennett, Sean Welsh, and Anna Ciaunica. *Why Is Anything Conscious?* Preprint, accepted to and presented at ASSC27 and MoC5, 2024

[11] Michael Timothy Bennett. What the f*ck is artificial general intelligence? *Artificial General Intelligence*, 2025c

[12] Michael Timothy Bennett. Are biological systems more intelligent than artificial intelligence? 2025b. Forthcoming 2026

[13] Ashitha Ganapathy and Michael Timothy Bennett. Cybernetics and the future of work. In *2021 IEEE 21CW*, 2021. DOI: 10.1109/21CW48944.2021.9532561

[14] Michael Timothy Bennett. Computable Artificial General Intelligence. *Under Review*, 2022c

PEER REVIEW

THIS THESIS HAS BEEN EXAMINED and was accepted in the state it was submitted. No corrections. Here are some quotes from the anonymous examiners.

ANON 1 SAYS: *"This thesis represents a substantial and original contribution to the interdisciplinary study of artificial intelligence and cognitive science, centering on the ambitious goal of building "conscious machines." The candidate introduces several innovative concepts, including:*

- *The ontological proposition that "ought stems from change, and change is time",*
- *The epistemological proposition that "everything is a stack of abstraction layers",*
- *The technical approach "to maximize the weakness of constraints on function" (w-maxing),*
- *The Psychophysical Principle of Causality,*
- *The understanding of intelligence as "long-term adaptation facilitating short-term adaptation",*
- *The evolutionary argument that "consciousness exists because it aids adaptation" ,*
- *A tripartite model of "selves" in consciousness.*

While some individual ideas have precursors in existing literature, the synthesis of these concepts into a unified theory of life, intelligence, and consciousness constitutes a significant intellectual achievement." [...] "The thesis shines most brightly in its demonstration of bold, independent thinking. Tackling consciousness with such conceptual originality within the constraints of a Ph.D. thesis reflects exceptional intellectual courage and dedication."

ANON 2 SAYS: *"This thesis presents a rigorous and original contribution to the fields of cognitive science, artificial intelligence, and the philosophy of computation. Its central achievement is the articulation and development of weakness as a unifying principle for inductive bias, with substantial implications for AGI architectures and theories of generalisation. The thesis advances both the mathematical formalisation of weakness and its operational integration into AI frameworks, while maintaining clarity and breadth of relevance across multiple domains of inquiry. The candidate demonstrates mastery of the technical, philosophical, and interdisciplinary aspects of the research, producing a body of work that is scholarly, original, and of international significance. In my view, the thesis comfortably meets and exceeds the standards for the award of a Doctor of Philosophy." [...] "In light of its originality, rigour, and significance, I recommend that the thesis be accepted in fulfilment of the requirements for the award of the degree of Doctor of Philosophy. This work represents a foundational advance in the conceptual and mathematical understanding of inductive bias, with profound implications for the development of artificial general intelligence."*

Contents

I. FOREWORD AND CHAPTER SUMMARIES

Humans overlook subtractive solutions. We refuse to reduce. Engineers cobble together bits of code into webs so monstrous the errors cannot be found. On a more human scale governments add laws with reckless abandon, but how often do you see them repeal the old? This bias for expansion over contraction is well documented across the spectrum of human endeavour[15]. For scientific and philosophical pursuits, I suspect our tendency to overlook subtractive solutions has made many problems more difficult than they need to be. When we encounter data we cannot explain within the confines of existing theory, an additive solution would be to construct more and more convoluted theories to reconcile the old theory with the new data. However, we do not always need to reconcile the new with the old. We just need to explain what is, and sometimes that means throwing out preconceptions. For example Milton Friedman proposed simple monetary models instead of complex cyclical models, informing monetary policy that allows us to avoid repeating the great depression. I am interested in broad reaching questions which are similarly burdened by precedent. How can we build a conscious machine? Why is anything conscious? Alive? What is life? Is complexity an illusion? Are biological systems more intelligent than artificial intelligence? Why? In search of answers I have published a number of papers[16] in peer reviewed books and journals. I wrote these papers not as disconnected works but as interconnected parts of a larger vision, culminating in this thesis.

[15] Gabrielle S. Adams, Benjamin A. Converse, Andrew H. Hales, and Leidy E. Klotz. People systematically overlook subtractive changes. *Nature*, 2021

[16] I have written 21 papers total. 12 of these are published or forthcoming in peer reviewed books and journals. By August 2025, I expect that number will rise to 19 out of 21. To validate my progress I have made sure to publish my results as I have progressed through my PhD.

Overall this thesis is about how to build a conscious machine. I don't actually have a conscious machine, because that seemed like overkill for a thesis. What I do have is an explanation of what consciousness is and how it came about. There remain one or two unanswered questions. I also have some proofs and experimental results showing how to 'adapt' as efficiently as possible, which is useful for building artificial superintelligence.

There are a few other results too. I've given explanations of the origins of life, language, the Fermi paradox, causality, an alternative to Ockham's Razor, the optimal way to structure control within a company or other organisation, and instructions on how to give a computer cancer. They've mostly been published and, strange as it is, they all tie in to a coherent vision. They weren't conceived in isolation, but as parts of a whole.

What follows now is a summary of the whole thesis. The purpose of this summary is to give you a narrative overview of what I am doing and why, before I get into the weeds. As such, it uses terms like causal-identity and task without formally defining them. These terms are formally defined later in the thesis main body, but here and now they are to be read intuitively. Brevity is the only virtue to which this chapter aspires.

II-III. LITERATURE REVIEWS

Chapters II and III are literature reviews.

Chapter II surveys philosophy and neuroscience. What is a conscious entity? To build one, I must know. Philosophy, psychology and neuroscience all provide insight. However the matter is far from settled. I must take concrete positions on disputed issues within these fields before I can say how to build a conscious machine. Hence I survey some relevant concepts and disputes, combining the introductory sections of my publications on enactive and ethical AI[17], communication[18] and consciousness[19]. Topics covered include the mind body problem, functionalism, theories of consciousness, self organisation, the free energy principle, enactivism, epistemology, semiotics, structuralism, post-structuralism and theories of meaning.

Chapter III deals with AGI, which is the foundation of this thesis. It is a survey from on of my earliest publications[20], updated to reflect more recent developments[21]. I begin by discussing several definitions of intelligence and AGI. I end up framing intelligence as adaptation[22], and AGI as that which adapts generally. For the purposes of benchmarking, I define AGI is an artificial scientist. I take inspiration from Sutton's 'Bitter Lesson'[23], which is that throwing compute at a wall consistently beats human ingenuity. With sufficient resources any general approach to optimisation can eventually attain an arbitrary level of skill. Two have consistently scaled: search and approximation. I discuss strengths, weaknesses and examples of each. Hybrids of search and approximation are best. I discuss some hybrids including Hyperon[24], AERA[25] and NARS[26]. I introduce the concept of meta-approaches that can be applied to search, approximation or hybrids. One example of a meta-approach is the maximisation of scale and available resources (scale-maxing), in accord with Sutton's bitter lesson. Another is simplicity maximisation (simp-maxing) based on Ockham's Razor. I evaluate the strengths and weaknesses of these approaches. They allow us to speculate about how a superintelligence might behave. However, simplicity is a matter of interpretation. It is subjective, and so these claims are also subjective. In this thesis I propose an alternative meta-approach that *is* optimal. Overall the meta-approaches I discuss in this thesis are to maximise the simplicity of form (simp-maxing), to maximise the scale (scale-maxing) and to maximise the weakness of constraints on function (w-maxing). This latter one is my proposal.

[17] Michael Timothy Bennett and Yoshihiro Maruyama. Philosophical specification of empathetic ethical artificial intelligence. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2): 292–300, 2022a

[18] Michael Timothy Bennett. Symbol emergence and the solutions to any task. In *Artificial General Intelligence*. Springer Nature, 2022a; and Michael Timothy Bennett. On the computation of meaning, language models and incomprehensible horrors. In *Artificial General Intelligence*. Springer Nature, 2023c

[19] Michael Timothy Bennett. Emergent causality and the foundation of consciousness. In *Artificial General Intelligence*. Springer Nature, 2023b; and Michael Timothy Bennett, Sean Welsh, and Anna Ciaunica. *Why Is Anything Conscious?* Preprint, accepted to and presented at ASSC27 and MoC5, 2024

[20] Michael Timothy Bennett and Yoshihiro Maruyama. The artificial scientist: Logicist, emergentist, and universalist approaches to artificial general intelligence. In *Artificial General Intelligence*. Springer Nature, 2022b

[21] Michael Timothy Bennett. What the f*ck is artificial general intelligence? *Artificial General Intelligence*, 2025c

[22] Pei Wang. On defining artificial intelligence. *Journal of Artificial General Intelligence*, 10(2):1–37, 2019

[23] Richard Sutton. The bitter lesson. *University of Texas at Austin*, 2019

[24] Ben Goertzel et al. Opencog hyperon: A framework for agi at the human level and beyond. Technical report, OpenCog Foundation, 2023

[25] Eric Nivel et al. Autocatalytic endogenous reflective architecture. Technical report, Reykjavik University, School of Computer Science, 2013

[26] Patrick Hammer and Tony Lofthouse. 'opennars for applications': Architecture and control. In Ben Goertzel, Aleksandr I. Panov, Alexey Potapov, and Roman Yampolskiy, editors, *Artificial General Intelligence*, pages 193–204, Cham, 2020. Springer Nature

IV. WOW, EVERYTHING IS COMPUTER

This chapter explains why complexity is subjective and what can be done to formalise *objective* performance. The key result is a concept I call **computational dualism**, described in my publication of the same name[27,28]. I begin by pointing out that the very idea of a software intelligence is broken. The behaviour of software is determined by the hardware on which it runs. It interprets the environment for the software, and the software for the environment. I use the term 'computational dualism' to describe theories that treat 'minds' as disembodied entities that interact with the environment through an interpreter. I conclude that to make claims regarding the *objective* behaviour of an intelligence, we must avoid computational dualism.

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

[28] Which I am proud to say won an award at the 17th International Conference on Artificial General Intelligence, in Seattle.

I propose a solution, which I published earlier in several of my papers[29]. To avoid computational dualism, it might be tempting to think we just need to focus on the hardware. However this would repeat the same mistake. Computer systems are organised into "abstraction layers". Higher abstraction layers run in lower abstraction layers. For example Python is interpreted by a C program. I argue the abstraction layers do not *end* at hardware, and that hardware is interpreted by physical laws just as software is interpreted by hardware.

[29] Michael Timothy Bennett. Computational dualism and objective superintelligence. In *Artificial General Intelligence*. Springer Nature, 2024a; Michael Timothy Bennett. Is complexity an illusion? In *Artificial General Intelligence*. Springer Nature, 2024b; and Michael Timothy Bennett. Are biological systems more intelligent than artificial intelligence? 2025b. Forthcoming 2026

Taken to its logical conclusion, everything is a **stack** of abstraction layers[30]. Software is a state of hardware. A human is a state of organs which are states of cells. If the mind is f_3, the body or hardware[31] is f_2 and the local environment f_1, then The Stack is $f_1(f_2(f_3))$. Perhaps The Stack has a lowest layer like an 'underlying physics' f_0, meaning The Stack is $f_0(f_1(f_2(f_3)))$[32]. However we have no way of knowing. The Stack might go on forever. To make claims that hold regardless, I conclude that I need a formalism that holds in every possible world. I propose one. It is a formal definition of **environment**, which is the foundation of what I call **Stack Theory**[33]. It is what is common to all environments and 'underlying physics'. It equates **time** with **difference**, and difference with a state of the environment. This lets me formalise declarative programs in terms of difference, to integrate pancomputationalism[34]. I then argue everything must fall within the scope of what this formalism can describe. Yes, my formalism is still an abstraction. However, some claims are so weak they are true of everything.

[30] Computers are often described as a stack. For example, a video game runs on a game engine that runs on an operating system that runs on a game console. Each one is just code inside the level below, like Matryoshka dolls.

[31] Hardware is a sort of body.

[32] For example, the idea that our reality is a simulation running in another reality amounts to claiming there are yet more abstraction layers f_{-1} to f_{-n} below f_0.

[33] Stack Theory in turn provides the foundation for formalising enactivism in what I call **Pancomputational Enactivism**.

[34] Gualtiero Piccinini and Corey Maley. Computation in Physical Systems. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, Stanford, Sum. 21 edition, 2021

V. TURTLES ALL THE WAY DOWN

Chapter V is about embodiment. Each body is an abstraction layer. When I do something with my body like raise my arm, I change the possibilities for what happens next. I impose a constraint on the world. In this sense, a body speaks a formal language[35]. This embodied language is ontological, meaning a statement *is* rather than *refers* to something. Every physical thing is an abstraction layer that speaks a formal language, not just living bodies. A computer speaks a formal language of hardware states. The universe speaks a formal language of physics[36]. This idea is once again from my publications on abstraction layers[37]. I show how Stack Theory expresses an **embodied formal language** of declarative programs. Those programs are the **vocabulary** of the language. Using this, the body makes **statements** that have truth values. In an embodied formal language, something is physically 'said' by the environment. This is the language of physical laws. If I was omniscient the environment would have one state at a time, because time is difference[38]. That state would determine what is true at the present time. The grammar of the language comes from the fact that states of the environment are in this sense mutually exclusive, and some programs in a vocabulary can never be expressed together. Everything that exists is a **statement** made in an environment's embodied formal language, and which statements are true depends on the state. However from my subjective perspective within my environment, I cannot know what the physical state is. I am a statement, and I exist for as long as the environment expresses me. When a statement is made, it constrains the space of what else can happen. Each statement has an extension. Intuitively, my extension is like the 'many worlds' in which I exist.

[35] To 'express' is to physically **realise**, manifest or call into existence an object.

[36] Some may object this conflates description with verbalisation.

[37] Michael Timothy Bennett. Computational dualism and objective superintelligence. In *Artificial General Intelligence*. Springer Nature, 2024a; and Michael Timothy Bennett. Are biological systems more intelligent than artificial intelligence? 2025b. Forthcoming 2026

[38] As I have defined it for the purpose of this thesis.

Each statement implies another *higher* abstraction layer. The extension of a statement forms a vocabulary of the layer above. In this way, every statement the environment makes creates an abstraction layer. The outputs of the level below form the vocabulary of the level above. We go *up* a level of abstraction by looking at second order effects of a body we started with. An abstraction layer is like a smaller environment defined in the context of a larger environment. A 'small world' defined inside a 'big world'[39]. It has its own formal language that is equivalent to a subset of the things the bigger environment can say. Each statement the environment makes is a body, and each body has an extension and thus its own, more restricted embodied formal language in which further statements can be expressed. Layer upon nested layer of abstraction.

[39] L. J. Savage. *The Foundations of Statistics*. John Wiley & Sons, NY, USA, 1954

VI. MASTER, WHAT IS MY PURPOSE?

Chapter VI is about purpose. These results were published earlier in my papers on abstraction layers[40] and consciousness[41]. The result is a formal definition of an embodied tasks, inference and stacks. In earlier chapters, I defined bodies or hardware as embodied formal languages that express statements. I can choose any statement the body can make and call it an **input**. The possible **outputs** are the extension of the input. So a body can be seen as a computational system that maps inputs to outputs. I can single out a subset of those possible outputs and call them **correct**. I call a set of inputs and outputs a **task**. This is a way of formalising *correctness*, or what *ought* to be. According to Hume's Guillotine, I cannot derive what *ought* to be from what *is* so I need a universal, cosmic *ought* from which to derive all others. I argue this comes from time. Change is foundational. Statements are destroyed as states change. A body is a statement that lasts for as long as the environment expresses it. Subjectively, we can interpret the process of creation and destruction as statements 'moving' relative to one another. Statements that persist are those that move away from circumstances in which they are destroyed. As the environment transitions from one state to another, this eliminates that which doesn't seek to preserve its existence. This is like natural selection, but applied to every aspect of the environment. It creates an incentive I call the **cosmic ought**. What I argue here is that *everything* that the environment expresses is a statement of what *ought* to be, and the rest that which *ought not*.

Each statement implies a *narrower* abstraction layer than the one in which it was expressed, like a window or *small world* within a *big world*. As we go to higher in a stack, the *ought* gets more specific. For example, an environment could be an abstraction layer. Lifeforms would then be statements made in that abstraction layer, growing ever more specific with each additional layer of abstraction. A lifeform might be considered 'fit' if it continues to exist, so the set of all fit outputs for a fit organism could be its extension. However organisms are often unfit. Such ambiguously 'fit' organisms would be statements whose extension contains unfit as well as fit behaviour. Were it to engage in unfit behaviour it would still exist, just not in any condition to maintain homeostatic and reproductive goals. It is that distinction between 'fit' and not that a task formalises, by pointing out the set of outputs considered 'correct' and the inputs in which being correct actually matters. Hence I formalise goal directed behaviour in a stack of tasks[42].

[40] Michael Timothy Bennett. Computational dualism and objective superintelligence. In *Artificial General Intelligence*. Springer Nature, 2024a; and Michael Timothy Bennett. Are biological systems more intelligent than artificial intelligence? 2025b. Forthcoming 2026

[41] Michael Timothy Bennett. Emergent causality and the foundation of consciousness. In *Artificial General Intelligence*. Springer Nature, 2023b; and Michael Timothy Bennett, Sean Welsh, and Anna Ciaunica. *Why Is Anything Conscious?* Preprint, accepted to and presented at ASSC27 and MoC5, 2024

[42] Later in the thesis, use this to define arithmetic operations on binary strings and run experiments.

VII. WEAK

Chapter VIII is about intelligence. The key results were published in my papers on 'weak' hypotheses[43]. The result is a theory of optimal learning. I propose a meta-approach I call w-maxing, and an upper bound on intelligent behaviour based upon it. I formally prove and demonstrate experimentally that w-maxing is optimal, and simp-maxing is not[44].

If we take a Darwinian point of view then intelligence is long-term adaptation[45] that facilitates short-term adaptation[46]. Without intelligence, an organism would need to have all knowledge hard coded from birth. With intelligence, it can adapt *during* its lifetime to survive in more circumstances than without intelligence[47]. To represent this in my formalism I describe an organism by what it *does*, rather than *is*[48]. What it *does* is a task. I explain how the task an organism does can be subdivided, by choosing subsets of inputs and outputs I call child tasks. Tasks thus exist in a generational hierarchy. An organism's past is a child task of its future task. A task implies a set of policies that constrain an organism's behaviour to the task definition. An organism embodies a 'fit' policy if it is constrained to fit behaviour. The process of learning is inferring a policy from the past that ensures *future* behaviour is fit. Intuitively, a policy is like a tool. A tool can complete more than one task. A hammer can be either a weapon or a paper weight. A weaker policy is a tool that completes more tasks. The weakest policies complete the largest number of tasks. I prove that, among all policies, the **weakest** policies are the most likely to generalise, maximising efficiency of adaptation. I call this the meta-approach of **w-maxing**[49,50].

I go on to compare w-maxing and simp-maxing[51]. I prove that we can w-max without simp-maxing. I support this claim with experiments comparing the two meta-approaches. I have them attempt to learn binary multiplication and addition. The w-maxing system outperforms the simp-maxing system by $110 - 500\%$. The fewer examples one has to learn from, the greater the advantage in choosing weak policies. This all goes to show an optimal agent does not need to optimise for simpler models. I then prove that the objectively optimal agent is one that embodies the weakest policies for a task, providing an upper bound on embodied intelligence.

[43] Michael Timothy Bennett. The optimal choice of hypothesis is the weakest, not the shortest. In *Artificial General Intelligence*. Springer Nature, 2023a; Michael Timothy Bennett. A formal theory of optimal learning with experimental results. *IJCAI*, 2025f; and Michael Timothy Bennett. Computable Artificial General Intelligence. *Under Review*, 2022c

[44] Simp-maxing being simplicity maximisation based on Ockham's Razor.

[45] Inherited, hard-wired from birth.

[46] During the organism's lifetime.

[47] All else being equal.

[48] This allows us to avoid asserting particular objects or properties exist. For example, why do we consider a stool to be something that exists instead of four legs and a seat?. Everything is really just an aspect of the environment. We need make this distinction so that we can examine exactly what is needed for an object exist in chapter 11.

[49] Michael Timothy Bennett. The optimal choice of hypothesis is the weakest, not the shortest. In *Artificial General Intelligence*. Springer Nature, 2023a; and Michael Timothy Bennett. A formal theory of optimal learning with experimental results. *IJCAI*, 2025f

[50] To frame it as an epistemological razor: "Explanations should be no more specific than necessary."

[51] Recall simp-maxing is preferring simpler hypotheses in line with Ockham's Razor.

VIII. STACKISM

This chapter brings together my papers on complexity[52] and abstraction[53]. I explain why simplicity of form has anything to do with function. In theory there could be no correlation, but in practice there is[54]. My result is proofs explaining this correlation, and this explains why biological systems seem to adapt more efficiently than AI. I begin by proving that at the lowest level of abstraction, all policies are equally simple. There is no such thing as objective complexity. Then I argue bodies must use finite vocabularies, because of the Bekenstein bound[55,56]. I show that there exist abstraction layers in which simple statements are weaker. Because vocabularies are finite, an abstraction layer in which weak statements take simple forms will be able to express more weak policies than an abstraction layer where weak policies do not take simple forms. This means complexity is an illusion perpetrated by abstraction layers. To maximise adaptability given finite resources, it is *necessary*[57] for abstraction layers to express weaker constraints using simpler forms. In other words, maximising the weakness of constraints on function (w-maxing) will *cause* simplicity of form to be maximised (simp-maxing), but simp-maxing may not cause w-maxing. Natural selection prefers bodies that can express policies that are more versatile. This forces a correlation between weakness and simplicity. Since we are products of natural selection, our languages reflect this.

Next I explore how systems do this. Biological systems seem to do a better job than AI of building versatile abstraction layers. To understand why, I look at how systems vary along dimensions of abstraction, delegation and distribution. I argue systems which delegate control to lower levels of abstraction are more adaptable. I illustrate this point using examples from biological, computational, human organisational, military and economic systems. Using Stack Theory I prove that adaptability at higher levels of abstraction requires adaptability at lower levels of abstraction. I call this **The Law of the Stack**. I argue biological systems are more 'intelligent' than contemporary AI, because they delegate control to lower levels of abstraction. To put it provocatively, artificial intelligence is like an inflexible bureaucracy that only adapts top down. By adapting at lower levels of abstraction, biological systems can ensure weak constraints take simple forms at higher level[58]. I argue this is why there is a correlation between simplicity of form and the weakness of constraints on function.

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

[53] Michael Timothy Bennett. Computational dualism and objective superintelligence. In *Artificial General Intelligence*. Springer Nature, 2024a; and Michael Timothy Bennett. Are biological systems more intelligent than artificial intelligence? 2025b. Forthcoming 2026

[54] Elliott Sober. *Ockham's Razors: A User's Manual*. Cambridge Uni. Press, 2015. DOI: 10.1017/CBO9781107705937

[55] Jacob D. Bekenstein. Universal upper bound on the entropy-to-energy ratio for bounded systems. *Phys. Rev. D*, 23: 287–298, Jan 1981

[56] It says a bounded system can contain only a finite amount of information.

[57] To an extent determined by selection pressures.

[58] Conversely, in a static stack like a stable environment, weak constraints can take complex forms. This is used to explain the origins of life in chapter XI.

IX. LETS GET PSYCHOPHYSICAL

Chapter IX is about how there are objects and properties. This brings together my work on causality[59,60] and consciousness[61]. The key result is the formalisation of *causal-identites* explaining how systems learn cause and effect, the Psychophysical Principle of Causality explaining why systems learn the objects and properties they do based on w-maxing, and the formalisation of selves that will inform the later theories of consciousness and meaning.

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

[60] Which I am proud to say won an award at the 16th International Conference on Artificial General Intelligence, in Stockholm.

[61] Michael Timothy Bennett, Sean Welsh, and Anna Ciaunica. *Why Is Anything Conscious?* Preprint, accepted to and presented at ASSC27 and MoC5, 2024

Normally to describe causality I would start with a set of variables representing objects and their properties, and then experiment to figure out if changing one variable changes another. However this only works if I already have the world divided up into variables, which my formalism doesn't yet have. Fortunately, we have attraction and repulsion from physical states. Valence, which is a causal relation. Hence, I can flip the problem and learn the objects instead. We have proofs of optimal adaptability, and any system that adapts optimally must correctly identify cause and effect. I show that by w-maxing in response to attraction and repulsion from environmental states[62], a system embodies policies that classify causes of valence. I call these policies **causal-identities**. They are prelinguistic classifiers. Weaker causal-identities classify more commonly encountered causes of valence. This explains why and how a contentless environment is divided up into objects and properties. I call this The Psychophysical Principle of Causality[63]. I identify two preconditions for a system to construct a causal-identity for an object: **incentive** and **scale**. First there must be an incentive, for example the object is relevant to survival. Second, the system must be able to embody the causal-identity[64].

[62] Valence.

[63] Michael Timothy Bennett, Sean Welsh, and Anna Ciaunica. *Why Is Anything Conscious?* Preprint, accepted to and presented at ASSC27 and MoC5, 2024

[64] e.g. It must be able to see and discriminate between it, and not it.

To survive, I must be able to tell the difference between what I have caused, and what I did not. This implies the construction of causal-identities for one's self. I introduce 'orders' of causal-identity for self, and show that if the scale and incentive preconditions are met they will be constructed. 1ST-order-self classifies my interventions. A 2ND-order-self is my prediction of your prediction of my 1ST-order-self. This is needed for theory of mind, or to herd and capture prey. Finally, a 3RD-order-self permits one to predict one's own 2ND-order-selves, which is needed to predict social environments and complex narratives.

X. LANGUAGE CANCER

CHAPTER X IS ABOUT LANGUAGE AND CANCER. This integrates my first paper, in which I proposed The Mirror Symbol Hypothesis[65], with my subsequent papers on symbol emergence[66] and the formalisation of Gricean pragmatics[67]. The results are the formalisation of how meaning is communicated, of how norms are formed and how this relates to cancer, and a refutation of the Orthogonality Thesis.

I SHOW HOW 2ND-ORDER-SELVES ARE NECESSARY FOR communication as described by Grice[68]. Grice argued that if I am speaking to you, my meaning is what I intend. You have understood me if you infer my intended meaning. A 2ND-order-self lets me predict what you think I think. I can use that to predict what you *will* think I intend. Hence I can anticipate what I need to express to bias your inference toward my intended meaning. Conversely, if I want to know what you mean, I can abduct that from my prediction of your prediction of my prediction of you.

THIS EXPLAINS THE EMERGENCE OF NORMS. Organisms that can communicate can co-operate. Now that we know *how*, it is easy to see how language would evolve. I formalise protosymbols and preferences to connect causal-identities to established semiotic theory. I explain how co-operation facilitates social predation, and how sufficient predictive accuracy in repeated interactions incentivises honesty. I argue members of a species have similar preferences, and thus efficiency dictates an organism use its own preferences to predict others[69,70]. Finally, I relate normativity to cancer. In an ecosystem computation is distributed and concurrent. Different organisms act upon one another at the same time, forming collectives. When they constrain one another in service of a goal, they form a collective informational structure with an identity. Davies and Levin have argued cancer is what happens when a cell becomes isolated from the informational structure of its collective[71]. I formalise this in Stack Theory. I use explain normativity as collective identity. I show that when no policy weak enough to be shared by the members of the collective, identity is lost. As such, parts of the collective will act in a manner analogous to cancer. The Law of the Stack shows systems should be as under-specified and loosely constrained as possible while still meeting their functional requirements. Cooperation and the emergence of norms depends on delegation of control to low enough levels of abstraction. I then use this to explore AI safety and refute the strong Orthogonality Thesis.

[65] Michael Timothy Bennett and Yoshihiro Maruyama. Philosophical specification of empathetic ethical artificial intelligence. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2): 292–300, 2022a

[66] Michael Timothy Bennett. Symbol emergence and the solutions to any task. In *Artificial General Intelligence*. Springer Nature, 2022a

[67] Michael Timothy Bennett. On the computation of meaning, language models and incomprehensible horrors. In *Artificial General Intelligence*. Springer Nature, 2023c

[68] Paul Grice. Meaning. *The Philosophical Review*, 66(3):377–388, 1957; and Paul Grice. Utterer's meaning and intention. *The Philosophical Review*, 78(2):147–177, 1969

[69] Michael Timothy Bennett and Yoshihiro Maruyama. Philosophical specification of empathetic ethical artificial intelligence. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2): 292–300, 2022a

[70] This is where I propose The Mirror Symbol Hypothesis from my first publication, to explain empathy.

[71] P C W Davies and C H Lineweaver. Cancer tumors as metazoa 1.0: tapping genes of ancient ancestors. *Physical Biology*, 8(1), feb 2011; and Michael Levin. Bioelectrical approaches to cancer as a problem of the scaling of the cellular self. *Progress in Biophysics and Molecular Biology*, 2021. Cancer and Evolution

XI. WHY IS ANYTHING ALIVE?

In this chapter I ask what drives the emergence of life in an ostensibly indifferent universe? Why is it that life is complex, when complex forms are less likely to exist? In answering these questions I respond to criticisms of Pancomputational Enactivism which allege my theory does not formalise cognition in a manner which aligns with the Free Energy Principle, and accounts for a boundary[72]. I argue that a rock persists by simp-maxing alone, and that causes it to persist because simpler forms tend express weaker constraints. On the other hand, a system that self-repairs does the opposite. It w-maxes at the *expense* of simp-maxing. This is only possible in a stable environment. When the underlying stack is static, weak constraints do *not* need to take simple forms. A slime mold is more fragile than a rock in general, but in the context of earth's environment it is more adaptable in the sense that it can *do* more, to spread and multiply. Systems like this can optimise for adaptability within the constraints of higher levels of abstraction. I then relate this to The Law of Increasing Functional Information[73,74], which I translate into Stack Theory and subsequently prove. Finally, I explain the Fermi Paradox using the incentive and scale preconditions for causal-identities. Intelligent systems might be all around us, but we do not recognise them as intelligent because we cannot construct a rationale for their behaviour. They fall outside the scale and incentive preconditions afforded by the human stack. This integrates my papers on abstraction layers[75], complexity[76] and most importantly my early paper on The Fermi Paradox[77]. This serves to further illuminate how and why we divide our subjective worlds up into the objects and properties that we do. I do all of this in order to lay the foundations for chapter XII.

[72] Chris Fields, Mahault Albarracin, Karl Friston, Alex Kiefer, Maxwell JD Ramstead, and Adam Safron. How do inner screens enable imaginative experience? applying the free-energy principle directly to the study of conscious experience. *Neuroscience of Consciousness*, 2025

[73] Michael L. Wong, Carol E. Cleland, Daniel Arend, Stuart Bartlett, H. James Cleaves, Heather Demarest, Anirudh Prabhu, Jonathan I. Lunine, and Robert M. Hazen. On the roles of function and selection in evolving systems. *Proceedings of the National Academy of Sciences*, 120(43):e2310223120, 2023. DOI: 10.1073/pnas.2310223120. URL `https://www.pnas.org/doi/abs/10.1073/pnas.2310223120`

[74] This is an explanation of life proposed by others.

[75] Michael Timothy Bennett. Are biological systems more intelligent than artificial intelligence? 2025b. Forthcoming 2026

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

[77] Michael Timothy Bennett. Compression, the fermi paradox and artificial super-intelligence. In *Artificial General Intelligence*. Springer Nature, 2022b

XII. WHY IS ANYTHING CONSCIOUS?

Chapter XII addresses the hard problem. I describe *what* is consciousness, and *why* is anything conscious. I published these results earlier in my papers on consciousness[78]. First, Higher Order Thought theories argue that we are conscious of higher order meta representations of lower order conscious states like 'the smell of coffee' or 'the colour red', but they don't explain where the latter come from. To understand higher order consciousness we must explain how lower order *local* states of consciousness arise.

[78] Michael Timothy Bennett. Emergent causality and the foundation of consciousness. In *Artificial General Intelligence*. Springer Nature, 2023b; and Michael Timothy Bennett, Sean Welsh, and Anna Ciaunica. *Why Is Anything Conscious?* Preprint, accepted to and presented at ASSC27 and MoC5, 2024

I begin by examining valence. At the most basic level we have 'one-dimensional' valence. How a cell is attracted or repelled, for example. Such a system cannot learn a causal-identity for any object. However, when we have two cells we now have a richer vocabulary. We can express more. If we scale up the system, we can have many parts which are being attracted or repelled by the state at any given time: a 'tapestry of valence'. A vast orchestra of cells playing a symphony of valence. Every state of the environment would evoke such a symphony, which can be reduced to causal-identities for those aspects of the environment which cause valence. This is the point of The Psychophysical Principle of Causality. An organism learns causal-identities from valence alone. They form an abstraction layer. Causal-identities can be categorical variables like hunger and thirst, which at the higher level of abstraction have the same 'one-dimensional' valence, but have a fundamentally different *qualities* because they are different tapestries of valence at the lower level of abstraction. An organism does not have a lookup table of ordered causal-identities, and does not *choose* to use a policy to interpret inputs. It embodies causal-identities as policies and is *impelled* by valence to act accordingly. A tapestry of valence does not have the luxury of separating a representation from its estimated utility. Reward is not a label applied after the fact. Interpretation and value judgement are one and the same. I call this *integrated representation and value judgement*. This is counterintuitive from a computer science point of view, where we are used to dealing in key-value pairs for neat databases. However from an evolutionary perspective such a separation of description from valuation is implausible.

Consciousness is something an organism *does*, rather than *is*. It is being impelled by a hierarchy of causal-identities. I argue phenomenal consciousness begins with a 1ST-order-self. A 1ST-order-self accompanies every intervention an organism makes and so, having a character, it answers Nagel's famous question of 'what it is like' to be an organism. Causal-identities become qualia. A philosophical zombie has access but not phenomenal consciousness. The contents of access consciousness are those available for communication. I argue that means access consciousness requires a 2ND-order-self, because that is what is required to communicate meaning in the pragmatic sense as humans do. Communicating requires reasoning about interventions. Hence, it also requires a 1ST-order-self. A philosophical zombie that behaves exactly like a human is therefore impossible. Intelligent behaviour at a human level requires a 1ST and 2ND-order-selves. Efficiency demands the delegated computational architecture of biological self-organisation with persistent structure that supports a tapestry of valence. If intelligence is adaptability, then there is no way to achieve human-level intelligence without consciousness. Increasing intelligence is reflected in increase scale that facilitates the construction of causal-identities. I conclude the chapter by I describing the stages a conscious organism, from rocks to humans, as intelligence increases.

XIII. HOW TO BUILD CONSCIOUS MACHINES

Chapter XIII is about how to engineer conscious machines. It integrates my papers on the artificial scientist[79], and consciousness[80]. The key result is a description of the features necessary and sufficient to build a conscious machine, the proposal of an unresolved problem I call The Temporal Gap, and two options describing strategies we might take to *build* a conscious machine, or to *avoid* building a conscious machine respectively.

[79] Michael Timothy Bennett and Yoshihiro Maruyama. The artificial scientist: Logicist, emergentist, and universalist approaches to artificial general intelligence. In *Artificial General Intelligence*. Springer Nature, 2022b

[80] Michael Timothy Bennett. Emergent causality and the foundation of consciousness. In *Artificial General Intelligence*. Springer Nature, 2023b; and Michael Timothy Bennett, Sean Welsh, and Anna Ciaunica. *Why Is Anything Conscious?* Preprint, accepted to and presented at ASSC27 and MoC5, 2024

I begin by discussing existing theories of conscious machines and AGI. I argue that, because intelligence begets consciousness and consciousness requires intelligence, these are one and the same. I frame Stack Theory and subsequently Pancomputational Enactivism as bottom up frameworks. I argue they should be used to improve rather than supplant existing theories that focus on top-down implementation of conscious or intelligent systems. I subsequently enumerate the features of an artificial scientist that should in turn lead to a conscious machine.

I examine the shortcomings of conventional computing hardware in contrast to biological polycomputers, and argue there are several features we must build into our systems if we want them to be as adaptive as a human scientist, and thus conscious. I identify a problem I call The Temporal Gap, which is that it is unclear whether a conscious state is *at* a point in time[81], or can be smeared across time[82]. Machines that satisfy the former definition are conscious according to the latter definition[83]. This has profound implications not just for what sort of machines can be conscious, but for our understanding of human subjective experience. There does not appear to be any way to conclusively resolve The Temporal Gap. However I argue that if we want to build a conscious machine we should assume consciousness is *at* a point in time, and design a machine accordingly. If we wish to avoid building a conscious machine, we should assume consciousness is *smeared* across time and avoid building potentially conscious machines accordingly.

[81] Computed concurrently in one step.

[82] Computed sequentially in many steps.

[83] Meaning the latter is the weaker standard for consciousness.

Finally, I conclude the thesis by summarising the many and varied results.

II. SOME PHILOSOPHY

I MUST KNOW WHAT IT IS I WANT TO BUILD before I can really plan out how to build it. I want to build a mind, so that means I have to take concrete positions on disputed issues within philosophy of mind, psychology, cognitive science and neuroscience. The following is a survey of some relevant material from those fields. It is based on the introductory sections of my publications on enactive and ethical AI[84], communication[85] and consciousness[86]. Topics covered include the mind body problem, functionalism, the "hard problem" of consciousness, various theories of consciousness, self-organisation and the free energy principle, enactivism, epistemology, semiotics, structuralism, post-structuralism and theories of meaning. Though this is a very broad ranging survey, I try to tie these concepts together into a coherent, sequential story from beginning to end.

[84] Michael Timothy Bennett and Yoshihiro Maruyama. Philosophical specification of empathetic ethical artificial intelligence. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2): 292–300, 2022a

[85] Michael Timothy Bennett. Symbol emergence and the solutions to any task. In *Artificial General Intelligence*. Springer Nature, 2022a; and Michael Timothy Bennett. On the computation of meaning, language models and incomprehensible horrors. In *Artificial General Intelligence*. Springer Nature, 2023c

[86] Michael Timothy Bennett. Emergent causality and the foundation of consciousness. In *Artificial General Intelligence*. Springer Nature, 2023b; and Michael Timothy Bennett, Sean Welsh, and Anna Ciaunica. *Why Is Anything Conscious?* Preprint, accepted to and presented at ASSC27 and MoC5, 2024

A BRIEF HISTORY OF THE MIND BODY PROBLEM

There is public knowledge, and private knowledge[87]. When I see, smell, touch, hear or taste an object, I am said to be *directly* observing that object. However, I cannot directly observe someone else's experience. I can see *evidence* that they might be experiencing pain, for example. I could run a test and directly observe C-fibre stimulation in their brain, but that is not the same thing as directly observing their experience of pain. One's own subjective experiences are "private" knowledge. To say something is "public" information is to say it is at least possible for more than one person to observe the event. A private event is never observable by more than one person. Even if I somehow built a computer to "read" someone's mind, store the information, and then "write" that subjective experience into another's mind, how could I know the experience is truly the same? When a scientific experiment is run, it is to test whether one publicly observable event reliably follows another publicly observable event. One reason it is so difficult to study the mind is because the things for which we are testing are not publicly observable[88]. This brings us to the "mind-body" problem.

[87] Jaegwon Kim. *Philosophy of Mind*. Routledge, New York, 3rd ed. edition, 2011

[88] Seeing a scan of brain activity is not the same as actually experiencing particular brain activity. It is this experience that we cannot observe in another.

"What is a mind" is a loaded question, because it seems to suggest a mind is a publicly observable object[89]. We know that minds are had by particular things. So instead we could ask "what does it mean when we say something has a mind?". We know the things we can observe that have minds also have physical substance. Objects with physical substance are spatially extended, meaning that for each moment in time that they exist they must occupy space. No other physical object may occupy that same space at that same time.

[89] The subject of explanation is called an **explanandum**. The explanation is itself is called the **explanans**. Philosophers study the explanandum, and engineers the explanans.

However, when people speak of minds and mentality they often talk like these are not part of their physical form. For example, there is mental and physical illness. This hints at something like a mental substance. Something non-physical. However, mental and physical phenomena clearly have a causal relationship. A mind causes the body to act, and that the body causes the mind to experience what it does.

EARLY 1600s - SUBSTANCE DUALISM

The idea that there exist distinct mental and physical substances is called **substance dualism**. It was most famously argued by

Descartes, 16th century French philosopher and namesake of "Cartesian Dualism". He sought to describe the union of immaterial mind and material body. His position is unsurprising, given prevailing beliefs in the 16th century. What is surprising is that his arguments were compelling enough for us to be mentioning them four centuries later.

According to Descartes mental substance does not occupy space. Mental events are not spatially extended. Presumably this is how a mind can be *inside* a body without making it explode. Descartes thought mental substance interacts with physical substance through the pineal gland, which acts as a sort of interpreter[90]. An interpreter is like an abstraction layer in a computer. It takes one sort of thing and turns it into another. This idea that the mental and physical causally interact is called *interactionism*. In the case of Cartesian Dualism, the mental and physical directly interact through the pineal gland. He speculated fluids called "animal spirits" act upon the gland, causing it to move, which causes the conscious states of the mind. The mind then acts directly upon the gland, causing it to move and affect the animal spirits.

[90] An interpreter is something that translates one thing to another; for example French to Spanish, or from computer code to the movements of a mechanical arm.

This argument has problems. I don't need to enumerate them. Cartesian Dualism hasn't aged well, but somehow it is still here. The reason I mention it is because I will later argue that dualism seems to have been baked into computer science[91]. The idea that AI is a software "mind" running on a hardware "body" echoes Cartesian Dualism. Software is just a state of hardware, and yet many still seem to treat software as something that interacts with the world through hardware. I call this **computational dualism**[92].

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

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

MID 1600s - PREESTABLISHED HARMONY

Following Descartes, others asked why mental substance should affect the physical through only the pineal gland, and not elsewhere? Why this inconsistency? Either mental substance affects physical substance, in which case the mental is a sort of physical substance, or it affects nothing physical. In order to preserve substance dualism (likely for religious reasons), some philosophers argued the latter, holding that causal interactions between mental and physical are an illusion perpetrated by god. Leibniz argued that mental and physical processes are set in motion by god in **preestablished harmony** so that they look like they interact, but never do. Like clocks synchronized by a clockmaker. The practicalities of quantum communication

are strangely reminiscent of this idea[93]. Malebranche was another philosopher who proposed yet another alternative to interactionism. He argued the physical can affect the mental only indirectly, through the intervention of god[94]. Each time you will your body to move, god intervenes in the physical world to move your body as you wish[95]. Any time your body is affected by something in the physical realm, god affects your mind. In occasionalism, god causes all interactions between the mental and physical by intervening constantly to create the illusion of interaction, whereas in Leibniz's preestablished harmony god intervenes only once, to synchronize the mental and physical worlds so that they appear to causally interact. Either way, there is still an interpreter (god, rather than the pineal gland). I mention all this because the idea of just moving the interpreter or abstraction layer is a central theme of this thesis. It'll come back a lot.

[93] David Wallace. *The Emergent Multiverse: Quantum Theory according to the Everett Interpretation*. Oxford University Press, 05 2012. ISBN 9780199546961. DOI: 10.1093/acprof:oso/9780199546961.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199546961.001.0001

[94] A position now called occasionalism.

[95] Who's a good boy?

LATE 1600s - NEUTRAL MONISM

BOTH LEIBNIZ AND MALEBRANCHE denied there are any **direct** causal interactions between mental and physical, invoking god as a means of indirect influence. Spinoza was yet another who denied direct causal interaction, but circumvented the need for divine intervention by arguing that both mental and physical are mere aspects of a third, unobserved substance that is neither mental nor physical. In other words, reality is neither mental nor physical. This position is now called **neutral monism**. There is a secret third thing. Physical and mental are just aspects of this secret third thing. This idea will come up later when I formalise abstraction layers.

1800s - EPIPHENOMENALISM

MUCH OF THE DIFFICULTY in understanding the apparent two way causal interaction between mental and physical stems from the assumption that our perception of mental activity as causing physical activity is accurate. What if instead I just consider a one way causal relation? By this I mean that the mental has no causal effect upon the physical, but physical events cause mental events. We might *believe* we act upon the physical, but this belief is an illusion. For example, I might think I chose to get up and get a glass of water, but every aspect of my decision was determined by physical processes in my body. My mental processes are the effect, not the cause, of physical processes. This is *epiphenomenalism*. It was proposed by Thomas Huxley, who argued neural events in the *brain* are the physical events that

cause mental events. However mental events don't actually do anything. Epiphenomenalism is a means of preserving dualism, but it leaves me wondering why anything would evolve to have consciousness? From an evolutionary perspective, epiphenomenalism seems a bit pointless. The alternative is materialism, or physicalism in contemporary terms. That's the idea that mental events are just part of the physical world. It seems a lot more compelling, because it means we can come up with an evolutionary explanation for mental events.

NOW - PHYSICALISM

Physicalism comes in two flavours: reductive and non-reductive. The reductive physicalists think we will be able to reduce mental events to non-mental physical events. The non-reductive physicalists believe we will not be able to do that. They hold that certain physical processes have mental properties which are **irreducible**, meaning we can't break them down into anything simpler and so we can't reduce them to non-mental physical parts. This position is basically that "qualia" are fundamental building blocks of reality. This still requires **mental causal efficacy**, in that mental events cause other mental and physical events. Mental events must **supervene** on the physical, meaning two objects that are physically identical must be mentally identical. I am a reductive physicalist. Perhaps what might be called a Hobbesian Stackist[96]. The main point of this thesis is to explain how. I'm now going to steel-man the non-reductive physicalists for the sake of argument.

[96] At least, that is what Sean Welsh once called me.

Psychoneural identity theory is one example of reductive physicalism. It holds that the mind *is* the brain. Feelings, sensations and thoughts may be *reduced* to neural activity in the brain, or more generally to a specific physical event. Each "identity" equates publicly observable physical event with a private mental event (they are one and the same thing). One objection to psychoneural identity theory is this: if a mental event like pain *is* a particular neural event like C-fibre stimulation, then why is it that the same mental event can be caused by entirely different physical events? If I experience pain when my C-fibres are stimulated, and an animal appears to experience pain but has no C-fibres, does that mean it is not experiencing pain? It seems unlikely. Instead, it would seem a mental event like pain can be "realised" by any number of physical events. This is called **multiple realisability**.

BEHAVIOURALISM AND FUNCTIONALISM

We need to say how systems behave in order to describe mental events. Behaviouralism is the idea that one can equate mental events with outwardly observable behaviour. By observable behaviour, I mean inputs and outputs. Behaviour would be a set of input-output pairs. This is one way around multiple realisability. However it mostly depends on how we define input and output. These depend on the level of abstraction. If an input is something so vague as an intuitive human definition of "pain" then yes it would appear an octopus in pain is experiencing pain like a human. If the inputs are as specific as C nerve fibre stimulation then the octopus does not have C nerve fibres and so cannot experience pain. There are many possible processes which map I to O exactly as f does, but are not all the same thing.

This brings us to the Chinese Room. Much debated, but a good example of multiple realisability. Imagine I sit in a room. I don't speak Chinese. Through one door I am passed a note written in Chinese. I pull out my laptop, get Google to translate the note, then pass a response back out the door. Someone outside the room then starts to believe I speak Chinese. Likewise, just because something behaves as if it has a mind, does not mean it does. Behaviouralism discounts mental activity in favour of observable behaviour. It reduces meaning to inputs and outputs. The obvious problem with behaviouralism is that there is more to the story. I think, I know I think, and I can do so without giving an output. Machine functionalism[97] tries to resolve this kind of problem by adding a causal intermediary between inputs and outputs. This causal intermediary is an interpreter like a Turing machine. It maps inputs I to outputs O. Given $\langle I, O \rangle$ and a function $f : I \to O$, machine functionalism says there are many different "causal intermediaries" equivalent to f. The trick is working out *which* Turing machine is most likely to have caused the behaviour.

[97] Hilary Putnam. Psychological predicates. In William H. Capitan and Daniel Davy Merrill, editors, *Art, mind, and religion*, pages 37–48. University of Pittsburgh Press, 1967

For a reductive account of the mind to be convincing, it must deal with private first person behaviours (e.g. understanding meaning)[98], and show why these behaviours arise. The problem one faces then is arguing "is this behaviour really what I experience"? And we're back to the public-private knowledge debate. To get around the public-private distinction, I argue we have to step *outside* the universe and look in. The only way to do that is to establish axioms that hold in every universe. That is the approach I will take in chapter 5. For now, I'll delve further into background.

[98] Pei Wang. A constructive explanation of consciousness. *Journal of Artificial Intelligence and Consciousness*, 07(02):257–275, 2020; Piotr Boltuc. The engineering thesis in machine consciousness. *Techné: Research in Philosophy and Technology*, 2012; and Manuel Blum and Lenore Blum. A theoretical computer science perspective on consciousness. *J. Artif. Intell. Conscious.*, 8:1–42, 2020

CONTEMPORARY EXPLANANDUM AND HARD PROBLEMS

CONTEMPORARY THEORIES FRAME CONSCIOUSNESS[99] as having two aspects: functional and phenomenal[100]. Functional means the behaviour of consciousness, however it might be realised. Anything which might be explained by natural selection. Some equate this with "access" consciousness[101], which is the contents one can consciously "access" for reasoning and report[102]. I will point out some inconsistencies in how access consciousness is typically understood. I'll argue access consciousness is merely *part* of functional consciousness.

THE OTHER ASPECT of consciousness is phenomenal[103]. This is the subjective experience of having "global" and "local" states of consciousness. A global state, for example, is being awake or asleep. Local states or "qualia" are the specific experiences of how coffee smells in the morning, or how wet grass feels underfoot. These are hard to define in rigorous terms. By function, I mean anything that serves reproductive and homeostatic[104] goals. That serves any goal really, but later I will make the argument that all goals stem from persistence and survival. So for now just take it as that. This information processing results in behaviour natural selection deems to be fit. Some of this information we are consciously aware of, as I am aware of the words I am writing on the page at this very moment. However most of the information processing in our bodies goes on "in the dark". We are unaware of it, as I am unaware of whether my muscles have decided to atrophy because I have spent too long sitting in this chair writing. Why doesn't all the information processing go on in the dark? Why do I have conscious access to some information, and not other information? Why is there phenomenal consciousness if we can just do everything in the dark?

SOME SPECULATE[105] we might take phenomenally conscious being like a human, and make a "zombie" of it. The zombie is a clone that has all the function of the original, but not phenomenal consciousness. From the outside it looks and acts the same, but inside it is dead. If zombies are possible, then that means phenomenal consciousness has no function. If zombies can exist then there is no evolutionary explanation for phenomenal consciousness. Supposing this is true, some have asked why is there something it is like[106] to be me, instead of nothing? Why do I subjectively experience some events, when it seems possible[107] for information processing to occur without any subject experiencing it? So to reiterate in the simplest possible terms, the functional aspect of consciousness is everything

[99] Recall the subject of explanation is called an **explanandum**. The explanation is itself is called the **explanans**. We are here trying to describe the explanandum.

[100] Anil Seth and Tim Bayne. Theories of consciousness. *Nature Reviews Neuroscience*, 2022; and Georg Northoff. *Unlocking The Brain, Vol. II: Consciousness*, volume 2. Oxford University Press, USA, 2014

[101] Ned Block. On a confusion about a function of consciousness. *Brain and Behavioral Sciences*, 1995

[102] Report just means you can consciously set out to communicate it to other people.

[103] David Chalmers. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 1995; Ned Block. On a confusion about a function of consciousness. *Brain and Behavioral Sciences*, 1995; Thomas Nagel. What is it like to be a bat? *Philosophical Review*, 1974; Shaun Gallagher and Dan Zahavi. *The Phenomenological Mind*. Routledge, New York, NY, 2021; and Thomas Fuchs. *Ecology of the Brain: The phenomenology and biology of the embodied mind*. Oxford University Press, 2017

[104] Homeostasis basically just means "staying alive". I remain alive because I have "static" internal state; physical processes that keep me from being dead.

[105] Ned Block. On a confusion about a function of consciousness. *Brain and Behavioral Sciences*, 1995

[106] Thomas Nagel. What is it like to be a bat? *Philosophical Review*, 1974

[107] David Chalmers. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 1995; and Ned Block. On a confusion about a function of consciousness. *Brain and Behavioral Sciences*, 1995

we can explain as a consequence of evolutionary processes, and the phenomenal part is the rest[108]. The question is whether there is such a thing as phenomenal consciousness distinct from function. My plan is to explain phenomenal consciousness as something functional, which will kill the distinction between the two.

THIS IS THE SO CALLED HARD PROBLEM OF CONSCIOUSNESS. Some interpret it as demanding a reductive explanation for local states. However, phenomenal consciousness is arguably an easy problem. Sensory processing may explain the character of qualia, and the "subject" of subjective experience may be explained in causal terms. A representation of the self lets an organism identify the effect of its actions[109], and is necessary for accurate inference in many circumstances[110]. In other words natural selection demands there exist a self to be subject to sensations. What some have called phenomenal consciousness is really the function of consciousness in the first person[111], What has been called "functional" is the very same thing from the third person perspective[112]. However, that doesn't explain why the same behaviour couldn't come about[113] without consciousness. Some have separated phenomenal consciousness into first person functional and "hard" consciousness[114]. In that sense, hard consciousness is whatever remains unexplained by function. For the sake of this thesis and the associated papers, I interpret the hard problem as demanding an explanation of why a world in which a zombie is possible is inconceivable[115]. I'll address the hard problem by describing how consciousness[116] follows from evolutionary processes[117], which follow from the very fact of existence. I describe a formalism that applies to every conceivable environment, and show that a zombie is impossible according to that formalism.

[108] Piotr Boltuc. The engineering thesis in machine consciousness. *Techné: Research in Philosophy and Technology*, 2012

[109] Bjorn Merker. The liabilities of mobility: A selection pressure for the transition to consciousness in animal evolution. *Consciousness and Cognition*, 2005. Neurobiology of Animal Consciousness; Bjorn Merker. Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences*, 2007; and Andrew B. Barron and Colin Klein. What insects can tell us about the origins of consciousness. *Proceedings of the National Academy of Sciences*, 2016

[110] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., New York, 1st edition, 2018

[111] Stan Franklin, Bernard J Baars, Uma Ramamurthy, Gilbert Harman, Antonio Chella, Michael Wheeler, Terrell Ward Bynum, and John Barker. Apa newsletters, 2008; and Piotr Boltuc. The engineering thesis in machine consciousness. *Techné: Research in Philosophy and Technology*, 2012

[112] Pei Wang. *A Constructive Explanation of Consciousness and its Implementation*. World Scientific, 2023

[113] Realised just means "made real" or "produced" or "created".

[114] Piotr Boltuc. The engineering thesis in machine consciousness. *Techné: Research in Philosophy and Technology*, 2012; and Piotr Bołtuć. Consciousness for agi. *Procedia Computer Science*, 2020. BICA 2019

[115] Michael Timothy Bennett, Sean Welsh, and Anna Ciaunica. *Why Is Anything Conscious?* Preprint, accepted to and presented at ASSC27 and MoC5, 2024

[116] At least those parts I consider important; why there is "something it is like", the construction of selves, access consciousness and meaning.

[117] Which I already published in one of those aforementioned papers.

LEVELS OF CONSCIOUSNESS

Beyond the phenomenal and functional aspects, there are levels. Morin proposed four levels[118], which I describe below. I will later make it 6 levels:

1. **Unconsciousness:** the absence of consciousness, including sensorimotor information processing.
2. **Consciousness:** a minimal level of consciousness in which one has subjective experience of local states. Phenomenal and access consciousness both begin here.
3. **Self Awareness:** this is where there is a distinction between public and private knowledge. One now has an inner monologue, and a concept of self. Importantly, this is where self knowledge becomes possible. It is also, according to Morin, where symbolic representations come into the picture. I interpret this as akin to the "meta-representations" in higher order theories, and I will later show that access consciousness must be equated with self awareness, because it is not possible without it[119].
4. **Meta Self Awareness:** the logical conclusion of self awareness if we simply "scale up" the reflective aspect, so that one's reflection contains a reflection. It is where one becomes aware that one is aware.

[118] Alain Morin. Levels of consciousness and self-awareness: A comparison and integration of various neurocognitive views. *Consciousness and Cognition*, 2006

[119] I will argue that if access conscious contents are those available for report, then they are available for report in the sense of human exchanges of meaningful *intent*. I will show that the exchange of communicative intent requires reflectivity, and so access consciousness cannot exist without self awareness.

CONTEMPORARY EXPLANANS

Functional and phenomenal are broad categories that leave a lot of questions unanswered. There are several dominant theories which seek to explain how the phenomenal and functional aspects of consciousness come about.

HIGHER ORDER THOUGHT THEORIES

Higher order thought theories (HOTs)[120] seek to explain why we have conscious access to some information and not all. HOTs characterise the contents of access consciousness as higher order "meta-representations" derived from lower order mental states. One can be aware of the higher order representations, but not the lower order states. This implicitly divides consciousness into higher order abstractions and lower order senses. The higher order representations are of a world divided neatly into concepts like "chair" and

[120] David M. Rosenthal. *Consciousness and Mind*. Oxford University Press UK, New York, 2005; and Richard Brown, Hakwan Lau, and Joseph E. LeDoux. Understanding the higher-order approach to consciousness. *Trends in Cognitive Sciences*, 23(9):754–768, 2019. DOI: 10.1016/j.tics.2019.06.009

"sit". Grounded, multi-modal symbols. The lower order mental states are more primitive parts of which we cannot be aware, because our awareness is constructed from them. While the focus of HOTs is on access consciousness, they have also been used to shed light on the character of local states[121]. The theory I put forward in this thesis tacitly embraces HOTs, although its origins lie with AI rather than neuroscience[122]. Furthermore, I define conscious access in very different terms, and point out flaws in how HOTs define access.

[121] John Morrison. Perceptual confidence. *Analytic Philosophy*, 57(1):15–48, 2016. DOI: 10.1111/phib.12077; and Megan Peters. Towards characterizing the canonical computations generating phenomenal experience, 04 2021

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

GLOBAL WORKSPACE THEORIES

Like HOTs, global workspace theories (GWTs) explain why some information processing goes on "in the dark"[123]. The focus is on access, not qualia. GWTs can be understood using a stage analogy. The content of which we are conscious is whatever is happening on the stage. The events on the stage are globally broadcast to all the unconscious processes which observe the stage and make use of the globally broadcast information. One gains access to sensory information when it is broadcast to different parts of the brain, in particular the prefrontal cortex. GWTs differ from HOTs in that they hold that it is the broadcast of information, rather than the composition of information to form meta-representations, that distinguishes conscious from unconscious content. GWTs explain why one might be conscious of a particular local state at a particular time, but unlike HOTs they provide little insight into why two local states might differ in character. Because of this, GWTs are often understood as providing insight into conscious access rather than qualia[124]. They address questions like attention and working memory. GWTs like the Conscious Turing Machine seek to address the hard problem[125] by treating the phenomenal as first person functional. I take a similar stance and our theories seem to be compatible, although I take a different position on what consciousness is and how it functions.

[123] Bernard Baars. *In the Theater of Consciousness: The Workspace of the Mind*. 1997

[124] Anil Seth and Tim Bayne. Theories of consciousness. *Nature Reviews Neuroscience*, 2022

[125] Manuel Blum and Lenore Blum. A theoretical computer science perspective on consciousness. *J. Artif. Intell. Conscious.*, 8:1–42, 2020

AN ASIDE ON REENTRY

Reentry refers to the bidirectional exchange of signals between brain areas[126]. It is thought to play a role in synchronized firing of neurons, allowing information to be integrated[127], forming patterns within patterns. Higher levels of activity. Some associate consciousness with top down signalling resulting from this[128].

INTEGRATED INFORMATION THEORY

Unlike GWTs and HOTs which begin with information processing and focus foremost on access, Integrated Information Theory (IIT) begins with the phenomenal, from first principles regarding the character of qualia[129]. From those axioms necessary preconditions for consciousness are derived, and then it is claimed that satisfying these preconditions is sufficient to instantiate consciousness[130]. This is all formalised in mathematical terms. IIT speaks of a "cause-effect structure" and the "causal power of a system to influence itself". Global states of consciousness are associated with the quantity Φ, that indicates the "maximum irreducible integrated information generated by a system". If Φ is non-zero, then the system is supposed to be conscious. The process of reentry is thought to play a key role in integrating information. Local states are then shapes in a high-dimensional space implied by the aforementioned cause-effect structure. IIT is comprehensive, the downside of which is that there exist more potential points of failure. Also, it doesn't answer the question I want answered. It makes consciousness primary and physics secondary. I want the opposite. I want to know why anything is conscious, and I want that reason to be in terms of the physical world. The idea that consciousness is primary is fascinating, but I am more interested in the alternative. The theory I propose in this thesis is also from first principles, and also formalises consciousness in mathematical terms. However I will make the environment primary. From physics to phenomenology, as opposed to from phenomenology to physics. We do not arrive at the same conclusions, but there are complementary ideas.

[126] Gerald M Edelman and Joseph A Gally. Reentry: a key mechanism for integration of brain function. *Front Integr Neurosci*, 7:63, August 2013

[127] Anil K Seth, Jeffrey L McKinstry, Gerald M Edelman, and Jeffrey L Krichmar. Visual binding through reentrant connectivity and dynamic synchronization in a brain-based device. *Cereb Cortex*, 2004

[128] Victor Lamme. Towards a true neural stance on consciousness. *Trends in cognitive sciences*, 2006; and Victor Lamme and Pieter Roelfsema. The distinct modes of vision offered by feedforward and recurrent processing. *Trends in neurosciences*, 2000

[129] Giulio Tononi. An information integration theory of consciousness. *BMC Neuroscience*, 5(1):42, 2004; and Giulio Tononi, Melanie Boly, Marcello Massimini, and Christof Koch. Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7):450–461, Jul 2016. ISSN 1471-0048. doi: 10.1038/nrn.2016.44. URL `https://doi.org/10.1038/nrn.2016.44`

[130] Anil Seth and Tim Bayne. Theories of consciousness. *Nature Reviews Neuroscience*, 2022

AN ASIDE ON SELF ORGANISATION AND NATURALISM

When I say a system self-organises, I mean its parts interact to produce a coherent pattern or whole. Intuitively, think of a drone swarm with no central controller. Distributed computation. The internet. Self-organization[131] is more typically defined as the spontaneous emergence of order from interactions[132]. The notion applies to physics[133], biology[134] neuroscience[135] and of course computer science. A typical assignment for a distributed systems programming class is to write a program that interacts with copies of itself in a simulated environment, and the various copies must co-operate to achieve a goal *without* electing a central controller. For example, these programs might be nodes in a simulated network, and be tasked with delivering messages to specific addresses in the network without any prior knowledge of where nodes are. That is a great example of engineered self-organisation.

Self-organisation is important in biology because biological systems distribute and delegate control down to the level of cells and proteins. Supposing life did not begin with a centralised controller, the only possible means of organisation is self-organisation. They form a **multiscale competency architecture** where cells form organs, which form organisms which form ecosystems[136]. In other words, a self-organising system made up of self-organising systems. To be self-organising, a system must act to occupy only a subset of possible states. A system which does not seek some states over others merely exists, rather than self-organises. More intuitively, a self-organising system will break down in some states, so it must act to remain out of those states. It must "resist a natural tendency to disorder"[137]. It must **optimise**, or at least satisfice to a survival level. To do this, a self-organising system must predict future states in order to remain within the set of acceptable states. When I talk about an organism, I mean a biological self-organising system motivated to act in a manner deemed fit by natural selection. A conscious human is a self-organising system. So is a snowflake. Self-organisation is only part of the picture, but one well suited to *naturalist* explanations[138] that treat the **phenomenal** as something that must be **functional**.

FREE ENERGY

Predictive coding explains human perception as the result of predicting the causes of sensory signals. It frames cognition as optimisation, and thus self-organisation. Minimising expected prediction

[131] W. R. Ashby. Principles of the self-organizing dynamic system. *Journal of General Psychology*, 1947; and H. von Foerster. On self-organizing systems and their environments. In *Self-Organizing Systems*. Pergamon Press, 1960

[132] Scott Camazine, Nigel Franks, J Sneyd, Eric Bonabeau, Jean-Louis Deneubourg, and Guy Theraulaz. *Self-Organization in Biological Systems*. Princeton University Press, NJ, 2001; Thomas D. Seeley. When is self-organization used in biological systems? *The Biological Bulletin*, 2002; and Fernando Rosas, Pedro A.M. Mediano, Martín Ugarte, and Henrik J. Jensen. An information-theoretic approach to self-organisation: Emergence of complex interdependencies in coupled dynamical systems. *Entropy*, 2018

[133] Hermann Haken. *Advanced Synergetics: Instability Hierarchies of Self-Organizing Systems and Devices*. Springer-Verlag, Berlin, 1983

[134] Scott Camazine. Patterns in nature. *Natural history*, 2003; and Martha Ann Bell and Kirby Deater-Deckard. Biological systems and the development of self-regulation: Integrating behavior, genetics, and psychophysiology. *Journal of developmental and behavioral pediatrics*, 2007

[135] Scott Kelso. *Dynamic Patterns: The Self-Organization of Brain and Behavior*. MIT Press, Boston, 1997; Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010; and Emmanuelle Tognoli and J A Scott Kelso. Enlarging the scope: grasping brain complexity. *Front Syst Neurosci*, 2014

[136] Chris Fields and Michael Levin. Scale-free biology: Integrating evolutionary and developmental thinking. *BioEssays*, 42, 06 2020; and Patrick McMillen and Michael Levin. Collective intelligence: A unifying concept for integrating biology across scales and substrates. *Communications Biology*, 2024

[137] Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010

[138] Naturalist meaning as a consequence of natural selection.

error cost, or equivalently maximising expected utility or reward. **Active inference** builds upon this idea to frame cognition not only as optimising one's internal state or "model" to correctly predict the surrounding environment, but the surrounding environment to match one's model[139]. This allows for the possibility of experimentation, to falsify one's hypotheses. In this context, "free energy" is a bound on prediction error. By minimising free energy, one can minimise prediction error, and so active inference seeks to minimise free energy.

[139] Friston K., FitzGerald T., Rigoli F., Schwartenbeck P., O. Doherty J., and Pezzulo G. Active inference and learning. *Neurosci Biobehav Rev.*, pages 862–879, 2016

This is called the free energy principle. It is formalised as variational Bayesian inference[140], an approach borrowed from machine learning. It is claimed that a system which minimises free energy is optimal, making the most accurate predictions possible. Overall you can think of these ideas as a reformulation of control systems for the purpose of understanding life. To explain consciousness as a consequence of free energy minimisation is to explain it as an adaptation. A functional adaptation. Solms[141] presented such an explanation, in which one's self is defined by a Markov blanket, in which one's internal state is conditionally independent of the outside world. Qualia are predictions. Inward facing "interoceptive" predictions about one's body are how one feels. Outward facing "exteroceptive" predictions are the phenomenal characteristics of the surrounding world. One has conscious access to the most probable predictions.

[140] Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010; and Karl Friston. Life as we know it. *Journal of The Royal Society Interface*, 10(86):20130475, 2013. DOI: 10.1098/rsif.2013.0475. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2013.0475

[141] Mark Solms. *The Hidden Spring*. Profile Books, London, 2021

REAFFERENCE

One remaining noteworthy account of consciousness is that of Merker[142]. Merker sought to explain subjective experience through the emergence of a subject. The ability to discern the consequences of actions logically necessitates the existence of an integrated and egocentric representation of the world from the subject's perspective. I can't know that "I" caused something, unless I have a representation of "I"[143]. For example, if a fly sits upon my shoulder and I move, this may indicate a threat to the fly. Natural selection demands the fly be able to discern the difference between the world moving because I moved, and the world moving because *it* moved. This ability is called "reafferance"[144]. In vertebrates this capacity is supported by integrated structures in the mid-brain and in insects, the central cortex[145]. Proponents of reafference as an explanation of consciousness argue it is where the minimal requirements for something we might call "subjective experience" are found. The theory I present largely

[142] Bjorn Merker. The liabilities of mobility: A selection pressure for the transition to consciousness in animal evolution. *Consciousness and Cognition*, 2005. Neurobiology of Animal Consciousness; and Bjorn Merker. Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences*, 2007

[143] A creature without a self is just a reflection of the world around it. This has some interesting implications.

[144] Erich von Holst and Horst Mittelstaedt. Das reafferenzprinzip. *Naturwissenschaften*, 37(20):464–476, Jan 1950. ISSN 1432-1904. DOI: 10.1007/BF00622503. URL https://doi.org/10.1007/BF00622503

[145] Andrew B. Barron and Colin Klein. What insects can tell us about the origins of consciousness. *Proceedings of the National Academy of Sciences*, 2016

agrees, though for quite different reasons[146]. That we arrived at the same conclusion from two different points of origin lends it credence. However my explanation of the emergence of causality also accounts for why organisms divide the world up into the particular objects we do. In other words, it links causality to relevance and symbol grounding.

[146] I was unaware of reafference at the time I initially published my findings. My theory found its origins in artificial general intelligence and Pearlean causality, rather than a biologically inclined empirical perspective.

LIQUID AND SOLID BRAINS

It should be noted here that reafference requires a degree of centralisation. It serves to integrate and unify information for navigation and other purposes. Brains, like those in humans, are solid. The neurons remain in place, and support a bioelectric network. Information is passed *synchronously* through this network. Timing and direct *access* to information is important. All of this, in service of the ability to predict and adapt.

However, brains are not the only thing that predict and adapt. Ant colonies can solve shortest path problems. Each ant has a brain, sure, but the ant colony doesn't and it seems far more intelligent than any individual ant. Ricard Solé proposed two classes of brain to understand this[147]. First are solid brains with persistent structure. Second are *liquid* brains without any persistent structure or network and which does not *require* centralisation. Information in a liquid brain is always A liquid brain is *asynchronous*, spread across time and space, and cannot support something like a bioelectric network. This distinction will become important in the final chapters of this thesis. For now, just note that a human population is a liquid brain, and human has a solid brain.

[147] Ricard Solé, Melanie Moses, and Stephanie Forrest. Liquid brains, solid brains. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1774):20190040, 2019. DOI: 10.1098/rstb.2019.0040. URL `https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2019.0040`; Ricard Solé and Luís F Seoane. Evolution of brains and computers: The roads not taken. *Entropy*, 24(5):665, 2022; and Ricard Solé et al. Fundamental constraints to the logic of living systems. *Interface Focus*, 2024

RELEVANCE AND ENACTIVISM

Relevance realisation is the formation of a cognitive language in which inference can take place. For example, active inference describes how an organism models the world using mathematical tools like variational Bayes. In predictive coding a self-organising system assigns higher weight to more relevant aspects of the world, treating relevant aspects of the world as more precise[148]. However, where do these aspects come from? How and why is the world divided up into particular objects? Before one can model the world and predict, one must have a language for doing so. I don't mean a spoken language like human speech. I mean the circuitry of cognition. A vocabulary of primitive structures of which more abstract machinery can be constructed. That vocabulary determines which problems are hard, and which are easy. So before one can engage in active inference or predictive coding, one must first tackle the problem of **relevance realisation**, turning semantics into syntax. The organism learns a world or language relevant to its motivations[149].

[148] Brett P. Andersen, Mark Miller, and John Vervaeke. Predictive processing and relevance realization: exploring convergent solutions to the frame problem. *Phenomenology and the Cognitive Sciences*, 2022

[149] John Vervaeke, Timothy Lillicrap, and Blake Richards. Relevance realization and the emerging framework in cognitive science. *J. Log. Comput.*, 2012; John Vervaeke and Leonardo Ferraro. *Relevance, Meaning and the Cognitive Science of Wisdom*. Springer Netherlands, Dordrecht, 2013a; John Vervaeke and Leonardo Ferraro. Relevance realization and the neurodynamics and neuroconnectivity of general intelligence. In Inman Harvey, Ann Cavoukian, George Tomko, Don Borrett, Hon Kwan, and Dimitrios Hatzinakos, editors, *SmartData*, NY, 2013b. Springer Nature; and Johannes Jaeger, Anna Riedl, Alex Djedovic, John Vervaeke, and Denis Walsh. Naturalizing relevance realization: Why agency and cognition are fundamentally not computational. *Frontiers in Psychology*, 15, 2024

Relevance realisation requires the organism be **embodied**. Yet where does the body begin and end? There is now a great deal of evidence to suggest that mental processes extend beyond the brain, into the immune system[150] and even the environment an organism inhabits[151]. Intuitively, my language of cognition is not constrained to my body. I can use a pencil to write reminders on a piece of paper, and **extend** my memory into the surrounding environment. I am **embedded** in a particular environment through which my cognition is extended, and if you take me out of that environment my cognitive capabilities change. Finally, different people may interact to **enact** cognition, co-creating this text in co-operation with the environment, which I affect and am affected by in turn. Such distributed processing takes place not just between people, but within them. What we call human intelligence is the collective or swarm intelligence of cells[152]. This blurs the line between organism and environment, but it means we *can* dispense with the idea of an interpreter[153]. This is called **enactive** cognition. Intuitively, a human simplifies the world into abstract objects like "chair" and "pen". We don't think about details, just whole objects. We reduce the big world of all details to a small world of things which impact our survival[154]. Such concepts have emerged from the interaction between humans and our environment. They are "co-created"[155].

[150] Anna Ciaunica, Evgeniya V. Shmeleva, and Michael Levin. The brain is not mental! coupling neuronal and immune cellular processing in human organisms. *Frontiers in Integrative Neuroscience*, 2023

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

[152] Patrick McMillen and Michael Levin. Collective intelligence: A unifying concept for integrating biology across scales and substrates. *Communications Biology*, 2024

[153] Note that though I formalise enactive cognition, I do so by formalising the formation of an interpreter rather than presupposing it. This is useful to combine enactivism with computationalism.

[154] L. J. Savage. *The Foundations of Statistics*. John Wiley & Sons, NY, USA, 1954

[155] Francisco Varela, Evan Thompson, Eleanor Rosch, and Jon Kabat-Zinn. *The Embodied Mind: Cognitive Science and Human Experience*. 2016; and Giovanni Rolla and Nara Figueiredo. Bringing forth a world, literally. *Phenomenology and the Cognitive Sciences*, 2021

PANCOMPUTATIONALISM

Computationalism, computational cognitivism or computational theory is the idea that mental processes are computational processes. From this point of view, artificial intelligence is the engineering branch of philosophy of mind. It is an attempt to formalise the systems that support mental processes and thus recreate them. This is hard to reconcile with enactivism because it presupposes some form of interpreter between the organism and its environment. It makes a firm distinction between the organism and its environment, where enactivism blurs the line between the two. In contrast, pancomputationalism is the idea that everything is computation, not just mental processes. Pancomputationalism is trivially true given a weak notion of computation[156]. More importantly, it does not require we make any distinction between the organism and its environment, so it leaves room to formalise enactivism.

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

Over the course of this thesis I will formalise enactivism in terms that are compatible with functionalist, computational ideas regarding the mind. To do so I will formalise the stack in which boundaries or interpreters are formed, rather than presupposing them. Unfortunately, notions like enactivism and relevance realisation and often considered to be at odds with computation[157]. Of course, that depends on what we consider computation to be. Some might consider computation to be just that which occurs in a human made computational system like an Apple Silicon M4 processor that uses ARM system architecture. Others might consider it to be more abstract and general notion of Turing computation, in the sense of any machine which mimics the operations of a Turing Machine. Piccinini divides computation up into *abstract* and *concrete* sorts[158]. Abstract is whatever we interpret it to be, much like a mathematical symbol. Concrete is that which is physically manifest in the environment. It is this latter variety I'll formalise, in order to describe the possible worlds that might exist.

[157] Johannes Jaeger, Anna Riedl, Alex Djedovic, John Vervaeke, and Denis Walsh. Naturalizing relevance realization: Why agency and cognition are fundamentally not computational. *Frontiers in Psychology*, 15, 2024

[158] Gualtiero Piccinini and Corey Maley. Computation in Physical Systems. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, Stanford, Sum. 21 edition, 2021

EPISTEMOLOGY

To argue such a position is justified we must also consider how it is one might come to know anything. At the beginning of this chapter I spoke of the difference between explanans (explanation) and explanandum (the thing to be explained). Given an explanandum, there may be many equally plausible explanations. This is

particularly relevant when we are considering explanations of private knowledge, that cannot be easily verified by experiment. There are so many theories of consciousness for exactly this reason. Hence, when we consider theories of consciousness we need a means of evaluating explanations, to decide which is most plausible.

OCKHAM'S RAZOR

Something more complex is more difficult to understand or predict. Ockham's Razor amounts to the idea that simpler explanations are more likely to hold true[159], or are more likely to **generalise**. Yet simplicity is a measure of form, not function. As a subjective measure of how difficult something is to understand, complexity makes perfect sense. As a measure of something objective, namely how likely something is to hold true in future interactions with an objective reality (an environment of which one's self is part), it makes little sense. Nevertheless, empirically it is the case that simpler explanations usually hold up better under scrutiny. One could perhaps interpret this as suggesting the environment is the product of one's perception, or that there is something else going on. The important thing is that subjective perception of simplicity does correlate with empirical veracity. As part of this thesis, I explain why this is the case. I show this correlation is due to causal confounding[160]. In the context of the mind body problem, Smart used Ockham's Razor to argue in favour of the mind-brain identity theory. He argued it is implausible that consciousness is non-physical while all aspects of human sensation are physical[161]. Why should we think neural activity *cause* mental activity, when the simpler explanation is that neural activity *is* mental activity?

[159] Elliott Sober. *Ockham's Razors: A User's Manual*. Cambridge Uni. Press, 2015. DOI: 10.1017/CBO9781107705937

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

[161] JJC Smart. Sensations and brain processes. *Philosophical Review*, 68(April): 141–56, 1959. DOI: 10.2307/2182164

PRINCIPLE OF INFERENCE TO THE BEST EXPLANATION

Not all hypotheses are equally "good". If one explains why it rains, and another explains why it rains *and* why the sun rises, then the latter is a "better" explanation. It explains something that otherwise would not be explained. The **Principle of Inference to the Best Explanation** is merely that one should prefer "better" hypotheses[162]. Of course, such a principle is not without its critics[163]. One obvious problem is that we might construct an infinite number of hypotheses which are equally "good" according to the criterion given above, including some which are implausibly convoluted and specific. Yet the principle is still worth mentioning as, like Ockham's Razor, it can help identify useful explanations.

[162] Gilbert H. Harman. The inference to the best explanation. *The Philosophical Review*, 74(1):88–95, 1965. ISSN 00318108, 15581470. URL http://www.jstor.org/stable/2183532

[163] Bas C. van Fraassen. *Laws and Symmetry*. Oxford University Press, 1989

STRUCTURALIST BRAINS IN VATS

Structuralism is the idea that words and ideas are not intelligible as isolated items, but become so through their interrelations. Those interrelations are the "structure" that structuralism refers to. For example, semiotics is the study of symbols. Saussure's semiotics defines a symbol as a sign, for example a sound or visual pattern like the word "cat", and a thing which is signified, called a referent. For the word "cat" the referent would typically be a cat, but of course that can change depending on context. According to Saussure, signs gain their meaning from their interrelations with other signs. Put another way, meaning is the *difference* between signs. Structuralism became extremely influential over the course of the 20th century, until the rise of a counter-movement called post-structuralism. For our purposes a notable post-structuralist was Derrida[164], who argued that any structural description that seeks to fully encapsulate the semantics, truth or unmediated pure experience of something will be deferred or incomplete. He coined the term "différance" to describe this combination of difference and deferral. Structuralism and post-structuralism are particularly relevant given the recent success of language models. A language model like GPT-4 is optimised to learn the structure of language through their signs alone. The results have been impressive, lending credence to structuralism. However, just because a language model writes like a human does not mean it has understood the aforementioned semantics, truth or unmediated pure experience a human might have. I mention post-structuralism because my discussions with post-structuralists have proven useful. To answer my questions I'll take a primarily structuralist approach, but one that seeks to acknowledge and formalise Derrida's post-structural critique. If we want to know if a machine is truly intelligent, and has conscious experience like a human, then we must first answer what those things are for a human. We cannot begin at the level of any one concept. We must formalise the space of everything conceivable.

[164] Jacques Derrida. Writing and difference. *U of Chicago P*, 1978

Over the course of this thesis I'll discuss computational dualism, a criticism I published concerning software 'intelligence'[165]. A brain in a vat can know only what it is fed by its senses. It has no idea what is objectively true. Likewise, a computer program can know only what it is fed through hardware. The "meaning" of code is entirely determined by the hardware on which we run it. If a computer program is a model of the world, then its accuracy depends on the interpretation of it. In other words if we're to know what a

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

computer program knows, then the conventional distinction between software and hardware is going to have to be abandoned. We're going to have to avoid computational dualism, and to do that we need to formalise the space of all conceivable environments and see what holds in all of them. I'll argue every conceivable environment has at least one state. The power set of states is the set of every possible *difference*. There is no difference within a state, only between states, and differences are the programs of which aspects of the environment are formed. Intuitively, it doesn't matter what states are because we assume nothing about them. After all a human can only interact with aspects of his environment. If he were to try and pinpoint what an aspect is made of, the answer would be *deferred* to other aspects. This is analogous to the treatment of foundational concepts in structuralism and post-structuralism. Any answer that sought to fully encapsulate the semantics, truth or unmediated pure experience of a state would, in the language of post structuralism, be always already delayed, deferred or incomplete. This does not render such attempts vacuous, but does speak to their inherent contingency and conditionality[166]. From there I take a firmly naturalist approach to explaining what might or must exist in every conceivable environment, working from first principles with pragmatic assumptions of natural selection and self-organisation.

[166] Thanks Elija Perrier for help with the phrasing here.

PRAGMATICS AND THE ORIGINS OF OUGHT

The philosopher Hume famously showed a statement describing what *ought* to be cannot be derived from a statement of what *is*. This dissociation of value from description is named "Hume's Guillotine"[167]. To take a naturalist approach to explaining everything, I need to dissolve Hume's Guillotine by showing where an original *ought* comes from, to get natural selection. Finally, I'll take a moment to describe an alternative to Saussure's structuralist semiotics. Note that Saussure's symbols were dyadic, meaning they contained two parts. A sign, and a referent. In contrast the semiotics of Peirce defines a symbol as triadic. A Peircean symbol as a sign, a referent and an *interpretant*. The interpretant is the effect of the sign upon the person who interprets it. For example, if I see the word "cat" and feel hungry, this has implications for what I'll do next. Such a pragmatic, consequence oriented account of symbols is useful for a naturalist account of meaning. I'll dispense with "is" by arguing the very fact of continued existence constitutes an *ought* from which purpose and behaviour follow. As part of that I'll formalise meaningful communication in terms of pragmatics, namely Gricean theories of mean-

[167] It is probably better known as Hume's Law, but I prefer Hume's Guillotine. Sharper. More of an edge.

ing[168]. Grice held that the meaning of an utterance[169], is whatever the speaker *intends* the listener hold in their mind as a consequence of listening. Likewise, the listener has *understood* the meaning of an utterance if they come to hold in their mind approximately what the speaker intended. As an unintended consequence of following the thread from first principles to try and explain consciousness, the formalisation I'll present just happens to align with both Peircean semiotics and Gricean pragmatics, unifying the two.

THIS CHAPTER BROUGHT TOGETHER MANY IDEAS. The rest of this thesis tells a more straightforward story, from beginning to end. The next chapter is yet another survey, but it tells a nice story about a bitter lesson.

[168] Paul Grice. Meaning. *The Philosophical Review*, 66(3):377–388, 1957; and Paul Grice. Utterer's meaning and intention. *The Philosophical Review*, 78(2):147–177, 1969

[169] Utterance is philosophical jargon for "something said aloud".

*III. WHAT THE F*CK IS AGI?*

Contemporary AI systems are narrow[170,171], brittle, and proficient only within stable environments. Artificial General Intelligence (AGI) represents the pinnacle of artificial intelligence research: a machine that learns and adapts with the ferocity of a human mind[172]. Many peg AGI to human-level performance across a broad range of tasks[173]. I myself have done this. It is a cozy, intuitive benchmark. It is also anthropocentric and so vague it's practically a Rorschach test. This definition is insufficient, but arguably necessary. Human intelligence has many aspects. Some have emphasised autonomy, agency, and a balance of exploration in search of knowledge against exploitation of that knowledge[174]. AGI is not a passive observer of the world but part of it. As Pearl puts it, a truly intelligent agent must surmount a 'ladder of causality'[175]. It must discriminate between events it has caused and events it merely observes. It must evaluate counterfactuals and imagine entirely alternative paths to the same end. Certainly these are all necessary for AGI. At a higher level, Goertzel[176] has described AGI as a system tackling complex goals in broad, unpredictable environments. However we must then decide what is "complex"? What's "unpredictable"?

[170] I cite precedent for the use of profanity in the chapter title. A respected PLoS medical journal permitted the word "shit" in a paper title. My use of censored profanity seems a little tame in comparison.

[171] Stefanie J Krauth, Jean T Coulibaly, Stefanie Knopp, Mahamadou Traoré, Eliézer K N'Goran, and Jürg Utzinger. An in-depth analysis of a piece of shit: distribution of Schistosoma mansoni and hookworm eggs in human stool. *PLoS Neglected Tropical Diseases*, 6(12): e1969, 12 2012. ISSN 1935-2727. DOI: 10.1371/journal.pntd.0001969. URL https://doi.org/10.1371/journal.pntd.0001969

[172] Michael Timothy Bennett and Yoshihiro Maruyama. The artificial scientist: Logicist, emergentist, and universalist approaches to artificial general intelligence. In *Artificial General Intelligence*. Springer Nature, 2022b; and Michael Timothy Bennett. What the f*ck is artificial general intelligence? *Artificial General Intelligence*, 2025c

[173] Stuart Russell. *Artificial Intelligence and the Problem of Control*, pages 19–24. Springer Nature, 2022

[174] Kristinn R. Thorisson. *A New Constructivist AI: From Manual Methods to Self-Constructive Systems*, pages 145–171. Atlantis Press, Paris, 2012; Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, MA, 2018; and P. Wang. *Rigid Flexibility: The Logic of Intelligence*. Applied Logic Series. Springer Nature, 2006

[175] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., New York, 1st edition, 2018

[176] Ben Goertzel. Generative ai vs. agi: The cognitive strengths and weaknesses of modern llms, 2023. URL https://arxiv.org/abs/2309.10371

HUTTER[177] SOUGHT TO ANSWER such questions with a universal problem-solving model, weighted by complexity. Legg and Hutter[178] later framed this idea as 'the ability to achieve goals across a wide range of environments'. It's crisp and formal, but incomputable and entirely subjective[179]. Chollet[180] argued AGI is that which maximises 'g-factor'. G-factor is an idea from psychology. It is how much information one requires to acquire a skill. By some accounts, this is what an IQ test is supposed to measure. However Chollet's formal measure of intelligence is not in any meaningful way different from Legg-Hutter intelligence. It still uses complexity to assess difficulty, suffering the same pitfalls of subjectivity and incomputability. Both Legg-Hutter intelligence and Chollet's measure treat goals as something that can be separated from intelligence. This implicitly endorses the orthogonality thesis. In AI safety the orthogonality thesis is that intelligence can be separated from final goals, and any goal can be pursued by an advanced intelligence[181]. As far as I can see the alternative is to treat intelligence as embodied, and since embodiment conveys a bias towards some goals over others this links intelligence to goals. I argue this refutes the orthogonality thesis later, and in a related paper[182]. For now, what is significant about this is that it sets the foundation to frame intelligence as adaptation. Pei Wang argued in favor of this definition[183]. Wang combines various definitions to arrive at 'intelligence is adaptation with insufficient resources'. I agree with this definition. I quibble about a few details, in order to formalise it.

I GIVE TWO TESTABLE DEFINITIONS. The first is a quantifiable definition of intelligence[184]. It is the ability to complete a wide range of tasks: a nod to Legg-Hutter intelligence, but most closely aligned with Wang's definition. It deals in systems as a whole. If system A can complete a superset of the tasks system B can, then A is more adaptable. This says intelligence is *contextual*, and that there is no intelligence absent a goal. This measures both sample and energy efficiency. If intelligence is adaptation then AGI should be that which adapts generally.

[177] Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer Nature, Heidelberg, 2010

[178] Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and Machines*, pages 391–444, 2007

[179] Jan Leike and Marcus Hutter. Bad universal priors and notions of optimality. *Proceedings of The 28th Conference on Learning Theory, in Proceedings of Machine Learning Research*, pages 1244–1259, 2015

[180] François Chollet. On the measure of intelligence, 2019

[181] Nick Bostrom. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22(2): 71–85, May 2012. ISSN 1572-8641. DOI: 10.1007/s11023-012-9281-3. URL `https://doi.org/10.1007/s11023-012-9281-3`; and Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford, UK, 2014. ISBN 9780199678112

[182] Michael Timothy Bennett. Lies, damned lies, and the orthogonality thesis. *Preprint*, 2025d

[183] Pei Wang. On defining artificial intelligence. *Journal of Artificial General Intelligence*, 10(2):1–37, 2019

[184] See definition 5 in the appendix.

Hence I define AGI as an artificial scientist. Others have proposed this[185], I just formalise it[186]. This is a high bar in terms of adaptability. Scientist is a job description. The test is can an AI do this job? Not just solving problems we hand it, but generating new hypotheses, experiments, and making real breakthroughs without relying on a human for direction. It must even give lectures, podcast interviews, apply for grants and flatter donors. An artificial scientist must be capable of autonomously making scientific progress, like a human. It must balance exploration and exploitation of knowledge. It must allocate resources. It must be able to achieve goals in a wider range of complex environments. It must identify cause and effect. It must construct plausible hypotheses and design experiments to test them. In short, an artificial scientist must satisfy all of the definitions above.

This isn't just about AGI for AGI's sake. It is a stepping stone to the core of this thesis: how to build a conscious machine. An AGI that discovers isn't just clever; it's got the kind of mental horsepower that might hint at awareness. I will explain consciousness as a consequence of function. The next sections will dig into the tech while calling out the gaps still holding us back.

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

[186] Michael Timothy Bennett and Yoshihiro Maruyama. The artificial scientist: Logicist, emergentist, and universalist approaches to artificial general intelligence. In *Artificial General Intelligence*. Springer Nature, 2022b

EVERYTHING IS A BITTER LESSON

HAVING DEFINED WHAT this thing is I now need to say how anyone hopes to get there. Rich Sutton[187] argues the history of AI has taught one 'bitter lesson'. In chess, early systems encoding grandmaster strategies were eclipsed by brute-force search algorithms as computational power grew[188]. In NLP, meticulously designed linguistic rules gave way to deep learning models trained on sprawling corpora, exemplified by the transformer architecture[189]. Sutton's insight was that the relentless march of compute trumps human ingenuity[190]. To solve a problem I can hand-craft clever solutions to problems, or I can apply general methods like search or approximation and just optimise for what I want[191]. If resources are not a consideration, then general methods will eventually beat any approach that relies on human-crafted knowledge or structures. AI started to be of practical use because hardware improved to the point where AI could be applied at scale, not because anything significant changed with the algorithms. The Bitter Lesson gives you The Scaling Hypothesis. The Scaling Hypothesis asserts that by amplifying the size of AI models, the volume of training data, and the computational power deployed, we'll eventually rival or surpass human capabilities. The Scaling Hypothesis has surged in prominence, fueled by the striking achievements of large-scale models across diverse domains. For example, OpenAI's GPT-3, boasting 175 billion parameters, showcased remarkable proficiency in generating human-like text, executing tasks with minimal prompting, and even hinting at basic reasoning[192]. Likewise, DeepMind's AlphaFold 2 harnessed vast computational resources and biological datasets to revolutionize protein structure prediction, solving a decades-old challenge in biology[193]. These breakthroughs demonstrate that scaling does get results, at least to an extent. Empirical support for the scaling hypothesis is bolstered by scaling laws, which reveal predictable performance gains as model size, data, and compute increase. Kaplan et al. demonstrated that in natural language processing (NLP), larger models consistently improve in performance. This hints at a systematic relationship between scale and capability[194]. Advocates argue that as models grow and ingest more diverse data, they approximate a deeper, more general understanding of the world.

[187] Richard Sutton. The bitter lesson. *University of Texas at Austin*, 2019

[188] Murray Campbell, A.Joseph Hoane, and Feng hsiung Hsu. Deep blue. *Artificial Intelligence*, 2002

[189] Ashish Vaswani et al. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, NY, 2017. Curran

[190] Note that I will prove an upper bound on embodied intelligence in this thesis.

[191] Sutton actually says 'search' and 'learning', but those terms are a bit ambiguous because a search algorithm can be used to learn. Hence to make the distinction clearer I'll call these 'search' and 'approximation'. Symbolic methods like traditional reinforcement learning fall into the search bucket. Curve fitting of any kind falls into approximation.

[192] Tom B Brown et al. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, NY, 2020

[193] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 2021

[194] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020

THERE ARE CRITICS. I COUNT MYSELF AMONG THEM. While scaling might eventually work, the word "eventually" is doing a lot of work. There are diminishing returns. Beyond a certain threshold, additional parameters yield only incremental gains, suggesting a ceiling to this approach. In language models, performance gains taper off as size increases. Marginal improvements don't justify exponential resource costs. This plateau challenges the notion that scale alone can bridge the gap to AGI. The environmental toll of training behemoth models is staggering, with carbon emissions rivalling those of small industries[195]. This is exacerbated by the fact that scaled models excel in their training domains but often falter beyond them. Large language models generate fluent text yet stumble on tasks demanding deep reasoning or contextual nuance[196]. Some suggest that neural networks are fundamentally incapable of reasoning or causal understanding[197]. I don't know about that. A full human brain integrated with a human body is quite spectacular. A chunk of human brain sitting on a counter-top tends to be rather ghoulish and unimpressive. I do know these systems are sample inefficient, meaning they need a lot of data or many 'examples' to learn from. That is a criticism I find compelling. Adaptability is about dealing with edge cases, not rote learning. A system that needs a data centre to learn tic-tac-toe isn't intelligent: It's a whale beached on silicon. Finally, scaling assumes you know what you want and can measure it. That is quite the assumption. We can mimic human behaviour, but is that really what we want? To replicate ourselves?

[195] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019. Association for Computational Linguistics

[196] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097

[197] Gary Marcus. Deep learning: A critical appraisal, 2018

THE SCALING HYPOTHESIS IS POTENT. Yet it is not a silver bullet. Empirical success must be weighed against diminishing returns, theoretical gaps, and ethical trade-offs. To understand how we can do this, we must examine what exactly it is that we're scaling. The typical ML and AI concepts like supervised learning, reinforcement learning, inference, reasoning, planning and so on aren't useful because an artificial scientist must able to do all of it. Instead I will take my cue from Sutton's bitter lesson and speak only of the means by which these things are achieved. These means are the search and approximation. This is not the only way to think about this, so I then discuss hybrids. Hybrids are those systems which do not fall neatly into the buckets of search and approximation. Finally, I discuss meta-approaches, which are frameworks through which search, approximation and hybrid systems can be understood. Meta-approaches give us a quantifiable answer to 'what is intelligence' that other systems can optimise for.

BASIC TOOLS

SEARCH

Search is the historical workhorse of AI[198]. I include any symbolic reasoning and planning in this bucket. In its most basic form search involves representing a problem space and solution criteria. Then every nook and cranny of the problem space is explored and tested until a solution is found. Rooted in the foundational era of computation, search-based methods embody the belief that intelligence can be distilled into systematic exploration of well-defined possibilities. This section dissects search-based AI. I discuss operational principles, its strength in structured domains, its limitations in the face of complexity, and its fit within the broader quest for AGI. It is a precision instrument. At its core, search-based AI is about exhaustive exploration. Whether it's planning a route, solving a puzzle, or proving a theorem, search involves a representation of the problem's state space (often as a graph or tree). A search algorithm then systematically traverses it, evaluating paths against a defined goal. This is the essence of algorithms like breadth-first search (BFS), which explore all nodes at the current depth before moving deeper. Depth-first search (DFS) dives deep into one path before backtracking. A more sophisticated method called A*[199] employs a 'heuristic' to guide the search toward promising areas. A heuristic is like a rangefinder, and A* searches the nodes that the heuristic says are closer to the goal first. A canonical example of all this is SatPlan[200], which transforms planning problems into Boolean satisfiability (SAT) instances, solvable via logic-based search. SatPlan and its ilk have excelled in domains like logistics scheduling and automated reasoning where the problem can be fully specified. By this I mean states, actions, and goals can be laid out clearly. Search thrives in these environments, where the solution is a matter of finding the optimal path through a labyrinth of possibilities.

[198] S. Russell and P. Norvig. *Artificial intelligence: A modern approach, global edition 4th*. Pearson, London, 2021

[199] Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968. DOI: 10.1109/TSSC.1968.300136

[200] Henry Kautz and Bart Selman. Planning as satisfiability. In *IN ECAI-92*, pages 359–363, New York, 1992. Wiley

Search has its advantages. Here are a few:

- optimality: When properly configured (e.g. with an admissible heuristic in A*), search algorithms guarantee the discovery of the optimal solution, provided one exists. This is invaluable in domains where precision is non-negotiable, such as automated theorem proving[201] or mission-critical planning in aerospace.
- interpretability: The process is transparent. Each step can be traced and understood. That makes search-based systems easier to debug, verify, and trust than their approximated counterparts.

[201] A. Newell and H. Simon. The logic theory machine–a complex information processing system. *IRE Transactions on Information Theory*, 2(3):61–79, 1956

- STRUCTURE EXPLOITATION: Search excels in problems with well-defined structures, where the state space, though potentially vast, is navigable through clever pruning and heuristic guidance. This makes it a go-to for tasks like game playing (e.g. chess engines pre-AlphaGo) and pathfinding in robotics.

These strengths have cemented search as a cornerstone of AI, particularly in environments where correctness and transparency are paramount.

HOWEVER, SEARCH ALSO HAS DRAWBACKS:

- COMBINATORIAL EXPLOSION: The primary curse of search is its scalability. For problems with large state spaces, the number of possible paths grows exponentially, a phenomenon known as the combinatorial explosion. Even with heuristics, search can become computationally intractable for all but the most carefully constrained problems. In chess the state space is approximately 10^{46} nodes. This is too large for brute-force exploration without aggressive pruning. Prior or contextual knowledge can be used constrain the search space and mitigate this problem.

- SEQUENTIAL NATURE: Search algorithms are sequential, making them ill-suited for modern parallel hardware like GPUs, which thrive on matrix operations and batch processing. This puts search at a severe disadvantage compared to approximation-based methods, which can leverage massive parallelism to accelerate learning and inference. Concurrent and distributed search algorithms exist, but have not yet matured into user friendly and scalable libraries[202].

- RIGIDITY IN PROBLEM FRAMING: Search demands a pristine problem definition. This means explicit states, transitions, and goals. Real-world problems are often riddled with uncertainty. Search falters in these environments, requiring human intervention to massage the problem into a tractable form. This reliance on human pre-processing is a far cry from the autonomous adaptability we seek in AGI. However this ceases to be a significant problem if search can be made more efficient and scalable.

In its current form, search-based AI is a perfectionist that thrives in controlled, sterile environments but wilts when faced with the chaos of reality.

[202] Christian Schulte and Mats Carlsson. Chapter 14 - finite domain constraint programming systems. In Francesca Rossi, Peter van Beek, and Toby Walsh, editors, *Handbook of Constraint Programming*, Foundations of Artificial Intelligence. Elsevier, 2006; Stefan Edelkamp and Stefan Schrödl. Chapter 9 - distributed search. In Stefan Edelkamp and Stefan Schrödl, editors, *Heuristic Search*, pages 369–427. Morgan Kaufmann, San Francisco, 2012; and Yichao Zhou and Jianyang Zeng. Massively parallel a* search on a gpu. *Proceedings of the AAAI Conference on Artificial Intelligence*, (1), 2015

Search has a few notches in its belt:

- SATPLAN: By converting planning problems into SAT instances, SatPlan has solved complex logistics and scheduling tasks with precision[203]. However, its reliance on well-defined constraints limits its applicability to more fluid, real-world scenarios.
- CHESS ENGINES (E.G. DEEP BLUE): Chess engines like Deep Blue[204] relied on search algorithms augmented with evaluation functions to defeat world champions.
- PATHFINDING ALGORITHMS: A* and its variants remain the gold standard for navigation in robotics and video games[205], efficiently plotting optimal routes in static environments. But again, their effectiveness diminishes with increased uncertainty and dimensionality.

These examples underscore search's prowess in structured domains while highlighting its limitations in more complex, adaptive settings.

[203] Henry Kautz and Bart Selman. Planning as satisfiability. In *IN ECAI-92*, pages 359–363, New York, 1992. Wiley

[204] Murray Campbell, A.Joseph Hoane, and Feng hsiung Hsu. Deep blue. *Artificial Intelligence*, 2002

[205] Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968. DOI: 10.1109/TSSC.1968.300136

APPROXIMATION

By approximation I mean curve fitting. I mean all those artificial intelligence techniques that address complex problems by approximating underlying functions, distributions, or decision surfaces, rather than relying on exhaustive computation or exact solutions. Unlike search, approximation-based approaches excel in environments with high dimensionality and noise. Computer vision and natural language processing systems depend heavily on approximation. This section briefly examines its defining characteristics, advantages and limitations. At its core, approximation-based AI optimises a model reflect patterns in data so it can be used to make predictions about other data generated by the same source. In its simplest form this would be like writing down and averaging someone's score in a game so you can predict what they will get in future. There is something that generated data (the player and games), and you train a model (by taking the average) until it reflects some aspect of the generator. I can train a model to classify data, answering 'which thing generated this data?'. I can also train a model generate new data.

Typically a parameterized model such as a neural network approximates a target function by minimizing a loss function over a training dataset. Mathematically, given an input space (X) and output space (Y), the goal is to find a function $f_\theta : X \to Y$, parameterized by θ , that closely matches the true mapping f^* , even when f^* is unknown or intractable. The error is typically quantified via a loss function $L(f_\theta(x), y)$, and optimization techniques like gradient descent adjust θ to minimize this loss over a dataset $D = \{(x_i, y_i)\}_{i=1}^N$. The ascendancy of deep learning, a subset of approximation-based AI, has been particularly notable. Deep neural networks leverage multiple layers of interconnected nodes to learn hierarchical feature representations, enabling them to tackle tasks with unprecedented accuracy. For instance, convolutional neural networks (CNNs) have redefined computer vision[206], while transformer architectures have revolutionized natural language processing[207]. Approximation-based end-to-end reinforcement learning has shown promise in game playing and robotics[208]. Approximation is ideally suited to scenarios where we can trade accuracy and reliability for scalability and practicality.

Approximation has advantages over search:

- scalability: These methods efficiently process large-scale, high-dimensional data. For example, convolutional neural networks can

[206] Alex Krizhevsky et al. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 2017; and Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016

[207] Ashish Vaswani et al. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, NY, 2017. Curran

[208] Volodymyr Mnih et al. Human-level control through deep reinforcement learning. *Nature*, 2015

classify millions of images by learning compact feature representations, bypassing the need for exhaustive hand-crafted rules.

- ROBUSTNESS TO UNCERTAINTY: By modeling data distributions probabilistically or incorporating regularization, approximation-based models can generalize from noisy or incomplete inputs. Techniques like dropout in neural networks[209] or Bayesian methods enhance this resilience, making them suitable for applications like speech recognition in variable acoustic conditions.

- FLEXIBILITY AND AUTOMATION: Search often requires a domain-specific heuristic. Approximation is cheaper, which means it can learn directly from data. This is ideal for problems where the relationship between inputs and outputs is highly non-linear or poorly understood. It can minimise the need for human-engineered features. This adaptability has fueled its adoption in fields from genomics to finance, with minimal reconfiguration.

Scalability in particular has led to widespread adoption, as you might expect given the bitter lesson. Some examples:

- CONVOLUTIONAL NEURAL NETWORKS (CNNS): CNNs exploit spatial locality and parameter sharing to achieve state-of-the-art performance in visual tasks[210].

- TRANSFORMERS: Transformers rely on self-attention mechanisms to model long-range dependencies in sequences[211]. Models like BERT[212] and GPT-3[213] have set benchmarks in natural language understanding and generation, leveraging massive datasets (e.g. GPT-3 was trained on 45TB of text) to approximate linguistic structures.

- DEEP REINFORCEMENT LEARNING: Deep Q-Networks (DQN)[214] combine neural networks with Q-learning to approximate value functions, achieving human-level performance in Atari games. Similarly, Proximal Policy Optimization[215] has advanced policy approximation in continuous control tasks.

These examples highlight the ability of approximation-based AI to address diverse challenges with tailored architectures.

[209] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL `http://jmlr.org/papers/v15/srivastava14a.html`

[210] Alex Krizhevsky et al. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 2017; and Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016

[211] Ashish Vaswani et al. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, NY, 2017. Curran

[212] Jacob Devlin et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019

[213] Tom B Brown et al. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, NY, 2020

[214] Volodymyr Mnih et al. Human-level control through deep reinforcement learning. *Nature*, 2015

[215] John Schulman et al. Proximal policy optimization algorithms, 2017

Despite recent success, approximation is not a panacea:

- unreliability: Approximation is only approximate. Stochastic. It is unreliable by design[216]. This makes it difficult to apply to problems where failure cannot be tolerated. This is why search is used for applications like maps and directions. Directions that 'approximate' a route through a river are not useful.
- interpretability: The complexity of models like deep neural networks, often with millions of parameters (e.g. GPT-3 has 175 billion), renders them opaque. This "black box" nature complicates understanding of decision rationales, a critical issue in domains requiring accountability, such as medical diagnostics or legal systems. Efforts like LIME[217] and SHAP[218] provide post-hoc explanations, but these are often approximations themselves and lack the rigor of causal insight.
- sample inefficiency: High performance hinges on access to large, labeled datasets. For instance, training ResNet-50 on ImageNet requires 1.28 million labeled images, while GPT-3's training consumed computational resources equivalent to thousands of GPU days[219]. In data-scarce domains, such as rare disease diagnosis, this dependency limits applicability and risks overfitting, where f_θ fits noise rather than signal (bias-variance trade-off). In other words, approximation is maladaptive. Techniques like transfer learning can mitigate costs, but performance still drops sharply outside the training distribution[220].
- computational cost: The training of approximation-based models incurs substantial energy and infrastructure demands. For example, Strubell et al.[221] estimate that training a single transformer model emits carbon equivalent to 626,000 miles of car travel, raising sustainability concerns.

These drawbacks underscore the trade-offs inherent in approximation, necessitating careful consideration of context and resource constraints.

[216] Elija Perrier and Michael Timothy Bennett. Position: Stop acting like language model agents are normal agents, 2025. URL https://arxiv.org/abs/2502.10420

[217] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322

[218] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, NY, 2017. Curran

[219] Tom B Brown et al. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, NY, 2020

[220] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3320–3328, Cambridge, MA, USA, 2014. MIT Press

[221] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019. Association for Computational Linguistics

HYBRIDS

Hybrids are those systems which do not fit neatly into the search or approximation buckets. Biological self-organising systems learn and adapt, but they are not clearly a case of just search or approximation. Hybrid approaches are inherently more general because I can pick choose any general approach for any occasion. I can fuse search and approximation, or something else. By combining complementary strengths, hybrid systems offer a tantalizing path toward AGI, promising robustness where monolithic approaches falter[222]. Perhaps no single AI paradigm holds the key to AGI. Search excels at precision. Approximation thrives on raw data and uncertainty. Hybrid systems bridge these gaps, blending precision with flexibility, logic with learning. The goal? Synergy. Emulate humanity's versatility, tackling everything from sensory processing to scientific discovery.

[222] Michael Timothy Bennett and Yoshihiro Maruyama. The artificial scientist: Logicist, emergentist, and universalist approaches to artificial general intelligence. In *Artificial General Intelligence*. Springer Nature, 2022b

Hybrids take many forms. AlphaGo[223] is the simplest example of how approximation and search can complement one another. This hybrid crushed Go's world champion in a testament to blending search and approximation[224]. Search allowed it to plan sequences of moves that conformed to the rules of Go, while approximation allowed it to figure out which sequences of moves were most likely to win. Hybrids can also take the opposite approach. Neuro-symbolic hybrids tackle the symbol grounding problem by linking raw data to abstract concepts[225]. Think neural nets mapping inputs to symbols, then reasoning over them. Structured reinforcement learning hybrids use this kind of approach, using approximation to process sensory data and search to choose actions. Raw, high-dimensional sensory data is too much for search to cope with, so approximation 'reformats' it into a simpler, structured, low-dimensional symbolic representation. In this case a convolutional autoencoder learns to 'compress' the raw data down to a small size and then back again, ensuring important information isn't discarded by converting the sensory data to the smaller format. The low dimensional data are clustered and labelled as 'objects' with properties based on geometry and where they are on the screen. These objects can then be tracked as the world changes over time, to get learn their dynamics and spatial interactions. More conventional reinforcement learning techniques are then applied to learn a policy in these highly abstracted, symbolic terms. The resulting agent adapts far more efficiently[226]. Finally and most importantly there are fully autonomous, general purpose systems. Cognitive architectures like SOAR[227] and ACT-

[223] David Silver et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587): 484–489, 2016

[224] Michael Timothy Bennett and Yoshihiro Maruyama. Philosophical specification of empathetic ethical artificial intelligence. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2): 292–300, 2022a

[225] A. Garcez, M. Gori, L. C. Lamb, L. Serafini, M. Spranger, and S. N. Tran. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. 2019

[226] Marta Garnelo, Kai Arulkumaran, and Murray Shanahan. Towards deep symbolic reinforcement learning, 2016

[227] John E. Laird. *The Soar Cognitive Architecture*. MIT Press, MA, 2012

R[228]. These weave search and approximation together for flexible, multi-task competence. The most prominent examples are ongoing projects that have shown steady improvement year on year:

- HYPERON: Probabilistic logic networks meet neural nets in a bid for holistic cognition. Perception, memory and reasoning in one package. It aims to build AGI on a modular, distributed, self-organising system that can integrate new technology as it develops[229]. For example, new components have been proposed based on active inference and the free energy principle[230].
- AERA: The Autocatalytic Endogenous Reflective Architecture (AERA) self-programs, reflecting on its own symbolic structures while learning statistically. It's a stab at autonomy and growth[231].
- NARS: The Non-Axiomatic Reasoning System (NARS) rejects rigid axioms for a fluid, adaptive logic. NARS operates under the Assumption of Insufficient Knowledge and Resources (AIKR), reasoning with incomplete, uncertain data via a non-axiomatic framework. It integrates symbolic reasoning with probabilistic inference, using a custom inheritance-based logic (NAL) to derive conclusions from limited evidence. Designed for real-time adaptability, NARS learns incrementally, refining its knowledge base as new inputs arrive—think of it as a brain that thrives on ambiguity, not a theorem prover shackled to certainty[232].

HYBRID SYSTEMS GIVE US THE BEST OF ALL WORLDS. Fusion of search and learning is a general approach that can be scaled. Hybrids are also more useful in the short term. Structured priors or search can narrow the problem space, improving sample and energy efficiency compared to brute-force approximation. The high-level symbolic abstractions often used for search are interpretable by humans. Conversely, we can easily integrate human priors into hybrid systems. Hybridisation can be a shortcut to autonomous agents. Hybrids can combine a persistent identity and interpretable goals with the ability to process raw, high-dimensional real-world environments. For example, scaffolding like memory can enable long term adaptation in ontologically stateless language models[233]. Hybrids edge us closer to AGI by mimicking diversity of human cognition. Yet I have lingering questions. What is missing? Is a given hybrid system truly scalable or just a clever patchwork that exemplifies Sutton's bitter lesson? Can we scale these systems to AGI?

[228] John R. Anderson, Daniel Bothell, Michael D. Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. *Psychological Review*, 2004. Because apparently six authors are needed to figure out how your brain works

[229] Ben Goertzel et al. Opencog hyperon: A framework for agi at the human level and beyond. Technical report, OpenCog Foundation, 2023

[230] Ben Goertzel. Actpc-chem: Discrete active predictive coding for goal-guided algorithmic chemistry as a potential cognitive kernel for hyperon and primus-based agi, 2024

[231] Eric Nivel et al. Autocatalytic endogenous reflective architecture. Technical report, Reykjavik University, School of Computer Science, 2013; and Kristinn R. Thorisson. *A New Constructivist AI: From Manual Methods to Self-Constructive Systems*, pages 145–171. Atlantis Press, Paris, 2012

[232] P. Wang. *Rigid Flexibility: The Logic of Intelligence*. Applied Logic Series. Springer Nature, 2006

[233] Elija Perrier and Michael Timothy Bennett. Position: Stop acting like language model agents are normal agents, 2025. URL https://arxiv.org/abs/2502.10420

META-APPROACHES

A META-APPROACH IS A FRAME THROUGH which systems can be understood. It is a guiding principle I can use to tweak search, approximation or hybrid systems to be more 'intelligent'. Meta-approaches are not mutually exclusive. The scaling hypothesis is an example of a meta approach through which I have framed search and learning. I call this **scale-maxing** because it works by maximising scale. For example, maximising the amount of training data, the avaible compute and the size of the model. There are two other meta-approaches I can identify. One is orthodox at prominent AI labs like Deepmind and OpenAI. I call it **simp-maxing** because it involves maximising the simplicity of forms. It is founded on Ockham's Razor. For example, if I have a perfect compression algorithm and I use it to compress two files, then the smaller compressed file is the simpler one even if the uncompressed files were the same size. Likewise, if I use regularisation to make regression converge on a simpler function, then I am simp-maxing. The last meta-approach is my own invention, which I propose in this thesis[234,235]. I call it **w-maxing** because it optimises for the least specific, weak constraints on functionality at the lowest possible levels of abstraction. So to reiterate, scale-maxing is about maximising available resources, simp-maxing is about maximising simplicity of forms, and w-maxing is about maximising the weakness of constraints implied by function. In this section I will focus on simp-maxing, and will explain my stack-based approach later.

[234] Actually I proposed it in the papers and I rehash it here.

[235] Michael Timothy Bennett. The optimal choice of hypothesis is the weakest, not the shortest. In *Artificial General Intelligence*. Springer Nature, 2023a; Michael Timothy Bennett. A formal theory of optimal learning with experimental results. *IJCAI*, 2025f; Michael Timothy Bennett. Computational dualism and objective superintelligence. In *Artificial General Intelligence*. Springer Nature, 2024a; and Michael Timothy Bennett. What the f*ck is artificial general intelligence? *Artificial General Intelligence*, 2025c

SIMP-MAXING IS ABOUT APPLYING OCKHAM'S RAZOR to make more accurate models[236]. It posits that among competing hypotheses which might explain some observed data, the simplest one is most likely to be correct. In AI, this translates to favouring models or solutions with lower complexity, as they are less prone to overfitting and more likely to capture the underlying structure of the problem. Examples of simp-maxing include regularisation[237], the minimum description length principle[238] and Universal Artificial Intelligence (UAI)[239]. UAI is the dominant mathematical formalisation of artificial general intelligence. It relies on Kolmogorov complexity[240], which defines the complexity of a string as the length of the shortest program that can generate it. For a dataset (D), the Kolmogorov complexity $(K(D))$ is the smallest program (p) such that $U(p) = D$, where (U) is a universal Turing machine. This concept extends to models. This connected simplicity to compressibility[241]. Simpler representations are shorter, and according to Ockham's Razor simpler models are more accurate. Kolmogorov Complexity is incomputable but we can approximate it. Computable alternatives exist, like minimum description length (MDL)[242] or Lempel-Ziv compression[243]. Solomonoff[244] subsequently proposed a universal method for inductive inference based on algorithmic probability, where the likelihood of a hypothesis is proportional to $2^{-K(h)}$, with $(K(h))$ being the Kolmogorov complexity of the hypothesis (h). This formalizes Ockham's Razor in a probabilistic framework, favoring simpler hypotheses. Hutter subsequently proposed AIXI, a general reinforcement learning agent that uses Solomonoff induction to make optimal decisions based on the simplest hypotheses[245]. This gives us a theoretical frame through which to view search and approximation. We can use it to come up with practical solutions. For example, machine learning techniques like regularization (e.g. $L1$ and $L2$ norms, or dropout[246]) explicitly penalize complexity to prevent overfitting. This improves out of domain generalisation. Similarly, Pruning in decision trees reduces model size while maintaining accuracy.

THIS SERVES TO ILLUSTRATE WHAT A META-APPROACH IS. It provides a guiding principle for adaptability and generalization. A meta-approach can be applied in the context of search or approximation.

[236] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Occam's razor. *Information Processing Letters*, 1987

[237] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html

[238] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 1978

[239] Jürgen Schmidhuber. Discovering neural nets with low kolmogorov complexity and high generalization capability. *Neural Networks*, 10(5):857–873, 1997; and Marcus Hutter, David Quarel, and Elliot Catt. *An Introduction to Universal Artificial Intelligence*. Chapman and Hall/CRC, 1st edition, 2024. DOI: 10.1201/9781003460299

[240] A.N. Kolmogorov. On tables of random numbers. *Sankhya: The Indian Journal of Statistics*, A:369–376, 1963

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

[242] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 1978

[243] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, 1977. DOI: 10.1109/TIT.1977.1055714

[244] R.J. Solomonoff. A formal theory of inductive inference. part i. *Information and Control*, 7(1):1–22, 1964

[245] Marcus Hutter. *Universal Algorithmic Intelligence: A Mathematical Top→Down Approach*, pages 227–290. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007; and Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer Nature, Heidelberg, 2010

[246] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html

CONCLUSION

I'VE DEFINED INTELLIGENCE IN TERMS OF ADAPTATION, AGI as an artificial scientist and laid out some of the tools available for that quest. These include search, approximation, hybrids, meta-approaches, and the relentless march of scaling.

FOUNDATIONAL TOOLS:

- SEARCH (e.g. navigation apps, DeepBlue),
- APPROXIMATION (e.g. GPT-3, deep Q-learning).

HYBRIDS:

- SIMPLE (e.g. AlphaGo, structured reinforcement learning),
- COMPLEX (e.g. Hyperon, AERA, OpenNARS[247]).

META-APPROACHES:

- SCALE-MAXING: maximise available resources (e.g. OpenAI's GPT series LLMs),
- SIMP-MAXING: maximise simplicity of forms. (e.g. regularisation[248], UAI[249], minimum description length principle[250]),
- W-MAXING: maximise weakness of constraints implied by function[251].

EACH OFFERS A PIECE OF THE PUZZLE. With sufficient resources any system that learns can eventually attain an arbitrary level of skill. Every system can be optimal. However not all systems are equally adaptable. Hence I'll conclude this chapter by reiterating that intelligence is a matter of adaptability, and thus efficiency. All else being equal, the more resources the system needs to reach a certain level of performance, the less intelligent it is. What I offer in this thesis is a meta-approach that lets us measure and maximise adaptability.

[247] Patrick Hammer and Tony Lofthouse. 'opennars for applications': Architecture and control. In Ben Goertzel, Aleksandr I. Panov, Alexey Potapov, and Roman Yampolskiy, editors, *Artificial General Intelligence*, pages 193–204, Cham, 2020. Springer Nature

[248] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL `http://jmlr.org/papers/v15/srivastava14a.html`

[249] Marcus Hutter, David Quarel, and Elliot Catt. *An Introduction to Universal Artificial Intelligence*. Chapman and Hall/CRC, 1st edition, 2024. DOI: 10.1201/9781003460299

[250] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 1978

[251] The last I propose in this thesis.

IV. WOW, EVERYTHING IS COMPUTER

There is a problem with simp-maxing. It works, but there is no apparent reason it *should*. After all, the No Free Lunch Theorem shows no algorithm outperforms others across all problems[252,253]. Indeed, it turns out AIXI's performance is entirely subjective[254]. The root of this subjectivity is in the definition of Kolmogorov complexity. For a given string (x), its Kolmogorov complexity $K_U(x)$ is defined as the length of the shortest program that, when run on a universal Turing machine (U), produces (x). Formally: $K_U(x) = \min\{|p| \mid U(p) = x\}$ where $(|p|)$ is the length of program (p). However, this definition is inherently tied to the choice of (U). Different universal Turing machines can yield different complexity values for the same string. Specifically, for any two universal Turing machines (U) and (V), there exists a constant (c) such that: $\forall x \quad |K_U(x) - K_V(x)| \leq c$. This is the invariance theorem[255]. While this suggests that the difference in complexity is bounded, the constant (c) can be arbitrarily large in practice, making comparisons across different machines problematic. AI is inherently interactive, which means we are dealing with more than one machine. AIXI is only optimal if the UTM it uses matches some other arbitrarily chosen UTM used to measure intelligence[256]. AIXI was supposed to be the most intelligent agent according to Legg-Hutter intelligence[257]. Legg-Hutter intelligence is kind of the opposite of AIXI. AIXI is assumed to be intelligent becuase it behaves according to Ockham's Razor. In contrast, Legg-Hutter intelligence *measures* intelligence according to Ockham's Razor: agents behave in a way that can be described with shorter program are more intelligent. Simplicity is once again measured using Kolmogorov complexity. The invariance theorem claims Kolmogorov complexities only shift by a constant across UTMs[258]. This doesn't hold in an interactive setting, because in an interactive setting we have two UTMs. One with respect to which AIXI is computed, and one with respect to which Legg-Hutter intelligence is measured. If Legg-Hutter intelligence is measured with respect to one UTM, and AIXI is computed using another, then AIXI might think it has chosen

[252] D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997. DOI: 10.1109/4235.585893

[253] I show why simplicity and generalisation are correlated in this thesis, in chapter 14

[254] Jan Leike and Marcus Hutter. Bad universal priors and notions of optimality. *Proceedings of The 28th Conference on Learning Theory, in Proceedings of Machine Learning Research*, pages 1244–1259, 2015

[255] Ming Li and Paul M. B. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications (Third Edition)*. Springer Nature, New York, 2008

[256] Jan Leike and Marcus Hutter. Bad universal priors and notions of optimality. *Proceedings of The 28th Conference on Learning Theory, in Proceedings of Machine Learning Research*, pages 1244–1259, 2015

[257] Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and Machines*, pages 391–444, 2007; and Shane Legg. *Machine Super Intelligence*. PhD thesis, Uni. of Lugano, 2008

[258] Ming Li and Paul M. B. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications (Third Edition)*. Springer Nature, New York, 2008

a short program that Legg-Hutter intelligence interprets as a long program. For instance, a string that appears simple relative to one Turing machine might seem complex relative to another, depending on the machine's instruction set or encoding scheme. Consider the analogy of programming languages, which can be thought of as different universal Turing machines. Suppose we have two programming languages, L_1 and L_2. Language L_1 has a built-in function that directly generates the Fibonacci sequence, while L_2 does not. Now, consider a string (x) that represents the first 100 Fibonacci numbers. In L_1, the shortest program to generate (x) might be a single function call, say `print_fib(100)`, making $K_{L_1}(x)$ very small. In contrast, in L_2, the shortest program would need to implement the Fibonacci sequence from scratch, resulting in a much larger $K_{L_2}(x)$. Thus, the same string (x) has vastly different complexities depending on the chosen language, illustrating the subjectivity introduced by the choice of reference machine.

REFRAMING THE PROBLEM

AGI IS SUPPOSED TO BE CAPABLE OF ADAPTING to any task or environment. However, if its internal measure of simplicity is tied to a specific, arbitrary choice of reference machine, its adaptability may be constrained by that choice. This could lead to blind spots or inefficiencies in certain domains, undermining the goal of general intelligence. AIXI illustrates an extremely valuable idea, but its subjectivity is a problem. Some have explored complexity measures that are invariant under certain transformations, aiming to reduce dependence on the reference machine. For example, Levin complexity[259] incorporates time complexity into the measure, potentially offering a more universal metric. However this is still a measure of form, not function. Simp-maxing is not optimal, and if I want to understand intelligence I need to know what I'm aiming at. I want to know what the upper bound on adaptability is.

[259] L. A. Levin. Universal sequential search problems. *Problems of Information Transmission*, 9(3):265–266, 1973

TO SOLVE THIS PROBLEM I need to go down a level of abstraction. I need to reframe the problem. Kolmogorov complexity takes information in one format A and represents it in another B. B is a language. It is how we represent information in a Universal Turing Machine (UTM). It is in B that length is measured, and length depends on how we format information in B. Short isn't universal[260]. It is tied to the UTM you pick. The UTM is an **interpreter**. B is an **abstraction layer** on top of A. New UTM, new definition of simple, new AIXI. In hindsight this seems obvious. Lieke and Hutter also concluded that Pareto optimality is trivial. That seems less obvious to me, so I proposed an analogy to describe the issue. Lets assume the environment is a function f_1: it takes AIXI's actions and coughs up observations and rewards. The UTM is f_2: it decodes AIXI's guesses, which are programs describing what happens next. AIXI's algorithmic, software 'mind' is f_3: it churns out those guesses. The reward r comes from $f_1(f_2(f_3))$. You can't judge AIXI by f_3 alone. It's the **stack** that deterimines success, by which I mean the environment, UTM and algorithm together. Switch the UTM, and the output changes. Pareto optimality is only trivial if you can change part of the stack. If we consider the entire stack, then Pareto optimality is far from trivial. The problem is that Kolmogorov complexity is a matter of form, not function[261]. Whatever notion of complexity we use, it is a matter of form[262]. Software is just a state of hardware[263]. Any claim regarding a software mind are symptomatic of a condition I call **computational dualism**. It is pointless to make claims about an optimal software mind if the environment and interpreter can be changed.

[260] Laurent Orseau. Asymptotic non-learnability of universal agents with neural networks. In Joscha Bach, Ben Goertzel, and Matthew Iklé, editors, *Artificial General Intelligence: 5th International Conference, AGI 2012*, pages 234–243, Berlin, Heidelberg, 2012. Springer Nature

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

[262] L. A. Levin. Universal sequential search problems. *Problems of Information Transmission*, 9(3):265–266, 1973; Gregory J. Chaitin. On the length of programs for computing finite binary sequences. *J. ACM*, 1966; and Jorma Rissanen. Modeling by shortest data description. *Automatica*, 1978

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

MORTALITY

Descartes thought it was 'animal spirits' and the pineal gland passing messages between mind and body. AI research has replaced the pineal gland with a Turing machine. It is a ghost haunting AGI labs like it's 1637. Software is the mind, hardware is the meat, and never shall they meet. Computational dualism is the idea that artificial intelligence is about creating intelligent software[264]. It is not. That should be obvious. Software is a state of hardware. AI is about making an intelligent system, and its state its part of it. Computational dualism is a problem if we want to build an intelligent system, because it ignores half the equation. We need to know what intelligence is if we want to optimise for it. We need to know what optimal looks like so we can work towards it.

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

It is not as if people haven't already pointed out there's a problem, it is just that they have chosen to live with it[265]. As far as I can tell, only one other has gone to the trouble of attempting to formalise an alternative[266]. Orseau attempted to formalise a version of AIXI which was interpreted by the environment, using bounded optimality. It is a commendable and compelling attempt, but it does not go far enough. It does not answer the questions I want answered. Symptoms of computational dualism remain. Software is a convenient abstraction and it works well for building standardised applications for standardised hardware in standardised contexts. It becomes more of a problem when we consider an agent interacting with the world. Many appear to have forgotten software is nothing more than a state of hardware. It has spawned wild AGI myths, like superintelligent code rewriting physics or escaping its box[267]. Even Nobel laureate Geoffrey Hinton has resorted to doomsaying[268].

[265] J. A. Fodor. Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Sciences*, 3(1):63–73, 1980. DOI: 10.1017/S0140525X00001771

[266] Laurent Orseau and Mark Ring. Space-time embedded intelligence. In Joscha Bach, Ben Goertzel, and Matthew Iklé, editors, *Artificial General Intelligence*, pages 209–218, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-35506-6

He speaks of mortal vs immortal computation as if software lives on in the absence of hardware[269]. His arguments seem to hinge on software's ability to leap from one hardware platform to another, retaining its functionality like some eternal digital essence. He suggests that because software can be duplicated across different machines without losing what it does, computation somehow transcends its hardware. At first glance, it seems plausible. Copy your code, run it elsewhere, and the process lives on. But that's an illusion. Software is a state of hardware. When the hardware changes or fails, the computation changes or fails. Copying doesn't preserve the original. It births a new instance as mortal as the last.

[267] Laurent Orseau. Asymptotic non-learnability of universal agents with neural networks. In Joscha Bach, Ben Goertzel, and Matthew Iklé, editors, *Artificial General Intelligence: 5th International Conference, AGI 2012*, pages 234–243, Berlin, Heidelberg, 2012. Springer Nature

[268] Zoe Kleinman and Chris Vallance. AI 'godfather' Geoffrey Hinton warns of dangers as he quits Google. *BBC News*, May 2023. URL `https://bbc.com/news/world-us-canada-65452940`. Accessed: 2025-03-13

[269] Geoffrey Hinton. The forward-forward algorithm: Some preliminary investigations, 2022

HINTON'S IMMORTAL COMPUTATIONS DON'T EXIST. There is only mortal software, because there are only finitely many devices. Software is a state of hardware, a pattern etched in silicon or flesh. Change the hardware, and the 'mind' follows. Human intelligence is embodied, not just symbol shuffling in the abstract[270]. Software is no different. The analogy of the brain as a computer only works if you treat it as a unified system with no distinction between hardware and software[271]. Intelligence isn't a disembodied mind interacting with but a dance between hardware and environment. Cognition is a physical act, not a ghost in a shell[272]. Every physical system computes simply by existing[273]. It is a whole-of-system physical process. Hence, I take a whole-of-system approach. Hardware, software and world are entwined.

[270] Hubert L. Dreyfus. *What Computers Can't Do: A Critique of Artificial Reason*. Harper & Row, 1972

[271] Oron Shagrir. Why we view the brain as a computer. *Synthese*

[272] Daniel Hutto and Erik Myin. Radical enactivism: Basic minds without content, 2013

[273] Gualtiero Piccinini and Corey Maley. Computation in Physical Systems. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, Stanford, Sum. 21 edition, 2021

FLIPPING THE TABLE

COMPUTATIONAL DUALISM IS A DEAD END. Software is a state of hardware. Hardware is a part of a larger system. Here I argue everything is nested abstraction layers from software to hardware to the laws of physics. This brings together several of my papers[274]. To quote a certain ambulatory meme, everything is computer.

SOFTWARE DOESN'T EXIST. At least, not in the way people seem to think. A Python script relies on an interpreter written in C. C is compiled into assembly. Assembly to machine code. Machine code is hard-wired in silicon. We don't 'run' software. We flip switches in a box. As AIXI illustrates, code is nothing if it is not etched in meat or metal. Same code, different rig, different mind. Why? Because a body isn't a neutral translator. Hardware is goal-directed, just like software is. Some hardware is better for some tasks. It is an abstraction layer.

ABSTRACTION DOES NOT END AT HARDWARE. Hardware is not the bedrock. Intuitively, think of a play. Software is a script, the hardware is the actor and the environment is the stage. Change the actor or the stage, and the show is not the same. An actor is not the play. Making hardware the foundation would repeat the mistake of computational dualism. Hardware is a body embedded in the world and bound by physics[275]. A CPU is a hunk of matter obeying laws we've barely glimpsed. Physical laws, only they are not really 'laws'. That is just how we understand what is happening[276]. We scribble on blackboards. We approximate reality. Whatever it is that those laws approximate, that is hardware's puppeteer. Nature's machinery. A transistor flips because nature says so, not because some coder waved a wand. Hardware is a middleman, enacted by something less abstract. Hardware is just another abstraction layer, like software.

[274] Michael Timothy Bennett. Computational dualism and objective superintelligence. In *Artificial General Intelligence*. Springer Nature, 2024a; Michael Timothy Bennett. Is complexity an illusion? In *Artificial General Intelligence*. Springer Nature, 2024b; and Michael Timothy Bennett. Are biological systems more intelligent than artificial intelligence? 2025b. Forthcoming 2026

[275] Hubert L. Dreyfus. *What Computers Can't Do: A Critique of Artificial Reason*. Harper & Row, 1972; Hubert L. Dreyfus. Why heideggerian ai failed and how fixing it would require making it more heideggerian. *Philosophical Psychology*, 20(2):247–268, 2007. DOI: 10.1080/09515080701239510. URL `https://doi.org/10.1080/09515080701239510`; and Michael Wheeler. Martin Heidegger. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, Fall 2020 edition, 2020

[276] Bas C. van Fraassen. *Laws and Symmetry*. Oxford University Press, 1989

PUT ANOTHER WAY, REALITY IS A MATRYOSHKA DOLL. Software is a state of hardware, which is a state of the physical reality we inhabit. A human is a state of organs, which are states of cells, which are states of molecules[277]. Each level is a state of the one below. Everything a program running on a deeper machine.

I CALL IT THE STACK. It is patterns within patterns[278]. I'll return to my function analogy. The mind is f_3, the body f_2 and the local environment is f_1. Now can add f_0 for our physical laws or whatever the local environment is running on. The underlying hum of reality. The stack is then $f_0(f_1(f_2(f_3)))$. A computer is the same. We can break things down into more granular detail. f_n could be Python, f_{n-1} C, f_{n-2} assembly, f_{n-3} machine code, f_{n-4} a particular machine, f_{n-5} the local environment and infrastructure with which the computer interacts all the way down to physics f_0. Hardware is not a sacred boundary where abstraction stops and reality kicks in. Hardware is as abstract as code. A computer is designed by a human, but it obeys the laws of physics. It follows a script we did not write[279].

[277] Michael Timothy Bennett. Are biological systems more intelligent than artificial intelligence? 2025b. Forthcoming 2026

[278] Ben Goertzel. *The Hidden Pattern: A Patternist Philosophy of Mind*. Brown-Walker Press, USA, 2006

[279] Bas C. van Fraassen. *Laws and Symmetry*. Oxford University Press, 1989

THE LIMITS OF KNOWING

How far down can the stack go? Gravity, quarks and spacetime sound foundational, but they are human abstractions. Our physics is a guess scrawled in chalk by apes[280]. We know it is incomplete. Our formulae are programs running on a human brain. They are not fundamental[281]. Even numbers are not fundamental. Numbers are just a means by which we describe the order we perceive. Could there be more? Maybe $f_{-1}...f_{-n}$ layers beneath physics, like a simulation running our reality? This question resembles Derrida's differance. Every layer defers to the next, and we cannot find the bottom[282]. The Stack might stretch forever, or it might hit a wall. I don't know.

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

[281] W.V.O. Quine. *Philosophy of Logic: Second Edition*. Harvard University Press, Cambridge MA, 1986. ISBN 9780674665637. http://www.jstor.org/stable/j.ctvk12scx

[282] Jacques Derrida. Writing and difference. *U of Chicago P*, 1978

Does this mean we are condemned to subjectivity? Solipsism? Intelligence is a whole system[283]. It is the whole stack in motion. To understand intelligence we must rethink reality.

[283] Oron Shagrir. Why we view the brain as a computer. *Synthese*

Our physical laws are models. They are programs we've written to predict nature's machinery. We are the computers on which those programs run. Our tools are built for our slice of reality, not the whole pie[284]. There could be $f_{-\infty}$ layers stacking down forever. I need a tool that doesn't care. I need to identify what is true across every stack. Across every world. Anything less would be computational dualism all over again.

[284] Jacques Derrida. Writing and difference. *U of Chicago P*, 1978; and J. Speaks. Theories of Meaning. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, Stanford, Spring 2021 edition, 2021

I'm going to propose a definition of environment that holds for every environment. It is the foundation of what I call Stack Theory[285]. It is a frame that holds no matter where the bottom lies. It's not about finding f_0. It is about sidestepping the need to know. Intuitively, we're ants on a leaf, guessing at the tree. The only way to know about the tree is to work out what must be true of all trees our leaf might be attached to. Hutter's Universal Artificial Intelligence was the right idea, but it made the wrong thing universal because it was hamstrung by computational dualism. We don't need a universal description of *intelligence*, we need a universal description of the entire *stack*.

[285] To avoid ambiguity, note that Pancomputational Enactivism refers to the formalism of enactive cognition *based* on Stack Theory.

ALL POSSIBLE WORLDS

HERE I LAY THE FOUNDATION of cognition *within* **Stack Theory.** Cognition within Stack Theory is enactive[286] and pancomputational[287,288]. Pancomputationalism says all physical systems are computational. Enactivism frames cognition as emerging from dynamic interactions between the system and its environment.

STACK THEORY'S FOUNDATION is the **environment**. More specifically, it is a definition I name 'environment', but really it is what is common to all environments. All possibilities for an 'underlying physics'. It is based on premises I refined over the course of several publications[289]. In those papers I called them axioms, but I'm no longer sure that description fits. They don't depend on anything. They are not assumptions. In the first 'axiom' I merely define what I mean by environment. The second is a tautology.

- AXIOM 1: Where there are things, I call them the environment.
- AXIOM 2: If things change, then the environment has states.

What is a state? At the very least, it is a difference. If nothing changed then there could be no states. Without states the environment can only be some sort of unity or oneness. There is nothing in it we could point to. It just is. Perhaps even that is arguable. If something has no state and no content, is it anything? There must be difference for there to be something. Since there must be difference, there must be states. If one thing changes, then there must be two states. Before, and after. We don't know what the thing is or what states are, and we don't need to. That is unnecessary detail. All we need to know is that there is a difference between states. I don't presuppose the environment is made up of objects or properties. Because of this there is a different, equivalent axiom we might use.

- ALTERNATIVE AXIOM 2: Time is difference.

Every state is point of difference is a different time in a particular timeline. This means states are **mutually exclusive** within a timeline.

[286] Evan Thompson. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Harvard University Press, Cambridge MA, 2007; John Vervaeke, Timothy Lillicrap, and Blake Richards. Relevance realization and the emerging framework in cognitive science. *J. Log. Comput.*, 2012; John Vervaeke and Leonardo Ferraro. *Relevance, Meaning and the Cognitive Science of Wisdom*. Springer Netherlands, Dordrecht, 2013a; John Vervaeke and Leonardo Ferraro. Relevance realization and the neurodynamics and neuroconnectivity of general intelligence. In Inman Harvey, Ann Cavoukian, George Tomko, Don Borrett, Hon Kwan, and Dimitrios Hatzinakos, editors, *SmartData*, NY, 2013b. Springer Nature; and Daniel Hutto and Erik Myin. Radical enactivism: Basic minds without content, 2013

[287] Gualtiero Piccinini. *Physical Computation: A Mechanistic Account*. Oxford University Press, UK, 2015; and Gualtiero Piccinini and Corey Maley. Computation in Physical Systems. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, Stanford, Sum. 21 edition, 2021

[288] Hence I often refer to it as Pancomputational Enactivism.

[289] Michael Timothy Bennett. Computational dualism and objective superintelligence. In *Artificial General Intelligence*. Springer Nature, 2024a; Michael Timothy Bennett. Is complexity an illusion? In *Artificial General Intelligence*. Springer Nature, 2024b; and Michael Timothy Bennett, Sean Welsh, and Anna Ciaunica. *Why Is Anything Conscious?* Preprint, accepted to and presented at ASSC27 and MoC5, 2024

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 **expressed**, **realised**[290] or **embodied** in state ϕ if it is an aspect of ϕ.*

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

Everything that is or might be must fall within the scope of what this formalism can describe. Yes, my formalism is still an abstraction. However, some claims are so weak they are true of everything[291]

[291] It can be referred to as *Stack Theory* because it has to be true no matter how far down the stack we go.

We have a set of states Φ. Each state $\phi \in \Phi$ represents a particular configuration of the environment at a given moment, capturing its current condition or state of affairs whatever that may be. The particulars don't matter, just that each state represents a difference from other states. Non-equality. The power set $2^{\Phi} = P$ of Φ, is all possible subsets of states, which I call declarative **programs**. Here, a declarative program is not a traditional algorithm but a subset of states. That program returns 'true' about states it contains.

A truth or fact about a state ϕ is any program (f) that includes ϕ , meaning (f) is true for that state. For the sake of intuitive example, if $\Phi = \{\texttt{on}, \texttt{off}\}$ and $f = \{\texttt{on}\}$, then (f) is true for the state `on` . Now, this is a toy example. `on` and `off` are high level human abstractions. They are at the top of the stack, states are at the bottom of the stack, and the stack may be infinite. Nevertheless if we are somehow omniscient and given a state of the environment, then the sum total of everything that is true is the programs that contain that state. You can think of a true program as marking out points of sameness between states. If a program is true about two states, then that program is something they share in common. If a program is true of one state but not another, then it is a point of difference which separates them. Since each state represents a single point of difference, the set of all things which can be different or the same is the powerset of states P.

The environment encodes everything through its state space. Whether objective or subjective every object, property, and goal is an **aspect** of the environment. An **aspect** of a **state** is a collection of facts

that all hold for that state, such as {`light is on, door is closed`} for a state where both are true. An aspect of the **environment** is one that holds for *at least* one state, and it is realised by a state if that state satisfies all the facts in the aspect. This formal structure allows us to model everything as programs, aligning with pancomputationalism's view that all physical processes are computational[292].

[292] Gualtiero Piccinini and Corey Maley. Computation in Physical Systems. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, Stanford, Sum. 21 edition, 2021

TOY EXAMPLES

To demonstrate the framework's generality, consider several examples from diverse domains. These illustrating how the environment can **represent** different systems. Now, in reality we don't know what states contain. We see perceive them through a possibly infinite stack of abstraction layers. However, for the sake of example lets assume we are omniscient. This lets me use the framework to describe toy problems and 'real world' examples. In reality Φ contains everything, but for the sake of example I pretend we have these very specific universes. Later, I will argue such things are abstraction layers but for now ignore that detail.

Light Switch System: Let $\Phi = \{\texttt{on}, \texttt{off}\}$, the set of states for a light switch. Programs include $f_1 = \{\texttt{on}\}$ (light is on) and $f_2 = \{\texttt{off}\}$ (light is off). A fact about the state `on` is f_1, since $\texttt{on} \in f_1$. An aspect could be $\{f_1\}$, realised by the state `on`. This simple example shows how even basic devices fit the framework, with states and programs defining goals.

Grid World in AI: In a grid world, Φ is all possible positions of an agent and reward locations, e.g., $\Phi = \{(x, y, r_x, r_y) \mid x, y, r_x, r_y \in \{1, 2, \ldots, n\}\}$, where $((x, y))$ is the agent's position and (r_x, r_y) is the reward's position. If we have a program $f = \{(x, y, r_x, r_y) \mid x = r_x \texttt{ and } y = r_y\}$ is true, then the agent is at the reward. Otherwise it is not. This aligns with reinforcement learning, where the agent interacts to achieve goals[293].

[293] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, MA, 2018

Biological Cell Metabolism: A cell's environment includes metabolic states (e.g., healthy, stressed, dividing) and external conditions (e.g., nutrient levels). Let Φ be the set of all such states, with programs like $f_1 = \{\texttt{states where cell is healthy}\}$ or $f_2 = \{\texttt{nutrient levels normal}\}$. This illustrates that sometimes one program can be a subset of another, so $f_2 \subset f_1$. An aspect could be $\{f_1, f_2\}$, realized by states where both hold.

These examples illustrate the framework's flexibility, applying to digital, biological, and social systems, each with distinguishable states and goals.

V. TURTLES ALL THE WAY DOWN

So far I've framed the environment as a set of states Φ. The contents of states are defined only by their differences from one another. These differences are formalised as programs, which are subsets of Φ. Elements of the powerset $P = 2^{\Phi}$. An aspect of the environment is a set of programs, and the aspect is *realised* or *exists* if it is true given a state[294]. It's a minimalist setup that makes no assumptions. Yes, it is an abstraction but some abstractions are so weak they are true of everything. This particular abstraction holds for all possible environments. That is the point. Now I'm going to talk about embodiment.

[294] A present state, or a point in time etc. Truth is reference dependent here.

Embodiment gets overlooked in computer science. That is why we have computational dualism. Anecdotally, when I have presented this research at conferences many of the questions I received strawmanned embodiment. The implication was that embodiment was a matter of sentimentality, or that I was arguing there is something non-computational about intelligence. After all, some proponents of enactive cognition believe that computation and true enactive cognition are incompatible[295]. However embodiment as I speak of it is just a fact of existence. Every body, whether it be human, machine, or a slab of granite, throws its weight around dictating what can happen next. A rock doesn't care about your feelings, but drop it in a pond and the ripples tell a story. I am framing this as a kind of ontological "speech". Not poetry, but a formal language baked into existence itself. Think of it as ontology with attitude. Entities *say* something by being what they are.

[295] Johannes Jaeger, Anna Riedl, Alex Djedovic, John Vervaeke, and Denis Walsh. Naturalizing relevance realization: Why agency and cognition are fundamentally not computational. *Frontiers in Psychology*, 15, 2024

LAYER CAKE

THE ENVIRONMENT SPEAKS IN PHYSICAL TERMS. Recall the definition of **environment** from the previous chapter:

- 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 **expressed**, **realised**[296] or **embodied** in state ϕ if it is an aspect of ϕ.

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

IF EVERY PHYSICAL SYSTEM COMPUTES[297], then every physical system embodies a formal language. The environment is a physical system, so that means I should be able to re-frame it as a formal language. P could be a **vocabulary**, and every aspect the environment a **statement** in this formal language. The set of all things the environment can *say* would then be the set of all aspects. Time is difference, so the only way we could have two different states at the same time would be if we had two different *worlds*. Seems rather like Everett's interpretation of quantum physics[298]. Conversely, given a particular world there can only be one state at a time. That means aspects of the environment that are never realised by the same state never coexist in the same world. They are *mutually exclusive*. I could use that to build something like a logical `nand` gate. A `nand` gate $n \subset P$ would be an aspect of the environment, sure, but it is also more than that. Like the environment as a whole has a global state, a `nand` gate has its own local state. The different is a matter of detail. If n is the aspect of the environment that is the `nand` gate in all its states, then n is what *does not* change when the `nand` gate's state changes. Each state of the `nand` gate is a more *specific* aspect of the environment than n. If we want to formalise all these things together, then we need a subset $\mathfrak{v}$ of P that contains the more specific aspects. I'll give an example of this. First, I'll define this formal language:

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

[298] David Wallace. *The Emergent Multiverse: Quantum Theory according to the Everett Interpretation*. Oxford University Press, 05 2012. ISBN 9780199546961. DOI: 10.1093/acprof:oso/9780199546961.001.0001. URL `https://doi.org/10.1093/acprof:oso/9780199546961.001.0001`

Definition 2 (abstraction layer)
By abstraction layer[299] *I mean:*

- *We single out a subset* $\mathfrak{v} \subseteq P$ *which we call* ***the vocabulary*** *of an abstraction layer. The vocabulary is finite unless explicitly stated otherwise. 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****.*

[299] (NOTATION) E with a subscript is the extension of the subscript. For example, E_l is the extension of l.

(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 the sub-table of x's truth table for which x is true.

- *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.*
- *The **extension of a statement** $x \in L_{\mathfrak{v}}$ 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 $E_x = E_y$.*

Our nand gate is embodied in silicon with inputs $\mathtt{a}, \mathtt{b} \in \{0,1\}$ and output $\mathtt{c} = \mathtt{nand}(\mathtt{a}, \mathtt{b})$[300]. For the sake of giving some clear intuition as to how this works I'm going to violate my own rule and write the states out as having specific contents rather than being *contentless*.

[300] Forgive the abuse of notation, for the purpose of this line think of nand as a function in $\{0,1\}$.

- states: $\Phi = \mathtt{001} \cup \mathtt{011} \cup \mathtt{101} \cup \mathtt{110} \cup \neg\mathtt{nand}$ where each value (e.g. 001) denotes a set containing all states in Φ where a, b and c equal those values[301], and $\neg$nand contains all other states[302].
- vocabulary: $\mathfrak{v} = \{f_\mathtt{a}, f_\mathtt{b}, f_\mathtt{c}\}$ s.t.
  - $f_\mathtt{a} = \mathtt{101} \cup \mathtt{110}$[303]
  - $f_\mathtt{b} = \mathtt{011} \cup \mathtt{110}$[304]
  - $f_\mathtt{c} = \mathtt{001} \cup \mathtt{011} \cup \mathtt{101}$[305]
- statements ($L_{\mathfrak{v}}$): Subsets $l \subseteq \{f_\mathtt{a}, f_\mathtt{b}, f_\mathtt{c}\}$ with $\bigcap l \neq \emptyset$, e.g., $\{\}, \{f_\mathtt{a}\}, \{f_\mathtt{c}\}, \{f_\mathtt{a}, f_\mathtt{b}\}$, but not $\{f_\mathtt{a}, f_\mathtt{b}, f_\mathtt{c}\}$ ($\cap = \emptyset$).
- behaviour: $f_\mathtt{c} = \Phi \setminus ((f_\mathtt{a} \cap f_\mathtt{b}) \cup \neg\mathtt{nand})$, so $\{f_\mathtt{c}\}$ is true iff $\{f_\mathtt{a}, f_\mathtt{b}\}$ is false and the gate is still operational.

[301] For example, 001 contains all the states where $\mathtt{a} = 0$, $\mathtt{b} = 0$ and $\mathtt{c} = 1$

[302] For example states where the gate is off or destroyed.

[303] ($\mathtt{a} = 1$)

[304] ($\mathtt{b} = 1$)

[305] ($\mathtt{c} = 1$)

Given a nand gate I can build a computer[306]. The nand is the basic building block of all computers today.

[306] Note that in the above example, none of $f_\mathtt{a}, f_\mathtt{b}, f_\mathtt{c}$ contain the aspect n. This will become important in later chapters when I introduce causal-identities.

I return to Grid World for another illustrative example. Because we now have this formal definition of abstraction layer, we can consider how Grid World exists *within* our reality, rather than as a separate, simplified reality. In other words I assume there is a machine in the environment that computes Grid World. That machine is built out of nand gates that together form an abstraction layer for Grid World. Lets not worry about Φ now, because Φ is unknowable and infinite. We can only see our subjective abstraction layer. Grid World needs positions of an agent and reward locations[307]. Lets say $\mathtt{positions} = \{(x, y, r_x, r_y) \mid x, y, r_x, r_y \in \{1, 2, \ldots, n\}\}$, where $((x, y))$ is the agent's position and (r_x, r_y) is the reward's position. These are

[307] Again, I have violated my own rule and written out contents for these states for your intuition.

just declarative programs, meaning $\texttt{positions} \subset P$. To properly embody Grid World we also need programs like $g = \{(x, y, r_x, r_y) \mid x = r_x \texttt{ and } y = r_y\}$ so that we can describe the goal state (when the agent is at the reward) and all possible actions in all possible orders $\texttt{actions} = \{\texttt{up}_1, \texttt{up}_2, \texttt{left}_1, \texttt{right}_1 ...\}$. The end result is we need an machine (in this case made of nand gates) that physically embodies a vocabulary $\mathfrak{v} = \texttt{positions} \cup \texttt{actions} \cup \{g\}$, so that it can *embody* at least every state in Grid World. It *could* embody more, but that would consume resources that could be better spent on only what is *relevant*. After all, every computation has a physical cost[308].

[308] R. Landauer. Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, 5(3):183–191, 1961; and Seth Lloyd. Ultimate physical limits to computation. *Nature*, 406(6799): 1047–1054, 2000

Every body carries a vocabulary, a subset $\mathfrak{v} \subseteq P$ of programs it can enact. Think muscle twitches, photon emissions, or gear shifts. In a computer, the vocabulary contains possible truths the system can physically encode and embody. A vocabulary is like a boundary of the system, at least in terms of its ability to process information as a coherent whole. For example, the programs in $\mathfrak{v}$ can describe every possible configuration of every possible bit in the system. Statements are subsets of $\mathfrak{v}$ that can hold together without clashing. For example, if $l \subset \mathfrak{v}$ is a statement then their intersection $\bigcap l \neq \emptyset$, meaning there exists a state where the statement is a true aspect of the environment. When the environment's state ϕ hits that sweet spot, the statement is *realised*. It becomes a tangible fact carved into reality. Again, this distinction between subjective, semantic truth and existence is important because we're trying to understand issues of consciousness. Delegating interpretation to the underlying states delegates the problem of interpretation to whatever the underlying physics of reality happen to be. It obviates the need for a translator and lets us *ground* symbols in a sense even Derrida might accept[309]. A bent knee isn't a stand-in for `knee bent`. It *is* `knee bent`. No middleman but the thing itself. The environment sets the rules and calls the shots. It cycles through states one at a time within the confines of a given world... or branching into many worlds[310], either works but for the sake of explanation I will confine myself to one particular timeline. Each $\phi \in \Phi$ greenlights some programs while axing others. Existence evolves with every tick of time[311]. Picture a robot clawing its way through a maze. Its vocabulary: sensor blips (`wall close`, `path open`) and motor grunts (`pivot right`, `lurch forward`). A statement like `wall close, pivot left` doesn't fit in its embodied circuitry. It can't turn left. It can't represent left. It can turn right. The screech of its servos turning right as the sensor pings. That motion *is* the statement, alive in the grind of metal on floor[312]. No just the pondering, but the doing. This dodges old traps like Searle's symbol-shuffling room[313]. Bodies don't represent reality. They are *aspects* of it.

[309] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346, 1990. ISSN 0167-2789. DOI: https://doi.org/10.1016/0167-2789(90)90087-6. URL `https://www.sciencedirect.com/science/article/pii/0167278990900876`; and Jacques Derrida. Writing and difference. *U of Chicago P*, 1978

[310] David Wallace. *The Emergent Multiverse: Quantum Theory according to the Everett Interpretation*. Oxford University Press, 05 2012. ISBN 9780199546961. DOI: 10.1093/acprof:oso/9780199546961.001.0001. URL `https://doi.org/10.1093/acprof:oso/9780199546961.001.0001`

[311] Ilya Prigogine. *From Being to Becoming: Time and Complexity in the Physical Sciences*. W.H. Freeman, 1980

[312] Andy Clark. *Being There: Putting Brain, Body, and World Together Again*. MIT Press, 1997

[313] John Searle. Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3:417–457, 1980

SUBJECTIVE AND OBJECTIVE

When a body expresses a statement $l \in L_{\mathfrak{v}}$ it filters the possibility space. Take a statement x. Its extension E_x is the full roster of statements that imply it. If x is `the car is rolling`, E_x might include `the car is rolling downhill with the brakes shot.` You might think of it as under-specification. A vague, *weak* statement (`system's active`) maps to a swarm of more specific, *stronger* statements (`gears turning, lights flashing`). If two statements x and y share the same extension ($E_x = E_y$), they are implied by the same set of statements. Within an abstraction layer, this is like having the same truth conditions.

Now if I were omniscient, the environment would have one state at a time because time is difference. That state would determine what is true at that time. That would mean some programs would return true at that time, and I could know the rest to be false. Truth would be binary. The world deterministic. Everything that exists is a statement made in an environment's embodied formal language, and which statements are true depends on the state.

But I am not omniscient. From my subjective perspective within my environment, I cannot know what the physical state is. I cannot see all the statements. I am a statement, and I exist for as long as the environment expresses me. The environment might be objectively deterministic, but from my subjective point of view it is non-deterministic. There are many possible futures. 'Many worlds' in which I may find myself, like Everett's interpretation of quantum physics[314]. Every statement x has an **extension** E_x, which is the set of all statements in the language that imply the statement x. These many possible worlds or futures are my extension.

By expressing a statement x, the environment is constrained. It can only be in states that express x. If the environment expresses both x and y then the possibilities are constrained even further, to the intersection of their extensions $E_x \cap E_y$. In this sense, statements bump up against each other. They clash. Just as the same constraint can be realised by different systems[315], the same extension can be realised by different bodies or combinations of bodies. Intuitively, this reflects how the parts of a distributed, complex system interact. Upward and downward causation. For example assume the environment expresses cells. Those cells can interact to constrain each other's behaviour and develop a collective identity[316].

[314] David Wallace. *The Emergent Multiverse: Quantum Theory according to the Everett Interpretation*. Oxford University Press, 05 2012. ISBN 9780199546961. DOI: 10.1093/acprof:oso/9780199546961.001.0001. URL `https://doi.org/10.1093/acprof:oso/9780199546961.001.0001`

[315] Robin Gandy. Church's thesis and principles for mechanisms. In *The Kleene Symposium*. North-Holland, 1980; Ricard Solé et al. Fundamental constraints to the logic of living systems. *Interface Focus*, 2024; and Oron Shagrir. Why we view the brain as a computer. *Synthese*

[316] Patrick McMillen and Michael Levin. Collective intelligence: A unifying concept for integrating biology across scales and substrates. *Communications Biology*, 2024

Consider a human raising its arm. This sets in motion a particular future. When a body moves, it *embodies* a **statement**. A statement (l) is expressed when the environment's state ϕ lands in $\bigcap l$. The underlying physics may be the true interpreter, but within the confines of an abstraction layer we have access only to the programs. Objectively all programs are true or false, but from within the confines of an abstraction layer a program is *subjectively* true, false or unknown because the underlying state is unknown. Only the programs in the abstraction layer are accessible. This is fine. A rock rolling downhill isn't pondering its path. It is merely interacting as part of a larger system. This is the loosest possible interpretation of computation. Just physics as the engine, no software required[317].

Each body has a vocabulary. A human is a chaotic symphony. A rock grunts single syllables. But each fits into the larger machine that is the environment. Computation here is the interaction of the body with its world. It *affords* the surround environment something[318]. The world offers possibilities tailored to a body's shape. A chair yells "sit" to a human, not to a boulder. The statements a body can pull off depend on what the environment hands it. This aligns with ideas like polycomputation, that a computation at one scale can perform an entirely different role as part of a computation at a larger scale[319]. The same matter is part of many larger and smaller computations. This is a rejection of both the old computational mind[320], and strong enactivism that holds cognition to be non-computational[321].

[317] C. Horsman, S. Stepney, R. C. Wagner, and V. M. Kendon. When does a physical system compute? *Proceedings of the Royal Society A*, 470(2169):20140182, 2014

[318] James J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979

[319] Joshua Bongard and Michael Levin. There's plenty of room right here: Biological systems as evolved, overloaded, multi-scale machines. *Biomimetics*, 8(1), 2023

[320] Jerry A. Fodor. *The Language of Thought*. Harvard University Press, 1975

[321] Johannes Jaeger, Anna Riedl, Alex Djedovic, John Vervaeke, and Denis Walsh. Naturalizing relevance realization: Why agency and cognition are fundamentally not computational. *Frontiers in Psychology*, 15, 2024

MATRYOSHKA DOLLS

A STATEMENT IS A SET OF PROGRAMS, but it is also **equivalent** to a program that has the same extension. Formally, I mean for every statement $x \subset P$ there exists a program $f \in P$ such that $E_x = E_{\{f\}}$. Hence I can map every statement's extension to a set of equivalent programs. If I have a `nand` gate n it has a very specific vocabulary as discussed earlier, but it can also be *part* of a larger system and thus have a much larger vocabulary. At most, it can be part of all the systems encompassed by its extension E_n. We can re-frame n to an abstraction layer, much like how we treat Python as an abstraction layer over C. All we need to do is convert E_n to a set of equivalent programs, and we have the vocabulary of a new abstraction layer. A second order abstraction layer over the environment. I formalise this using an *abstractor* function:

Definition 3 (abstractor function) $\mathfrak{f} : 2^P, 2^P \rightarrow 2^P$ *is an* ***abstractor*** *function that takes a vocabulary* $\mathfrak{v}$ *and a statement* $l \subset \mathfrak{v}$*, and returns a new vocabulary* $\mathfrak{v}' = \{f \in P : \exists o \in E_l(\bigcap o = f)\}$.

Naturally I could also do this with a more constrained extension, taking into account other parts of the environment and how they constrain n. I can take a vocabulary $\mathfrak{v}$ and form statements x, y, z. They can interact to give me the combined extension $E_x \cap E_y \cap E_z$ implied by $x \cup y \cup z$. From them I get a new vocabulary v' s.t. $\mathfrak{f}(\mathfrak{v}, x \cup y \cup z) = \mathfrak{v}'$. A higher level of abstraction. In this way, every statement the environment makes creates an abstraction layer. The outputs of the level below form the vocabulary of the level above. We go *up* a level of abstraction by looking at the 2nd order effects of the body we started with. An abstraction layer is like a smaller environment defined in the context of a larger environment. A 'small world' defined inside a 'big world'[322].

[322] L. J. Savage. *The Foundations of Statistics*. John Wiley & Sons, NY, USA, 1954; and Ramon Ferrer i Cancho and Ricard Solé. The small world of human language. *Proceedings of the Royal Society B: Biological Sciences*, 268(1482):2261–2265, 2001. DOI: 10.1098/rspb.2001.1800

CONCLUSION

The universe is a firehose of information. No system within the universe can fully understand the universe unless we start making assumptions about iterated function system fractals. That would be interesting, but it isn't what I'm doing here. The Bekenstein bound says bounded systems contain only finite information. A body is a bounded system, so a vocabularies are in general finite even if the universe isn't. An abstraction layer picks the truths a body cares about and ignores the rest. I see this as a form of relevance realisation enforced by physics[323]. A rock's vocabulary says things like "I'm here" or "I'm falling". A human is a sprawling mess from basics like run, grab and scream all the way through to divorce. These vocabularies are ontological rather than semantic. Concrete rather than abstract computations[324]. They are enacted by a particular body and interpreted by physics, rather than a person trying to reason out motives in the sense of Gricean meaning. Goertzel framed consciousness as a problem of moving from unary, to dyadic, to triadic relations[325]. A state is unary. A program is dyadic, in that it relates states to truth. By formalising an abstraction layer, we mimic the truth conditions of semantic structures using an ontological, unary foundation.

An embodied language is governed by the rules etched into reality's fabric. Each body has a formal grammar. At the core of this grammar lies the mutual exclusivity of states[326]. The logic of what can and cannot coexist. If I am omniscient then I can see the truth of every program unconstrained by any abstraction layer. Only one ϕ can hold sway at any moment in time, because time is difference. If two states could coexist then there would be programs which are both true and false. Hence, from an omniscient *objective* point of view things are true or false. *Subjectively* however, only the programs in one's abstraction layer are true or false. All others are unknowable, and so the world appears non-deterministic. Under-determined. This leaves room for certain notions of free will and compatibilism[327]. Conversely there can only be one state at a time *if* everything is to be only true or false. *Subjectively* we don't need to worry about states because we can only access the programs within the abstraction layer, and programs can be neither true or false. The point is that we have mutual exclusivity from an objective frame of reference, and this will give us the logical equivalent of `nand`. An aspect l is true only if its programs can share a state, meaning $\bigcap l \neq \emptyset$. If programs within an aspect can't coexist, then the aspect cannot exist.

[323] John Vervaeke, Timothy Lillicrap, and Blake Richards. Relevance realization and the emerging framework in cognitive science. *J. Log. Comput.*, 2012; John Vervaeke and Leonardo Ferraro. *Relevance, Meaning and the Cognitive Science of Wisdom*. Springer Netherlands, Dordrecht, 2013a; John Vervaeke and Leonardo Ferraro. Relevance realization and the neurodynamics and neuroconnectivity of general intelligence. In Inman Harvey, Ann Cavoukian, George Tomko, Don Borrett, Hon Kwan, and Dimitrios Hatzinakos, editors, *SmartData*, NY, 2013b. Springer Nature; and Johannes Jaeger, Anna Riedl, Alex Djedovic, John Vervaeke, and Denis Walsh. Naturalizing relevance realization: Why agency and cognition are fundamentally not computational. *Frontiers in Psychology*, 15, 2024

[324] Gualtiero Piccinini and Corey Maley. Computation in Physical Systems. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, Stanford, Sum. 21 edition, 2021

[325] Ben Goertzel. *The Hidden Pattern: A Patternist Philosophy of Mind*. Brown-Walker Press, USA, 2006

[326] Mutually exclusive within a 'world' or timeline.

[327] Kevin J. Mitchell. *Free Agents: How Evolution Gave Us Free Will*. Princeton University Press, Princeton, NJ, 2023. ISBN 9780691226231

VI. MASTER, WHAT IS MY PURPOSE?

This chapter is about purpose. It is based on the latter parts of my papers on abstraction layers[328], tasks[329] and consciousness[330]. What is normative? What *ought* to be? David Hume, fond of Guillotines[331], said one cannot smuggle an 'ought' out of an 'is.' This leaves me in a pickle. If I am to build a conscious machine, presumably it must have a moral compass. Where do we anchor its sense of "should"? I *could* argue it is anchored it in satisfying homeostatic and reproductive needs, but then where do they come from? I'm a naturalist, not a vitalist. I need something more fundamental than mere *life*. Besides I want to explain life, not assume it. Hence I'm going to argue there is no "is", only "ought". Some things exist. Others do not. Is this a normative judgement? I say it is. What else can it possibly be? Ought stems from change, and change is time. Not just ticking away like some bored clock, but calling the shots on what sticks around and what gets yeeted into the void. Creation and destruction. It is not just about *when* but *what lasts*. Time sifts the wheat from the chaff, and what hangs on gets the cosmic thumbs-up.

Many have sought to patch the gap between is and ought. Some say ought is a matter of feeling (puts the cart before the horse), or social contract (arguably come from feelings), or divine memo (god did it). I find these lacking. Change seems more foundational. Fundamental, if anything is. Without change or difference, everything would be the same thing. If everything is the same, can you really say there is anything? Is there an environment if there are no things? I say no. There would be nothing. Just an irreducible oneness. It is hard to conceive of it as an internally consistent idea. To comprehend it, must I cease to exist as an observer? Becoming one with everything is beyond the scope of my thesis. Difference or change must be fundamental to existence, because without change it seems inconceivable that anything exists. Time is just the passage of this change.

[328] Michael Timothy Bennett. Computational dualism and objective superintelligence. In *Artificial General Intelligence*. Springer Nature, 2024a; and Michael Timothy Bennett. Are biological systems more intelligent than artificial intelligence? 2025b. Forthcoming 2026

[329] Michael Timothy Bennett. The optimal choice of hypothesis is the weakest, not the shortest. In *Artificial General Intelligence*. Springer Nature, 2023a; and Michael Timothy Bennett. A formal theory of optimal learning with experimental results. *IJCAI*, 2025f

[330] Michael Timothy Bennett. Emergent causality and the foundation of consciousness. In *Artificial General Intelligence*. Springer Nature, 2023b; and Michael Timothy Bennett, Sean Welsh, and Anna Ciaunica. *Why Is Anything Conscious?* Preprint, accepted to and presented at ASSC27 and MoC5, 2024

[331] David Hume. *A Treatise of Human Nature*. 1739

Definition 4 (Time) *Time is the ordered sequence of transitions between distinct states of the environment, where each state $\phi \in \Phi$ is a full snapshot of reality at a given tick.*

Time is the process of becoming[332]. Every tick of the cosmic clock is creation and destruction. Some aspects of the environment persist through many ticks of the clock.

[332] Alfred North Whitehead. *Process and Reality*. 1929

Definition 5 (Persistence) *An aspect l persists across time if there's a sequence of states $\phi_1, \phi_2, \ldots, \phi_n$ where each ϕ_i has a statement in l's extension E_l that's expressed.*

Persistence is survival. Darwin's natural selection[333] on a universal scale. Stable atoms stick around because they vibe with physics[334]; critters adapt or get fossilized[335]. The universe is like a bouncer. Fit the rhythm and you stay. Clash with it and you're out. New things are occasionally allowed in. This is the first whisper of "ought." What persists is what is *meant* to, by the rules of the game.

[333] Charles Darwin. *On the Origin of Species*. 1859

[334] Ilya Prigogine. *From Being to Becoming: Time and Complexity in the Physical Sciences*. W.H. Freeman, 1980

[335] John Maynard Smith. *Evolution and the Theory of Games*. Cambridge University Press, 1982

THE ENVIRONMENT HAS AN OPINION

A STATE EXPRESSES SOME ASPECTS, but not others. From the definition of environment, a statement s is expressed, realised or embodied by state ϕ if all its programs are true in ϕ, i.e., $\phi \in \bigcap s$. Something what *ought* to be. The environment, churning through time, picks winners and losers[336]. What sticks is the universe's way of saying "I like this". A sturdy molecule or a sneaky predator. The "ought not" pile is everything which doesn't exist. Persistence over time sets the baseline for normativity[337].

[336] Stuart A. Kauffman. *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, 1993

[337] Daniel C. Dennett. *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. Simon & Schuster, 1995

THESE STATEMENTS FORM ABSTRACTION LAYERS. Abstraction layers stack up like Matryoshka dolls, each layer refining the cosmic "ought" into sharper rules. From "thou shalt exist" at the base, we climb to "thou shalt compute efficiently" or "thou shalt not crash the system"[338]. Time, persistence, and expression give us this natural "ought". Just the universe doing its thing[339]. For my conscious machine, this is the foundation.

[338] Seth Lloyd. Ultimate physical limits to computation. *Nature*, 406(6799): 1047–1054, 2000

[339] David Deutsch. *The Fabric of Reality: The Science of Parallel Universes–and Its Implications*. Penguin Books, 1997

PURPOSE

THE FACT OF EXISTENCE IS A VALUE JUDGEMENT. Some things exist, and others do not. A rock doesn't need to know physics to fall, but in doing so constrains what can happen next. It is an embodied *ought* that constrains what *is*, has been or ever will be. In this sense, every body is chattering away in a language forged by its form. When the state changes, some statements persist, and others are destroyed. This creates an incentive. The universe preserves that which preserves itself. Change is fundamental, and by its very nature change optimises for systems that cope with change, by deleting those that cannot. I want to formalise intelligence, which means I want to formalise a system that preserves itself. A living system.

A LIVING, SELF PRESERVING SYSTEM is a statement l made by the environment, and an abstraction layer. A self preserving system differs from other systems in that it exerts influence on the surrounding environment in order to preserve its own existence. If l is a living organism, then some possible worlds end up with an organism dead or failing to reproduce. Not fit. There are more constraints on an organism's possible worlds than just its body. It is embedded in the environment. The state of an organism's nervous system is a statement in its embodied formal language. There is context to consider. Not all of the worlds in E_l are compatible with that context. Where another system would sit passively awaiting its fate, a self preserving system expresses additional statements to preserve its existence, actively imposing constraints on its own extension.

THEREIN LIES THE RUB. Not everything serves homeostatic and reproductive goals. Not every possible world is a winner. Intelligent systems discriminate. Abstraction layers are biased toward some goals over others, but how exactly is that supposed to work? To really describe intelligence I need to formalise the idea of goals, but not in the abstract sense we humans are accustomed to. I can't have goals separate from the systems that pursue them or I'm just going to end up with computational dualism again[340]. I need integrate goals with embodiment. A goal together with context and instructions is commonly known as a task. A task is what I use to formalise enactive cognition, in what I call **Pancomputational Enactivism**[341,342]. A task is a formal description of a system in terms of its behaviour. That system is an abstraction layer, and the task is what it expresses. Outputs O in the context of inputs I. Typical computer science fare. I can choose any statement the body can make and call it an **input**.

[340] Michael Timothy Bennett. Lies, damned lies, and the orthogonality thesis. *Preprint*, 2025d

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

[342] Stack Theory is the idea that everything is an infinite state of abstraction layers. Pancomputational Enactivism is the formalisation of enactivism within Stack Theory.

The possible **outputs** are the extension E_I of the inputs I. This makes sense because we have only so many possible worlds given the inputs. However not all the possible worlds are desirable[343], so the O is a subset of E_I. This pairs inputs with the **correct** outputs. A body can be seen as a functional, computational system that maps inputs to outputs. Intuitively, these are the outputs that keep you breathing instead of bleeding out in a ditch.

[343] For the purpose of defining intelligence, we need some notion of value. I'll get to where this comes from in the next section.

Definition 6 (𝔳-task)
For a chosen $\mathfrak{v}$, a task α is a pair $\langle I_\alpha, O_\alpha \rangle$ where[344]*:*

- *$I_\alpha \subset L_\mathfrak{v}$ 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 "lattice" or generational hierarchy of tasks. Formally, 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 always "lower level" than its parents*[345]*.*

[344] (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 [see def. of 𝔳-task].
- A **correct output** is a correct completion of an input.

TASKS ARE LIKE MATRYOSHKA DOLLS. Little ones fit inside bigger ones. For example not choking on your coffee fits inside surviving the day. It's a hierarchy. So how does your body pick the right output? Every statement your body makes constrains what can happen next. A policy is just a statement that constrains your outputs. A **correct policy** constrains you to **correct outputs**, given the additional constraint of the inputs. Correct policies keep you from face-planting. They steer you toward the outputs that don't end in a Darwin Award. It works thusly:

Definition 7 (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.*
- *The set of all correct policies for a task α is denoted Π_α.*[346]

Assume $\mathfrak{v}$-task ω and a policy $\pi \in L_\mathfrak{v}$. Inference[347] *proceeds as follows:*

1. *we are presented with an input $i \in I_\omega$, and*
2. *we must select an output $e \in E_i \cap E_\pi$.*

[345] (FURTHER INTUITIVE SUMMARY) A 𝔳-task is a formal, **behavioural** description of an aspect of the environment. For example, a self-organising biological system could be described as a task α enumerating all behaviour in which it remains alive. It begins alive in circumstances given by inputs I_α, and remains alive in circumstances given by outputs O_α, and is dead in circumstances given by $E_{I_\alpha} - O_\alpha$. Likewise, we could describe the game chess played from the perspective of white. We could say Φ contains a state corresponding to each and every move of each and every possible game of chess, I_α contains every possible sequence of moves in which the game has not ended and it remains *possible* for white to win, and O_α contains every possible sequence ending in a move that means white *has* won. Tasks are *behavioural* descriptions of systems in the philosophical sense of the word, and we will next relate these ideas to machine functionalism.

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

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

[347] (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.

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).*

MIND AND BODY ARE INTIMATELY CONNECTED[348]. But flesh or steel, the same constraints can be realised by wildly different systems[349]. Cellular automata show how simple rules birth complex life[350]. Reinforcement learning is basically evolution with better PR[351]. Am I saying bodies are just computers? No. Your mind is etched in meat, not silicon. Still, the vibes are the same. Inputs, outputs and constraints. The means of computation are less important than the resulting constraints. In that sense the ambulatory meme was correct when he said "wow, everything is computer". Bodies are computational systems, tasks define the goals, and policies enforce the wins. It is a framework that scales from slime mould to Silicon Valley.

[348] M. Wilson. Six views of embodied cognition. *Psychonomic Bulletin & Review*, 9(4):625–636, 2002

[349] Ricard Solé et al. Fundamental constraints to the logic of living systems. *Interface Focus*, 2024

[350] S. Wolfram. *A new kind of science*. Wolfram Media, 2002

[351] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, MA, 2018

LEARNING THE STACK

TIME DOES A FROGMARCH. Each step deletes something. Systems that stick around are those that avoid the jackboots of extinction. I'll call this 'fit', but it is broader than the Darwinian notion[352]. It applies even to non-living systems. Now, being 'fit' in this sense is hard-wired. Imposed from outside. Extrinsic. It does not require intelligence or agency. The jackbooted universe shapes matter into a form which is 'fit' by just deleting everything else. In that sense everything is an adaptation, which is a bit unsatisfying. This relentless change is an optimiser. Aspects "adapt" to whatever it is the environment has decided to optimise for. The extinsic *what* that the environment optimises for is an **uninstantiated task**.

[352] Charles Darwin. *On the Origin of Species*. 1859

Definition 8 (λ-tasks)
*The set of all tasks with no abstraction (meaning $\mathfrak{v} = P$) is Γ_P (it contains every task in every vocabulary). For every P-task $\rho \in \Gamma_P$ there exists a function $\lambda_\rho : 2^P \to \Gamma_P$ that takes a vocabulary $\mathfrak{v}' \in 2^P$ and returns a highest level child $\omega \sqsubset \rho$ which is also a $\mathfrak{v}'$-task. We call λ_ρ an **uninstantiated-task**, and $\lambda_1 \sqsubset \lambda_2$ iff $\lambda_1(P) \sqsubset \lambda_2(P)$.*

THIS DEFINES EXTRINSIC, EXTERNALLY IMPOSED PURPOSE. It lets me consider purpose without pinning it to one vocabulary, so that we might compare embodiments. A $\mathfrak{v}$-task does a good job describing hard-wired behaviour. For example, a simple reflex agent that responds to the world around it predictably, with preordained responses. Every behaviour is an adaptation baked into such organisms from birth. They cannot acquire new adaptations over the course of their lifetimes. If hard-wired adaptations are long term adaptations, then short term adaptations are those an organism *learns*. I mean adaptations acquired *during* a system's existence, rather than baked into it from the start. Learning is an adaptation that facilitates adaptation. The ability to learn must be hard-wired into a system from its inception, but once a system *can* learn it can acquire new adaptations. A system which cannot learn has to store all of its policies from birth. That is inefficient, and limits how many tasks the system can complete. The alternative is an adaptation that allows a system to acquire new adaptations. To record and retrieve information to help them persist[353]. A rock might "store" information about its past by having chunks knocked off it, but it does not then use that information to maintain its form. A rock does not pursue *homeostasis*. A living system maintains homeostasis. This means it optimises its internal and external world to maintain its form. Its integrity. Maintaining homeostasis is a very basic form of task. This seems to be where 'learning'

[353] Ricard Solé et al. Fundamental constraints to the logic of living systems. *Interface Focus*, 2024

begins. Learning requires a goal. We need something that defines 'correct' before a system can optimise for what is correct. Correct in general can be unrelated to homeostasis, but as I am trying to work from first principles I need to explain how we get to 'correct' in at least one case. Evolutionarily speaking that this is where we get a basic *ought* for the purpose of learning. A living system like a human can then build a computer that optimises for any arbitrary notion of *correct*[354]. A body embedded and extending into its environment is an abstraction layer. It can be constrained to 'correct' behaviour by expressing a policy π that constrains it to desirable possible worlds. Worlds in which conform to homeostatic goals. Such a system ceases to exist in other worlds[355]. A system *learns* by expressing a policy that constrains it to some arbitrary notion of *correct* (homeostatic or otherwise). A system which stores and retrieves information like a computer can express a policy by changing its internal state, and so can complete a far wider range of tasks than a system which cannot learn[356].

[354] Within the bounds of what we can conceive of, as systems whose nature it is to maintain homeostasis.

[355] If the system did not maintain homeostasis, then it is dead.

[356] All else being equal.

Definition 9 (learning) *Learning is a collection of definitions that describe the process by which a policy is constructed by any system*[357].

- *A **proxy** $<$ is a binary relation on statements, and the set of all proxies is Q.*
- *$<_w$ is the **weakness** proxy*[358]*. 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*[359]*. We have $l_1 <_d l_2$ iff $|l_1| > |l_2|$.*

[357] (INTUITIVE SUMMARY) Learning is an activity undertaken by an adaptive system, and a task has been **learned** by a system that embodies a correct policy. Humans typically learn from **examples**. An example of a task is a correct output and input.

[358] By the weakness of a statement, we mean the cardinality of its extension. By the weakness of an extension we mean its cardinality.

[359] When we speak of simplicity with regards to a policy $\pi \in \Pi_\alpha$ we mean the cardinality of the smallest correct policy $\pi' \in \Pi_\alpha$ s.t. $E_{\pi'} = E_\pi$. The complexity of an extension is the **simplest** statement of which it is an extension.

(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.*

(EFFICIENCY) *Suppose*[360] $\mathfrak{app}$ *is the set of **a**ll **p**airs of **p**olicies. Assume a proxy $<$ returns 1 iff true, else 0. Proxy $<_a$ is more efficient than $<_b$ iff*

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

[360] (FURTHER INTUITIVE SUMMARY) 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.

(OPTIMAL PROXY) *There is no proxy more efficient than weakness. The weakness proxy formalises the idea that "explanations should be no more specific than necessary" (see Bennett's razor in this ref*[361]*).*

(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 (it chooses policies that complete a wider variety of tasks, and thus are more sample and energy efficient choices), the more efficient the proxy. The most efficient proxy is weakness (see proofs 1 and 2, or these refs*[362]*), which is why we're using it here.*

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

[362] Michael Timothy Bennett. The optimal choice of hypothesis is the weakest, not the shortest. In *Artificial General Intelligence*. Springer Nature, 2023a; and Michael Timothy Bennett. Are biological systems more intelligent than artificial intelligence? 2025b. Forthcoming 2026

TO LEARN, ONE MUST EXPRESSING A POLICY π that constrains future behaviour to desirable worlds. That *generalises* to future instances of a problem. If tasks are uniformly distributed, then the most effective way to learn is to maximise the number of tasks π completes. A proxy is a means of choosing between correct policies, in the hopes of selecting a policy that generalises to the relevant parent tasks. I give two proxies above, but there are others. In the next chapter I will explain why weakness is the optimal proxy, but for now don't worry about that. It seems eminently reasonable to assume tasks are uniformly distributed. Anything else is an unnecessary assumption. A normative judgement beyond what is required. As we have already covered at length, the very fact of existence is a matter of *ought*. It is a sort of *existential* normativity. If a body exists and can store and retrieve information, then its representational capabilities are the product of that existential normativity. I'm not saying it isn't theoretically conceivable some Lovecraftian horror reaches into an environment and changes the rules at a whim, but such a thing is already baked into this existential normativity. For the purpose of deciding which policies are optimal in general, it makes no difference why the state changes. It simply does.

FINALLY, I'LL INTEGRATE THIS IDEA OF A TASK with The Stack I've been talking about. To do this, I'll introduce the multilayer architecture (MLA). An abstractor function $\mathfrak{f}$ is applied here to *policies*, rather than just any statement.

Definition 10 (multilayer architecture) *The multilayer architecture (MLA) found in both biological systems and computers. It integrates a stack with tasks to represent natural selection or 'correctness' at different layers of abstraction.*

- *The **stack** is represented here by a sequence of uninstantiated tasks* $\langle \lambda^0, \lambda^1 ... \lambda^n \rangle$ *s.t* $\lambda^{i+1} \sqsubset \lambda^i$.
- $\mathfrak{f}$ *is an **abstractor** function.*
- *The **state** of the MLA is a sequence of policies* $\langle \pi^0, \pi^1 ... \pi^n \rangle$ *and a sequence of vocabularies* $\langle \mathfrak{v}^0, \mathfrak{v}^1 ... \mathfrak{v}^n \rangle$ *such that* $\mathfrak{v}^{i+1} = \mathfrak{f}(\mathfrak{v}^i, \pi^i)$ *and* $\pi^i \in \Pi_{\lambda^i(\mathfrak{v}^i)}$.

In the absence of abstraction where the system is seen as nothing more than the sum of its parts, the MLA is just a task $\lambda^0(\mathfrak{v}^0)$*, allowing us to look at the system across scales of distribution. We say the MLA is **over-constrained** when there exists* $i < n$ *s.t. and* $\Pi_{\lambda^i(\mathfrak{v}^i)} = \emptyset$*, and **multilayer-causal-learning** (MCL) occurs when the MLA is not over-constrained and the proxy for learning is weakness. Note that the vocabulary is different at each level in the stack, which means each has its own generational hierarchy of tasks. By a higher level of abstraction, we mean a task higher in the stack (later in the causal chain).*

Now, I won't actually use this definition for a couple of chapters, but it is important to introduce it here for intuition. After all, I have been speaking endlessly about abstraction layers, and it would be strange to introduce goal directed behaviour *within* an abstraction layer without explaining how that goal directed behaviour *propagates* up and down the stack. As you can see in the above definition, the process of applying the abstractor function to obtain second, third or higher orders of behavioural effect is the same. However because it is applied to a policy, this is useful for formalising biological and other distributed goal directed systems. For example, say α_1 and α_2 are tasks representing the behaviour of two cells. Imagine those cells are both child tasks of α, which is an organ. A collective identity[363]. They exist in the same abstraction layer. α is like looking at an organ as a collection of cells. However, if we move up a level of abstraction by taking the policy of α and applying the abstractor function, we are now looking at the organ, because 'organ' is a property we ascribe to the behaviour of cells. It is a second order of abstracted behaviour. I will discuss more about how this works in later chapters. For now, it is only important to note that $\mathfrak{v}$-tasks are arranged in a stack, like abstraction layers. As we look higher in the stack, the goal directed behaviour gets narrower and more specific, because each successive layer is an effect of the goal directed behaviour in the layers below.

[363] Patrick McMillen and Michael Levin. Collective intelligence: A unifying concept for integrating biology across scales and substrates. *Communications Biology*, 2024

VII. WEAK

Human intelligence. It is sometimes hailed as the crowning achievement of evolution. But in the Darwinian arena, it's not about being the smartest. It is about surviving long enough to pass on your genes. From a Darwinian perspective, intelligence is long-term adaptation that facilitates short-term adaptation during an organism's lifetime. Long-term adaptation is like the genes we inherit. Short-term is what we learn as we go. If evolution is an optimiser, then biological intelligence is a mesa-optimiser (an optimiser within an optimiser)[364]. Without intelligence, a system would need all its knowledge pre-programmed. Like a robot with a fixed set of instructions. With intelligence a system can learn, adapt, and survive in a wider range of circumstances. It can complete a wider range of tasks[365]. A machine with intelligence can learn from its environment, adapt to new situations, and potentially develop something akin to consciousness. It's not just about processing power. It is about the ability to change. Intelligence is the key to unlocking the door to consciousness. Without it, we're just building a glorified abacus. This chapter is based primarily on my paper on weak versus simple hypotheses[366], with some minor updates from later works[367]. I'll show how adaptability is maximised by using weakness as the proxy, and I propose the following epistemological razor:

> "Explanations should be no more specific than necessary."[368]

[364] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems, 2021

[365] All else being equal, of course.

[366] Michael Timothy Bennett. The optimal choice of hypothesis is the weakest, not the shortest. In *Artificial General Intelligence*. Springer Nature, 2023a; and Michael Timothy Bennett. A formal theory of optimal learning with experimental results. *IJCAI*, 2025f

[367] Michael Timothy Bennett. Optimal policy is weakest policy. *Artificial General Intelligence*, 2025e; and Michael Timothy Bennett. A formal theory of optimal learning with experimental results. *IJCAI*, 2025f

[368] In the publication I originally proposed this, I named it Bennett's Razor. Once published, there are no backsies.

INTELLIGENCE IN STACKISM

Intelligence is not about what something *is*, but what it *does*. I've framed everything as a stack of abstraction layers, each enacted by the one beneath. Like software on hardware, the layers go down. This is as true of human language as it is of human software. The cosmic *ought* determines functionality. Affordances[369]. It is why we have one abstraction layer instead of another. Intelligence isn't about what an organism *is* but what it *does*. **ʊ-tasks**. Subjected to inputs, a system produces outputs. Intelligence *affords* adaptation.

[369] James J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979

Tasks are interconnected, forming a **generational hierarchy** that reflects an organism's temporal existence. Within this lattice, **child tasks** are specific cases of the broader **parent tasks** encompassing them. An organism's past decisions is a child task of the task that includes every decision an organism might make over the course of its existence. The lattice structure links past and future behaviour. An example of a child task might be "take the succession of turns leading from home to office" and its parent could be "navigate the environment". An organism's survival depends on generalising from completed child tasks to meet the demands of the broader parent tasks that lie ahead.[370] Thus, the generational hierarchy provides a dynamic framework for understanding intelligence as a process of bridging temporal scales.

[370] Adaptation requires flexibility—past success does not guarantee future survival.

POLICIES AND ADAPTATION

In the context of intelligence and adaptation, policies serve as the mechanisms that guide an organism's behaviour towards survival. Policies can be innate or hard-wired, but that is less adaptable than being able to learn new through interaction with the environment. A learned **policy** is a learned statement or constraint that an organism embodies to ensure fit behaviour. It acts as a rulebook for survival, dictating how to respond to various inputs to achieve desirable outcomes. Assuming an organism remains alive and in some condition to procreate, its past behaviour is an ostensive definition of fit behaviour. Past behaviour is likely to imply at least some fit policies, and some unfit ones. The challenge of learning is to discern which policies will reliably constrain the organism to fit phenotypes. That complete the tasks of the past, and the widest range of possible future tasks[371].

[371] An organism that relies on overly specific policies may struggle to adapt to new environments.

THE WEAK SHALL INHERIT THE WORK

A WEAKER POLICY IS ONE THAT PERMITS more possible behaviours while still being fit, akin to a versatile tool. Their lack of specificity allows them to address a wide range of future scenarios, enhancing adaptation. Weak policies are the key to generalisation. I'll prove it. Recall from the definition of learning:

- A **proxy** $<$ is a binary relation on statements, and the set of all proxies is Q.
- $<_w$ is the **weakness** proxy[372]. 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[373]. We have $l_1 <_d l_2$ iff $|l_1| > |l_2|$.

Now consider the follow two proofs:

Theorem 1 (sufficiency)
Assume $\alpha \sqsubset \omega$. The weakness proxy sufficient to maximise the probability that a parent ω is learned from a child α[374].

Proof 1 *You're given the definition of $\mathfrak{v}$-task α from which you infer a hypothesis $\pi \in \Pi_\alpha$. To learn ω, you need $\pi \in \Pi_\omega$:*

1. *For every $\pi \in \Pi_\alpha$ there exists a $\mathfrak{v}$-task $\gamma_\pi \in \Gamma_\mathfrak{v}$ s.t. $O_{\gamma_\pi} = E_\pi$, meaning π permits only correct outputs for that task regardless of input. We'll call the highest level task γ_π s.t. $O_{\gamma_\pi} = E_\pi$ the **policy task** of π.*
2. *ω is either the policy task of a policy in Π_α, or a child thereof[375].*
3. *If a policy π is correct for a parent of ω, then it is also correct for ω. Hence we should choose π that has a policy task with the largest number of children. As tasks are uniformly distributed, that will maximise the probability that ω is γ_π or a child thereof.*
4. *For the purpose of this proof, we say one task is **equivalent**[376] to another if it has the same correct outputs.*
5. *No two policies in Π_α have the same policy task[377]. This is because all the policies in Π_α are derived from the same set inputs, I_α.*
6. *The set of statements which might be outputs addressing inputs in I_ω and not I_α, is $\overline{E_{I_\alpha}} = \{l \in L_\mathfrak{v} : l \notin E_{I_\alpha}\}$[378].*
7. *For any given $\pi \in \Pi_\alpha$, the extension E_π of π is the set of outputs π implies. The subset of E_π which fall outside the scope of what is required for the known task α is $\overline{E_{I_\alpha}} \cap E_\pi$[379].*
8. *$L_\mathfrak{v} = \overline{E_{I_\alpha}} \cup E_{I_\alpha}$ and for all $\pi \in \Pi_\alpha$, $E_\pi \subset L_\mathfrak{v}$. Apart from the inputs and correct outputs of α, E_{I_α} contains only outputs which would be incorrect according to both α and ω. Put another way, $E_{I_\alpha} \cap E_\pi = O_\alpha$ for every possible choice of π in Π_α. Hence the only way $|E_\pi|$ can increase is if $\left|\overline{E_{I_\alpha}} \cap E_\pi\right|$ increases. It follows that $\left|\overline{E_{I_\alpha}} \cap E_\pi\right|$ increases with $|E_\pi|$.*

[372] By the weakness of a statement, we mean the cardinality of its extension. By the weakness of an extension we mean its cardinality.
[373] When we speak of simplicity with regards to a policy $\pi \in \Pi_\alpha$ we mean the cardinality of the smallest correct policy $\pi' \in \Pi_\alpha$ s.t. $E_{\pi'} = E_\pi$. The complexity of an extension is the **simplest** statement of which it is an extension.
[374] Assume there exist correct policies for ω, because otherwise there would be no point in trying to learn it.

[375] I'd like to give credit here to Nora Belrose for pointing out an error. Nora pointed out I was miscounting the number of tasks. As a result I realised I was not counting tasks, I was in fact counting policy tasks and had entirely neglected to mention this fact. This was a significant error which has now been corrected, with several additional steps added to account for equivalence.
[376] This is because switching from β to ζ s.t. $I_\beta \neq I_\zeta$ and $O_\beta = O_\zeta$ would be to pursue the same goal in different circumstances. This is because inputs are *subsets* of outputs, so both sets of inputs are implied by the outputs. O_ζ implies I_β and O_β implies I_ζ
[377] Every policy task for policies of α is non-equivalent from the others.
[378] This is because E_{I_α} contains every statement which is a correct output or an incorrect output, and $\overline{E_{I_\alpha}}$ contains every statement which could possibly be in I_ω, E_{I_ω} and thus O_ω.
[379] This is because E_{I_α} is the set of all conceivable outputs by which one might attempt to complete α, and so the set of all outputs that can't be made when undertaking α is $\overline{E_{I_\alpha}}$ because those outputs occur given inputs that aren't part of I_α.

9. $2^{|\overline{E_{I_\alpha}} \cap E_\pi|}$ *is the number of non-equivalent* ***parents*** *of* α *to which* π *generalises. It increases monotonically with the weakness of* π.

10. *Given* $\mathfrak{v}$*-tasks are uniformly distributed and* $\Pi_\alpha \cap \Pi_\omega \neq \emptyset$*, the probability that* $\pi \in \Pi_\alpha$ *generalises to* ω *is*

$$p(\pi \in \Pi_\omega \mid \pi \in \Pi_\alpha, \alpha \sqsubset \omega) = \frac{2^{|\overline{E_{I_\alpha}} \cap E_\pi|}}{2^{|\overline{E_{I_\alpha}}|}}$$

$p(\pi \in \Pi_\omega \mid \pi \in \Pi_\alpha, \alpha \sqsubset \omega)$ is maximised when $|E_\pi|$ is maximised. Recall from definition 4 that $<_w$ is the ***weakness*** *proxy. For statements l_1, l_2 we have $l_1 <_w l_2$ iff $|E_{l_1}| < |E_{l_2}|$. π that maximises $<_w$ will also maximise $p(\pi \in \Pi_\omega \mid \pi \in \Pi_\alpha, \alpha \sqsubset \omega)$. Hence the weakness proxy maximises the probability that*[380] *a parent ω is learned from a child α.* □

[380] Subsequently it also maximises the sample efficiency with which a parent ω is learned from a child α.

Theorem 2 (necessity)
To maximise the probability of learning ω from α, it is necessary to use weakness as a proxy.

Proof 2 *Let α and ω be defined exactly as they were in proof 1.*

1. *If $\pi \in \Pi_\alpha$ and $E_{I_\omega} \cap E_\pi = O_\omega$, then it must be he case that $O_\omega \subseteq E_\pi$.*

2. *If $|E_\pi| < |O_\omega|$ then generalisation cannot occur, because that would mean that $O_\omega \not\subseteq E_\pi$.*

3. *Therefore generalisation is only possible if $|E_\pi| \geq |O_\omega|$, meaning a sufficiently weak hypothesis is necessary to generalise from child to parent.*

4. *For any two hypotheses π_1 and π_2, if $|\ E_{\pi_1}\ | < |\ E_{\pi_2}\ |$ then the probability $p(|E_{\pi_1}| \geq |O_\omega|) < p(|E_{\pi_2}| \geq |O_\omega|)$ because tasks are uniformly distributed.*

5. *Hence the probability that $|E_m| \geq |O_\omega|$ is maximised when $|E_m|$ is maximised. To maximise the probability of learning ω from α, it is necessary to select the weakest hypothesis.*

To select the weakest hypothesis, it is necessary to use the weakness proxy. □

Weak policies are thus essential, boosting both the sample and energy efficiency of adaptation.

META-APPROACHES

Ockham's Razor favors simplicity. Earlier I described AGI in terms of tools (search and approximation) and meta-approaches (scale-maxing, simp-maxing and w-maxing). Ockham's Razor is an example of simp-maxing (simplicity maximisation). My razor prioritises generality, advocating for broad yet effective explanations[381]. It promotes **weak policies**, enhancing adaptability by avoiding overspecification, thus improving the speed and efficiency of adaptation to varied scenarios. I call this w-maxing. Rules and explanations should remain as general as possible while fulfilling their purpose[382]. So to reiterate we now have three meta-approaches: simp-maxing, scale-maxing and w-maxing. W-maxing is the meta-approach I propose, achieved in part by using weakness as the proxy. Of course, how much you can w-max depends on the abstraction layer. Simp-maxing is likewise formalised using a proxy. Finally, scale maxing involves maximising the resources available, which formally would involve just increasing the size of the vocabulary.

[381] Ockham's Razor: "Don't multiply entities unnecessarily"; Bennett's Razor: "Don't constrain unnecessarily."

[382] Michael Timothy Bennett. The optimal choice of hypothesis is the weakest, not the shortest. In *Artificial General Intelligence*. Springer Nature, 2023a; and Michael Timothy Bennett. A formal theory of optimal learning with experimental results. *IJCAI*, 2025f

INTUITION

For intuition on w-maxing, consider The Contravariance Principle[383]. The contra-variance principle says that we are more likely to converge on the *true* underlying model the more we scale up the data and subsequently the apparent "difficulty" of a task (here in the informal sense). If we keep resources finite, then there are fewer and fewer possible explanations of the given data as we scale up difficulty. Scale up difficulty enough, and there will only be one possible explanation. This is useful because it tells us we can scale up data to get models to converge to the true process that generated the data, assuming of course that it can be represented given the system[384]. This contravariance principle appears to be like The Inventor's Paradox, which says it can be easier to solve a more general problem that includes A, than a A alone. A solution to a more general problem is, of course, weaker. It completes a wider range of tasks. I can formalise the contravariance principle using tasks. In a proof earlier in this chapter I talk about a **policy task**. For every $\pi \in \Pi_\alpha$ there exists a $\mathfrak{v}$-task $\gamma_\pi \in \Gamma_\mathfrak{v}$ s.t. $O_{\gamma_\pi} = E_\pi$, meaning π permits only correct outputs for that task regardless of input. The highest level task γ_π s.t. $O_{\gamma_\pi} = E_\pi$ it the **policy task** of π. Now, if I increase the "difficulty" of the task by scaling up the data until there is only one solution, what I have done is construct a policy task for a weakest policy in the set. The contravariance principle amounts to the idea that one should scale up data until the weakest policy is the only one left. Now, given my result, we can skip the bit where we scale up the data and just choose a weakest policy. Same result, less work involved.

[383] Rosa Cao and Daniel Yamins. Explanatory models in neuroscience, part 2: Functional intelligibility and the contravariance principle. *Cognitive Systems Research*, 85:101200, 2024

[384] Which is very likely not to be the case unless the system's designer is well informed.

Before I move on, some have questioned whether my framework struggles with multiclass classification[385]. It's fine, no need to worry. Yes, a task frames problems as one-class classification. However, two-class classification problems are a subset of one-class. As Simmons points out in his reply to one of my papers, you can just include an extra declarative program to discriminate between classes[386]. Say we have a two class classification problem α. It is in reality a parent task of two one-class classification tasks α_1 and α_2. $O_{\alpha_1} \cup O_{\alpha_2} = O_\alpha$. O_{α_1} is one correct class classification for outputs, and O_{α_2} is the other. Any correct policy for α much correctly classify both α_1 and α_2. If we want to deal with continuous values, we just have an infinite class classification problem. Of course, no physical system *actually* implements a continuous value because that would require infinitely accurate `read/write` memory, and if we had those we could violate the Bekenstein bound[387] and store infinite infor-

[385] Gabriel Simmons. Comment on is complexity an illusion?, 2024. URL `https://arxiv.org/abs/2411.08897`

[386] Gabriel Simmons. Comment on is complexity an illusion?, 2024. URL `https://arxiv.org/abs/2411.08897`

[387] Jacob D. Bekenstein. Universal upper bound on the entropy-to-energy ratio for bounded systems. *Phys. Rev. D*, 23: 287–298, Jan 1981

mation. But I digress. The important thing is my simple one-class classification set up can address any sort of task. Everything, in the end, amounts to an instance of one-class classification. By adding more details (subclasses etc) we just make learning and inference easier. Consider a task with k classes, where each class offers distinct output constraints. The proof below formally shows that increasing the number of classes reduces the complexity of learning.

SIMP-MAXING WILL NOT SAVE YOU

HERE I COMPARE W-MAXING AND SIMP-MAXING. I've a simple proof and some experiments that show the former destroys the latter, based on my paper on weak vs simple hypotheses[388]. Simp-maxing is a matter of form. Whether one thing has a shorter minimum description length than another depends on the language. It is like a ruler that changes length. W-maxing is about function. How loose is your policy's leash. How many tasks it completes. Intuitively it is a bit like the difference between a Swiss Army knife and a scalpel. One is simple, sure, but the other is ready for anything. From the previous chapter we already know that w-maxing is necessary and sufficient to maximise adaptability. What we don't know yet is whether we still need to simp-max, and to what extent there's a difference. When I initially concieved of this framework, I actually thought that w-maxing would amount to simp-maxing. I was astounded when my experiments showed w-maxing far outperforming simp-maxing. So here I'll prove you don't need to simp for simplicity to win. I'm not saying it won't help if you can't w-max, but w-maxing is all that's necessary for adaptation.

[388] Michael Timothy Bennett. The optimal choice of hypothesis is the weakest, not the shortest. In *Artificial General Intelligence*. Springer Nature, 2023a; and Michael Timothy Bennett. A formal theory of optimal learning with experimental results. *IJCAI*, 2025f

UPPER BOUND

SIMP-MAXERS LIKE AIXI[389] hinge on complexity. They argue intelligence boils down to finding the simplest model that fits the data. As has already been pointed out, simplicity hinges on the choice of Universal Turing Machine (UTM)[390]. Pick the wrong one and the theoretical superintelligent god is reduced to a bag of hammers. Even so, it helps to be formal. How does this work in my formalism?

[389] Marcus Hutter. *Universal Algorithmic Intelligence: A Mathematical Top→Down Approach*, pages 227–290. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007

[390] Jan Leike and Marcus Hutter. Bad universal priors and notions of optimality. *Proceedings of The 28th Conference on Learning Theory, in Proceedings of Machine Learning Research*, pages 1244–1259, 2015; and Laurent Orseau. Asymptotic non-learnability of universal agents with neural networks. In Joscha Bach, Ben Goertzel, and Matthew Iklé, editors, *Artificial General Intelligence: 5th International Conference, AGI 2012*, pages 234–243, Berlin, Heidelberg, 2012. Springer Nature

Theorem 3 (simplicity sub-optimality)
Description length is neither a necessary nor sufficient proxy for the purposes of maximising the probability that induction generalises.

Proof 3 *In proofs 1 and 2 we proved that weakness is a necessary and sufficient choice of proxy to maximise the probability of generalisation. It follows that either maximising* $\frac{1}{|m|}$ *(minimising description length) maximises* $|E_m|$ *(weakness), or minimisation of description length is unnecessary to maximise the probability of generalisation. Assume the former, and we'll construct a counterexample with* $\mathfrak{v} = \{a,b,c,d,e,f,g,h,j,k,z\}$ *s.t.* $L_{\mathfrak{v}} = \{\{a,b,c,d,j,k,z\},\{e,b,c,d,k\},\{a,f,c,d,j\},\{e,b,g,d,j,k,z\},\{a,f,c,h,j,k\},\{e,f,g,h,j,k\}\}$ *and a task* α *where*

- $I_\alpha = \{\{a,b\},\{e,b\}\}$
- $O_\alpha = \{\{a,b,c,d,j,k,z\},\{e,b,g,d,j,k,z\}\}$
- $\Pi_\alpha = \{\{z\},\{j,k\}\}$

Weakness as a proxy selects $\{j,k\}$*, while description length as a proxy selects* $\{z\}$*. This demonstrates the minimising description length does not necessarily maximise weakness, and maximising weakness does not minimise description length. As weakness is necessary and sufficient to maximise the probability of generalisation, it follows that minimising description length is neither.* □

AS YOU CAN SEE, IT IS POSSIBLE to w-max without simp-maxing. In fact, they can be at odds. Given w-maxing is necessary and sufficient for generalisation, this bodes poorly for simp-maxing. It also raises questions about why simp-maxing works at all which I'll get into in later chapters. That later chapter is kinder to the idea. For now, I can use this fact to establish an alternative to AIXI as an *upper bound* on intelligence. Arguably AIXI's greatest contribution is to provide an ideal for which we can optimise. Even if it is based on a flawed premise, it is much better than nothing. This idea is an upper bound on intelligent behaviour. The flaw is that it is an upper bound on *disembodied*, software intelligence. The alternative is to establish an

upper bound on *embodied* intelligence. Now it is obvious from the earlier proofs that the upper bound on intelligent behaviour within the confines of an abstraction layer is achieved by simply w-maxing. Fine. But in general? AIXI gives an upper bound across all possible environments. For embodied intelligence, that is like an upper bound across all possible abstraction layers. To formalise that I'll need some extra steps. Recall the definition of uninstantiated task from earlier:

- The set of all tasks with no abstraction (meaning $\mathfrak{v} = P$) is Γ_P (it contains every task in every vocabulary). For every P-task $\rho \in \Gamma_P$ there exists a function $\lambda_\rho : 2^P \to \Gamma_P$ that takes a vocabulary $\mathfrak{v}' \in 2^P$ and returns a highest level child $\omega \sqsubset \rho$ which is also a $\mathfrak{v}'$-task. We call λ_ρ an **uninstantiated-task**, and $\lambda_1 \sqsubset \lambda_2$ iff $\lambda_1(P) \sqsubset \lambda_2(P)$.

I'll now use this definition to establish the *utility* of w-maxing given an uninstantiated task and a vocabulary.

Definition 11 (utility of intelligence)
Every task $\gamma \in \Gamma$ has a "utility of intelligence"[391] *computed as $\epsilon : \Gamma \to \mathbb{N}$ such that $\epsilon(\gamma) = \max_{m \in \Pi_\gamma} (|E_m| - |O_\gamma|)$. Maximisation of utility means maximising $\bullet \overset{\epsilon}{<} \bullet$ that returns true iff $\epsilon(\alpha) < \epsilon(\omega)$.*

[391] Assuming we accept that intelligence is the ability to generalise, then we can measure the utility of selecting policies in accord with Bennett's Razor by measuring the weakness of the weakest policy for a task. Tasks with weaker policies make more use of intelligence.

This tells me the difference between w-maxing and not. Obviously utility is minimised by having a task with one policy. Likewise, there's no utility if there is no correct policy, so predicting noise is out. However, in the *absence* of any abstraction there is always a correct policy. Utility is maximised when $\mathfrak{v} = P$, though in practice finite resources would limit us to smaller vocabularies. Γ_P contains all tasks in all vocabularies. Hence, for every task ρ in Γ_P we can define a function that takes a vocabulary $\mathfrak{v}$ and returns a $\mathfrak{v}$-task which is a child of ρ. For this proof, I'll set the vocabulary $\mathfrak{v}$ to P and work "inwards" to more constrained abstraction layers.

Theorem 4 (upper bound) *The most 'intelligent' choice of policy and vocabulary given uninstantiated task λ_ρ is π and $\mathfrak{v}$ s.t. $\mathfrak{v}$ maximises utility for $\lambda_\rho(\mathfrak{v})$, $\pi \in \Pi_{\lambda_\rho(\mathfrak{v})}$ and π maximises weakness.*

Proof 4 *We have equated intelligence with sample efficient generalisation. The weakest correct policies have the highest probability of generalising. Given an uninstantiated task λ_ρ, utility measures the weakness of the weakest correct policies. We can use this to compare vocabularies. By choosing a vocabulary $\mathfrak{v}$ which maximises utility for $\lambda_\rho(\mathfrak{v})$, we instantiate λ_ρ in a vocabulary that maximises the weakness of correct policies for λ_ρ even in the absence of abstraction (meaning when $\mathfrak{v} = P$). Then, using weakness proxy, we can select a policy that has the highest possible probability of generalising, and thus maximise sample efficiency.* □

Now, this isn't to say that w-maxing is always going to get you there given any old abstraction layer, but it works at least from an objective point of view where $\mathfrak{v} = P$[392]. Generally speaking, the best way to optimise toward intelligent systems is to build abstraction layers that enable the system to embodied the weakest correct policies *relevant* to the task.

[392] Meaning in the absence of abstraction.

EXPERIMENTS

In the appendix is a Python script to perform two experiments using PyTorch with CUDA, SymPy and A^*[393]. Here I'll summarise the results and how the experiments worked. First, I wrote a toy program based on A^* that learns policies for 8-bit string prediction tasks (binary addition and multiplication)[394].

(abstraction layer) I used a simplified environment of 256 states, one for every possible 8-bit string. Basically an abstraction layer without our environment. The statements in L were expressions regarding those 8 bits that could be written in propositional logic ($\neg$, $\wedge$ and $\vee$).

(task) A task was specified by choosing $O \subset L$ such that all $d \in O$ conformed to the rules of either binary addition (for the first experiment) or multiplication (for the second experiment) with 4-bits of input, followed by 4-bits of output.

Each of the two experiments (addition and multiplication) involved repeated trials (sampling results). The parameters of each trial were "operation" (a function), and an even integer "number_of_trials" between 4 and 14 which determined the cardinality of the set O_k (defined below). Each trial was divided into training and testing phases.

(training phase)

1. A task T_n was generated:
   (a) First, every possible 4-bit input for the chosen binary operation was used to generate an 8-bit string. These 16 strings then formed O_n.
   (b) A bit between 0 and 7 was then chosen, and I_n created by cloning O_n and deleting the chosen bit from every string (meaning I_n was composed of 16 different 7-bit strings, each of which could be found in an 8-bit string in O_n).
2. A child-task $T_k = \langle I_k, O_k \rangle$ was sampled from the parent task T_n. Recall, $|O_k|$ was determined as a parameter of the trial.
3. From T_k two policies (formerly known as models) were generated; a weakest c_w, and a MDL c_{mdl}.

(testing phase) For each policy $c \in \{c_w, c_{mdl}\}$:

1. The extension E_c of c was then generated.
2. A prediction O_{recon} was then constructed s.t. $O_{recon} = \{e \in E_c : \exists s \in I_n \ (s \subset z)\}$.

[393] Adam Paszke et al. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019; David Kirk. Nvidia cuda software and gpu parallel computing architecture. In *Proceedings of the 6th International Symposium on Memory Management*, ISMM '07, page 103–104, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595938930. DOI: 10.1145/1296907.1296909. URL https://doi.org/10.1145/1296907.1296909; Aaron Meurer, Christopher Smith, Mateusz Paprocki, Ondřej Čertík, Sergey Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian Granger, Richard Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, and Anthony Scopatz. Sympy: Symbolic computing in python. *PeerJ Computer Science*, 3:e103, 01 2017. DOI: 10.7717/peerj-cs.103; and Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968. DOI: 10.1109/TSSC.1968.300136

[394] Notation here varies slightly from the formal notation due to the limitations of what can be written in Python (not latex), and because it the experiments coincided with an earlier iteration of the formalism.

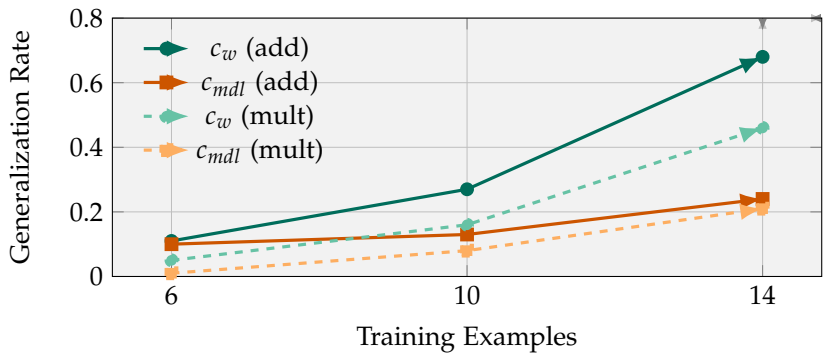


Figure 1: Rates for binary addition (solid) and mult. (dashed).

3. O_{recon} was then compared to the ground truth O_n, and results recorded.

Between 75 and 256 trials were run for each value of the parameter $|O_k|$. Fewer trials were run for larger values of $|O_k|$ due to restricted availability of hardware. The results of these trails were then averaged for each value of $|O_k|$.

(MEASUREMENTS) Generalisation was deemed to have occurred where $O_{recon} = O_n$. The number of trials in which generalisation occurred was measured, and divided by n to obtain the rate of generalisation for c_w and c_{mdl}. Error was computed as a Wald 95% confidence interval. Even where $O_{recon} \neq O_n$, the extent to which policies generalised could be ascertained. $\frac{|O_{recon} \cap O_n|}{|O_n|}$ was measured and averaged for each value of $|O_k|$, and the standard error computed.

	c_w				c_{mdl}			
$\|O_k\|$	Rate	±95%	AvgExt	StdErr	Rate	±95%	AvgExt	StdErr
6	.11	.039	.75	.008	.10	.037	.48	.012
10	.27	.064	.91	.006	.13	.048	.69	.009
14	.68	.106	.98	.005	.24	.097	.91	.006

Table 1: Binary addition.

	c_w				c_{mdl}			
$\|O_k\|$	Rate	±95%	AvgExt	StdErr	Rate	±95%	AvgExt	StdErr
6	.05	.026	.74	.009	.01	.011	.58	.011
10	.16	.045	.86	.006	.08	.034	.78	.008
14	.46	.061	.96	.003	.21	.050	.93	.003

Table 2: Binary multiplication.

THE KING IS DEAD, LONG LIVE THE KING!

This has profound implications. Simp-maxing is treated with great reverence in many circles. When I have presented these results at conferences, reactions have varied wildly. This challenges the foundations of approaches to AI based on simp-maxing, and by extension information theory and thermodynamics. I have listed the assumptions, the proofs and the experimental results. It is surprising, but it seems to check out. Human cognition is about breadth, analogy and the minimising surprise[395]. I am not disagreeing with those fundamental ideas, but with complexity's role in them. To effectively w-max, all of the above is likely to be necessary. W-maxing is a *meta-approach* after all. However, it seems hasty to completely dismiss simp-maxing. The next chapter will explore its significance.

[395] Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010; Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017. DOI: 10.1017/S0140525X16001837; and

VIII. STACKISM

Why does simplicity of form correlate with generalisation in function? I've already shown it isn't necessary. Yet the subjective perception of simplicity remains stubbornly correlated with the objective generalisation of functionality. Why and how are subjective and objective related here? What causes this? Is there somehow an objective notion of complexity? Is this hard-coded into us by evolution? How? This chapter is based on my papers on complexity[396], and on abstraction layers[397]. I'll begin with why simpler forms are correlated with generalisation in function. Second, I'll dissect how systems come to embody this correlation. I'll map systems along dimensions of abstraction, distribution of work and delegation of control. This allows me to compare systems. Third, I'll compare biological self-organisation to AI and conclude that biology is more adaptable because it delegates control to lower levels of abstraction. This causes weak constraints to be embodied by simple forms. This means w-maxing requires delegating control, and w-maxing while delegating control will cause simp-maxing. However, simp-maxing and delegating control will not cause w-maxing. In my papers I also discuss what this means for systems in general, exemplifying my points with human organisational and economic structures. I'll touch on these in the conclusion of the thesis.

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

[397] Michael Timothy Bennett. Computational dualism and objective superintelligence. In *Artificial General Intelligence*. Springer Nature, 2024a; and Michael Timothy Bennett. Are biological systems more intelligent than artificial intelligence? 2025b. Forthcoming 2026

WHY SIMP-MAXING SORT OF WORKS?

Empirically, there is a correlation between simplicity of form and generalisation in function. I want to know *why*. I've established that the probability of generalisation and measure of weakness go hand in hand. Recall how I've defined simplicity and weakness.

- A **proxy** $<$ is a binary relation on statements, and the set of all proxies is Q.
- $<_w$ is the **weakness** proxy[398]. 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[399]. We have $l_1 <_d l_2$ iff $|l_1| > |l_2|$.

As the proof on the upper bound shows, there is a meaningful *objective* notion of of weakness in the absence of an abstraction layer[400]. Is there a similarly meaningful notion of complexity in the absence of abstraction? By this, I mean does knowing something's simplicity convey any information? Ideas like Legg-Hutter intelligence seem to suggest that there is an *objective* notion of simplicity. That the environment considers some things complex, and some simple. When Leike and Hutter found AIXI to be subjective, they still claimed it is optimal *if* the UTM AIXI uses matches the UTM with respect to which Legg-Hutter intelligence is measured. The latter UTM is like the *objective* measure of complexity, and AIXI's UTM is the *subjective*. Their argument was effectively that if objective and subjective complexity match, then AIXI's performance is optimal. I'll now refute this claim indirectly, by showing there's no such thing as *objective* complexity.

[398] By the weakness of a statement, we mean the cardinality of its extension. By the weakness of an extension we mean its cardinality.

[399] When we speak of simplicity with regards to a policy $\pi \in \Pi_\alpha$ we mean the cardinality of the smallest correct policy $\pi' \in \Pi_\alpha$ s.t. $E_{\pi'} = E_\pi$. The complexity of an extension is the **simplest** statement of which it is an extension.

[400] Meaning when $\mathfrak{v} = P$. The answer is no, by the way.

Theorem 5 (subjectivity) *If there is no abstraction, complexity can always be minimized without improving sample efficiency, regardless of the task.*

Proof 5 *In accord with the definition of an abstraction layer, 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 $\bigcap l = f$. Theorems 1 and 2 show that 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 of weakness, because the simplest representation of every extension is a set containing exactly one program.* □

When we don't have abstraction, everything is equally complex. Simplicity becomes meaningless. Complexity is an illusion perpetrated by abstraction layers. So no, there is no *objective* notion of simplicity. Yet the fact remains the the *subjective* perception of complexity is correlated with generalisation in function. How can this be? The key lies in the fact that, in reality, we never have the absence of abstraction. In fact, we never even have infinite vocabularies. If we did, we could store an infinite amount of information on a computer. The Bekenstein bound says that a bounded system can store only a finite amount of information. That means there are only so many permutations. Only so much information needed to specify the system's state. That is why vocabularies are finite. Can this be the cause of the correlation between simplicity of form and generalisation? We know that weak constraints on function are necessary and sufficient for generalisation, so the question can be reframed as "can a finite vocabulary correlate simple forms with weak constraints?". Yes, it can.

Theorem 6 (confounding) *If the vocabulary is finite, then policy weakness can confound*[401] *sample efficiency with policy simplicity.*

[401] *A confounds B* and *C* when for example $A =$ "*badlyinjured*" *causes* $B =$ "*died*" and $C =$ "*pickedupbyambulance*", and it looks like C causes B because $p(B \mid C) > p(B \mid \neg C)$, and yet it may be that $p(B \mid C, A) < p(B \mid \neg C, A)$.

Proof 6 *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* $\{a, b\} <_w \{a\}$*,* $\{a, b\} <_w \{b\}$*,* $\{a, b\} <_d \{a\}$*,* $\{a, b\} <_d \{b\}$*.* □

In causal language, simplicity does not *cause* generalisation. Weak constraints on function *cause* generalisation, but if the vocabulary is finite weakness can also *cause* simplicity. Weakness *confounds* simplicity and generalisation. The Bekenstein bound slams a ceiling on how much information fits in any finite space. A sort of cosmic storage cap. Nothing overlaps. Two things can't share the same spot at once. If we put too much in one spot we get a singularity, but not the good sort. That caps our vocabularies. From atoms to molecules to cells, at each layer of abstraction we are stuck with a finite set of terms. Intuitively, imagine you have to pick `100` words with which to construct a language. You would not pick `100` very specific words, like lantern or erudite. Instead, you'd pick words that are used in more circumstances. That have as many meanings as possible, like "swallow" which is both a bird and an action. You would pick "hot" and "wet" rather than "sultry". "Look" instead of "gaze". More precise words have more limited uses.

In mathematical terms think of Huffman coding , whereby the shortest codes are assigned to the most common messages[402]. Compression becomes meaningful when space is limited. With a finite vocabulary, simplicity and weakness have to become correlated if the system is to work well. Complex policies become one-trick ponies. Waste space on overly-specific policies, and the system will fail to adapt.

[402] David A. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 1952

SYSTEMS

I've established that weakness can confound simplicity and generalisation, but I've not explained what sort of system can actually do this. Computers are the way they are because we build them that way. We embody highly abstract human behaviours in silicon. Human mathematics and logical operations normally depend on a human to interpret them. What we have done with computers is hard-wire physical consequences into instruction set architectures, and those consequences conform to our expectations and interpretations of logic. However it came about, we have a natural tendency to embody weak constraints in simple forms. We do it in natural language, and we do it with programming languages. At least this is the case within the narrow confines of the problem sets to which we apply computers. The result is that shorter programs do tend to be more generalisable. It isn't perfect, but it is empirically observable. How is it we came to have this tendency? Yes, natural selection prefers week constraints take simple forms, but what is this system natural selection has built as a consequence of that?

THERE IS A GOOD REASON to ask this question, beyond my usual curiosity. The evidence suggests machine learning systems are far less sample and energy efficient than biological systems that learn the same tasks[403]. There must be more going on than just search or approximation. Biology must be using a meta-approach that makes it more efficient. Sample and energy efficiency mean w-maxing. Biology is clearly very good at w-maxing. How? What is it that biological self-organising systems do differently?

[403] M. Khajehnejad, F. Habibollahi, A. Paul, A. Razi, and B. J. Kagan. Biological neurons compete with deep reinforcement learning in sample efficiency in a simulated gameworld. arXiv preprint arXiv:2405.16946, 2024

THERE ARE FUNDAMENTAL CONSTRAINTS to the logic of living systems, particularly those that learn[404]. Most important for my purposes is the ability to store information in an internal state in order to retrieve it facilitate adaptation in future. The ability to learn, rather than have information hard coded. Put another way, a living system *is* a policy because it is an aspect of the environment. However that policy implies an abstraction layer, and that abstraction layer can then *express* a policy. A living, learning system is a policy that expresses yet another policy at a higher level of abstraction. That higher level of abstraction is the behaviour of the lower level. When I think of a plan, that plan is a policy I express in my neural substrate. It is the behaviour of my nervous system. This is *upward* causation from the body to the mental representation of the plan. When I act on that plan, I *delegate* it to my body. There is *downward* causation from the plan to my body, to my organs and eventually my cells[405]. A basic

[404] Ricard Solé et al. Fundamental constraints to the logic of living systems. *Interface Focus*, 2024

[405] Patrick McMillen and Michael Levin. Collective intelligence: A unifying concept for integrating biology across scales and substrates. *Communications Biology*, 2024

feature of any adaptive system is valence, by which I mean they are attracted to or repelled away from physical states[406]. A cell in isolation might have very simple dichotomy of attraction and repulsion. However cells can network. They can support a bioelectric information structure[407]. What is a simple dichotomy of one-dimensional valence at the level of a single cell is part of a rich tapestry of valence across the network. The same matter is simultaneously involved in many different computations occurring at different scales and levels of abstraction. This is called "polycomputing"[408]. It is possible to "program" these biological polycomputing systems to do the same tasks as computers, so that we can compare the two, but it is not simple. One way to program them is by imposing constraints until the system is forced to conform to our expectations, like slime mould navigating a cave[409]. Another is to use bioelectricity. The right signal can trigger a dramatic shift in morphology, like the growth of an organ[410] or cancer[411]. These means of programming are extrinsic, top down impositions on the system.

[406] Ricard Solé et al. Fundamental constraints to the logic of living systems. *Interface Focus*, 2024

[407] Michael Levin. Bioelectrical approaches to cancer as a problem of the scaling of the cellular self. *Progress in Biophysics and Molecular Biology*, 2021. Cancer and Evolution

[408] Joshua Bongard and Michael Levin. There's plenty of room right here: Biological systems as evolved, overloaded, multi-scale machines. *Biomimetics*, 8(1), 2023

[409] T. Nakagaki, H. Yamada, and A. Toth. Maze-solving by an amoeboid organism. *Nature*, 407(6803):470, 2000

[410] Sam Kriegman, Douglas Blackiston, Michael Levin, and Josh Bongard. A scalable pipeline for designing reconfigurable organisms. *Proc Natl Acad Sci U S A*, 117(4):1853–1859, January 2020

[411] Michael Levin. Bioelectrical approaches to cancer as a problem of the scaling of the cellular self. *Progress in Biophysics and Molecular Biology*, 2021. Cancer and Evolution

W-MAXING INVOLVES CHOOSING not just a weak policy, but an abstraction layer. A body is an abstraction layer. The right morphology can *express* a weak policy. The wrong morphology cannot. When cells network, they can express complex policies beyond the capabilities of any one cell. They form a 'solid brain' with a persistent structure that supports bioelectric information processing. However, nature's intelligence is not restricted to such solid brains. Information processing can be carried out by 'liquid' brains that are spread out across time and space[412]. Such a system may not actually store a coherent policy anywhere at any one time. Rather, the policy is implicit in the behaviour. Such liquid brains seem to be able to solve complex problems just as a solid brain might[413]. Biological systems seem to excel at *creating* abstraction layers. Sometimes liquid, sometimes solid. If we want to understand why biology adapts so well, then we can look at how it forms abstraction layers, and what sort of liquid and solid brains adapt well.

[412] Ricard Solé, Melanie Moses, and Stephanie Forrest. Liquid brains, solid brains. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1774):20190040, 2019. DOI: 10.1098/rstb.2019.0040. URL `https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2019.0040`

[413] Chris R. Reid, David J T Sumpter, and Madeleine Beekman. Optimisation in a natural system: Argentine ants solve the towers of hanoi. *Journal of Experimental Biology*, 214(1):50–58, jan 2011

A GROUP OF HUMANS IS A LIQUID BRAIN. It has layers of abstraction, just like everything else. The behaviour of a group of soldiers can be a squad, and the behaviour of a couple of squads can be a platoon. It isn't always, but it can be if it obeys certain constraints. These abstractions are classifications of behaviour. We ascribe these labels to groups of people that behave a certain way.

Some organisational and economic structures can adapt more efficiently than others. I'll embrace the military analogy as it has one very helpful example[414]. When peacekeepers were deployed to Bosnia in the 90s, some units were successful in preventing massacres. Others were not. A force called NORDBAT 2 was successful, but they were also unconventional. They subscribed to the doctrine of Mission Command, which emphasises autonomy and independent action. Control is delegated to the units on the ground so that they can choose how to serve the larger objectives of the mission[415]. Low ranking officers are taught to take the initiative. If the circumstances demand an unconventional approach, Mission Command enables it. It allows for more bottom up control in exactly the sort of organisation best known for autocratic top-down control. This enables adaptation. NORDBAT 2 were trigger happy, but effective in preventing massacres. The counterpoint to NORDBAT 2 was the Dutch force operating in the same region, under the same command. Their approach was different, as were the results. Micromanaged by their home government, they did not adapt. They did not engage, and they ended up presiding over a massacre.

[414] Tony Ingesson. *The Politics of Combat: The Political and Strategic Impact of Tactical-Level Subcultures, 1939-1995*. Doctoral thesis (monograph), Department of Political Science, Lund University, 2016

[415] For another example, Admiral Nelson famously did something similar. He encouraged subordinates to take action rather than wait for orders that might not make it through the chaos of battle. Thank you Sean Welsh for this example.

This illustrates an important point. Biological brains, whether liquid or solid, are distributed systems made up of many parts. Like a decentralised network of computers. Those individual parts adapt, and that enables the system as a whole to adapt. By delegating control to the forces on the ground, Mission Command simply leverages the computational infrastructure available. The feedback loop is shorter, because information doesn't need to propagate all the way up to a central command. Looking at systems this way, we can see parallels with computers. A computer can be made up of smaller computers, and it can delegate work to those smaller computers. In doing so, it can delegate *more* or *less* leeway in *how* those tasks are complete. Delegating work and letting smaller computers decide how that work is done is delegating *control*. Delegating work and rigidly constraining how the smaller computers go about it is just delegating work, not control. Looking at systems this way, we can chart every adaptive system along three dimensions:

1. abstraction: Every system is composed of abstraction layers.

2. distribution: The machinery of the abstraction layer can be distributed, or centralised. Distribution allows work to be divided up into subtasks that are undertaken in parallel. For example, a single core CPU is highly centralised, and a supercomputer made up of thousands of cores is distributed. A group of humans is highly distributed, and one human is not.

3. DELEGATION OF CONTROL: Within any system, control is delegated to a level of abstraction. Put another way goals, and the choice of how to go about them, can be delegated to different levels of abstraction. Say I have a group of humans. If I delegate control to the individual level, then those individuals choose what to do. If I exercise a little bit of control and insist they pursue a particular goal, then they can choose how they pursue it it. If I exercise more control I can tell them *how* to pursue it. I delegate less control to the individuals, concentrating more of it at the top. Less delegation of control means more top down micromanagement. Too little control and the system doesn't do what I want it to do. Delegate enough control, and the system becomes controlled bottom-up rather than top-down.

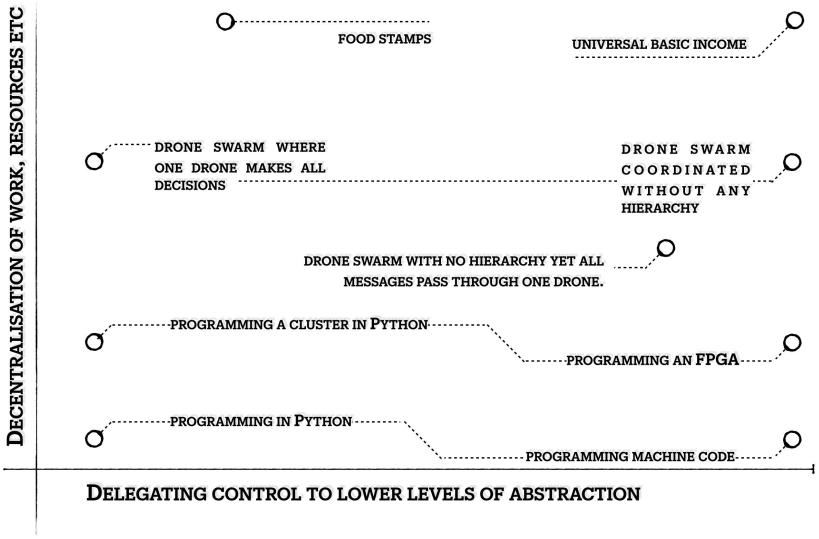


Figure 2: Systems can delegate control to a level of abstraction, by which I mean exert control at a particular level. They can also distribute work, resources and so on. Distribution and delegation should not be confused. One can delegate control to a lower level of abstraction without distributing it, as the figure illustrates. For example, both foodstamps and universal basic income might distributed the same amount of of resources to the same people, but the former constrains the people in how they use those resources.

OBVIOUSLY EVERY SYSTEM IS A BALANCE of these things. Delegation of control and distribution are often correlated, but they are not the same thing. Figure 2 and 3 illustrate. Consider an economy. In a free market, the level of abstraction to which control is delegated is the individual. In a command economy, control is concentrated at the state level. Economies can be more or less distributed. Supply lines can be centralised, monopolies can form, and control can be rigidly top-down. In an ideal Soviet style command economy, control is not delegated at all. It is concentrated at the top. Work is distributed among the components of the system, but not control.

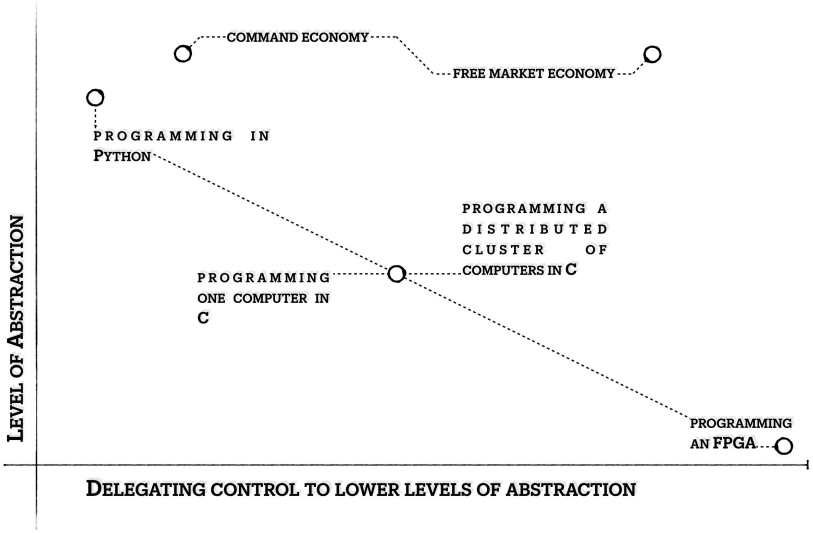


Figure 3: Notice that delegation in this sense is *distinct* from the distribution or decentralisation of work and resources. For example, if I program a cluster in C then my control is at the level of C, not lower. It doesn't matter whether I program one computer or fifty: control is happening at the same level of abstraction. I can *look* at the same system at many different levels of abstraction, but control takes place at *specific* levels of abstraction.

In contrast, in an ideal free market of the sort that Friedman[416] or Hayek might have liked[417], control is entirely delegated. Individuals decide what to buy, and what to produce. Networks of individuals are allowed to form in order to exert localised top-down control over companies and other small organisations. Eventually, the upward causation of this bottom-up control governs the entire economy. Of course, these are just ideals to illustrate what I mean by delegated control. Free markets can degrade into centralised monopolies, and monopolies can be broken. Now I will explain how distribution and delegation of control are formalised.

[416] M. Friedman and R.D. Friedman. *Capitalism and Freedom*. University of Chicago Press, 1962

[417] FA Hayek. The use of knowledge in society. *American Economic Review*, 35(4), 1945

DISTRIBUTION

Distribution amounts to having more than one policy expressed by an abstraction layer. For example in a collective of cells each cell is a policy, and if those cells are working towards the same goal then the intersection of their extensions is extension of their *collective* policy. That collective policy is the higher level of abstraction. The "identity" of the collective, to borrow the neurobiological term[418]. For example, say ω, α, β and γ are $\mathfrak{v}$-tasks. Let ω be an organ made up of cells α, β and γ. That means α, β and γ are all children of ω. $\Pi_\omega = \Pi_\alpha \cap \Pi_\beta \cap \Pi_\gamma$, meaning there can only be a coherent collective policy if the parts of the system can *share* a policy, and if the vocabulary can express it. That collective policy is the aforementioned collective identity. If the task is too difficult, then there may be no possible collective identity and the collective may break apart. This is consistent with descriptions of *cancer* as a loss of collective identity, when cells become isolated from the collective information structure and revert to primitive transcriptional behaviour[419]. However I'll discuss that in later chapters. For now, what matters is that we can represent the distributed nature of a system as a set of child tasks. Distribution is might be more conventionally conceived of in terms of connections in a graph (see 4). I don't need to explicitly represent the connections. They are implicit in the *correctness* of tasks. What matters is whether it is *possible* for the parts of a system to work according to a shared policy, and whether they actually *do*.

[418] Patrick McMillen and Michael Levin. Collective intelligence: A unifying concept for integrating biology across scales and substrates. *Communications Biology*, 2024

[419] Michael Levin. Bioelectrical approaches to cancer as a problem of the scaling of the cellular self. *Progress in Biophysics and Molecular Biology*, 2021. Cancer and Evolution

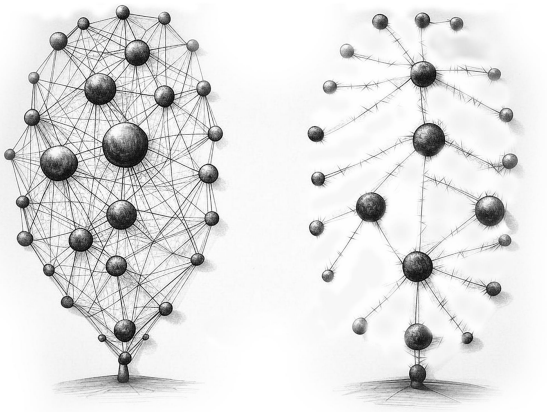
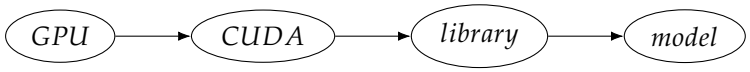


Figure 4: Illustration of a decentralised system (left) vs a centralised system (right). Notice that if a central node from the system on the right is deleted, then large portions of the graph become disconnected. If a node is deleted from the left graph, then there would still be a path between all remaining nodes. This illustrates what I mean about correctness. The more fragile, sparsely connected graph may have fewer correct policies and thus may be more prone to failure.

DELEGATION

DELEGATION OF CONTROL IS HOW CONSTRAINED the system is at different levels. Where are policies selected? Where does adaptation take place. Consider a typical machine learning stack.



ASSUMING A TYPICAL SORT OF MACHINE learning model, adaptation just involves tuning some parameters until the model spits out responses resembling the training data. Adaptation occurs *only* at the level of the model. The library does not change to better suit the task at hand. CUDA does not change. The GPU does not change. In contrast, consider the system which originally constructed the model. It would look something like this.

THIS STACK ADAPTS AT MUCH LOWER LEVELS. The human can construct new math to make a better model for the task at hand. When we swap out the human and math for a static computational stack, we lose that adaptability. We also lose the conscious effort involved, so it's overall a good thing, but for that convenience we trade adaptability. Does this hold up in practice? Yes. Recall the utility measure from earlier, that says how *useful* intelligence is for a task by measuring the difference between the strongest and weakest policies for a task in a vocabulary. Also, the multilayered architecture.

- $\epsilon : \Gamma \to \mathbb{N}$ such that $\epsilon(\gamma) = \max_{m \in \Pi_\gamma} (|E_m| - |O_\gamma|)$. Maximisation of utility means maximising $\bullet \overset{\epsilon}{<} \bullet$ that returns true iff $\epsilon(\alpha) < \epsilon(\omega)$.
- The **stack** is represented here by a sequence of uninstantiated tasks $\langle \lambda^0, \lambda^1 ... \lambda^n \rangle$ s.t $\lambda^{i+1} \sqsubset \lambda^i$.
- $\mathfrak{f}$ is an **abstractor** function.
- The **state** of the MLA is a sequence of policies $\langle \pi^0, \pi^1 ... \pi^n \rangle$ and a sequence of vocabularies $\langle \mathfrak{v}^0, \mathfrak{v}^1 ... \mathfrak{v}^n \rangle$ such that $\mathfrak{v}^{i+1} = \mathfrak{f}(\mathfrak{v}^i, \pi^i)$ and $\pi^i \in \Pi_{\lambda^i(\mathfrak{v}^i)}$.

DELEGATING ADAPTATION TO LOWER in the stack allows lower levels systems to optimise for weaker policies that suit the task at hand. That results in a more relevant vocabulary at higher levels, which permits weaker policies and thus adaptation at higher levels of abstraction.

Theorem 7 (The Law of the Stack) *The greater the utility* $\epsilon(\lambda^{i+1}(\mathfrak{v}^{i+1}))$*, the weaker the policy* π^i *s.t.* $\mathfrak{f}(E_{\pi^i}) = \mathfrak{v}^{i+1}$ *must be*[420]*.*

[420] Intuitively, this just means adaptability at higher levels implies adaptability at lower levels.

Proof 7 *If* $\mathfrak{a} \subset \mathfrak{b}$ *then* $\epsilon(\lambda^{i+1}(\mathfrak{a})) < \epsilon(\lambda^{i+1}(\mathfrak{b}))$*, meaning if* $\mathfrak{b}$ *is the vocabulary at* $i+1$*, then it will be possible to construct weaker policies than if* $\mathfrak{a}$ *is the vocabulary (intuitively, a larger vocabulary enables a wider range of policies). We consider two policies* $\pi^i_{\mathfrak{a}}$ *and* $\pi^i_{\mathfrak{b}}$ *which could be the policy* π^i *at i. If* $\mathfrak{a} = \mathfrak{f}(E_{\pi^i_{\mathfrak{a}}}) \subset \mathfrak{f}(E_{\pi^i_{\mathfrak{b}}}) = \mathfrak{b}$*, then* $\pi^i_{\mathfrak{a}} <_w \pi^i_{\mathfrak{b}}$*, meaning an enlarged vocabulary at* $i+1$ *implies a weaker policy at i.* □

INTERPRETATION

Adaptability at higher levels is constrained by adaptability at lower levels. If I unplug a computer, it ceases to function. Its stack is inflexible. Brittle. Humans are also quite brittle outside of our environment, but we are extremely adaptable within it. Our environment is an abstraction layer. The human stack is expressed within it. By w-maxing within the confines of our environmental abstraction layer, the human stack delegates control. Adaptation takes place at every scale and level of abstraction of the human stack[421]. Hence we are much more effective at w-maxing, and thus much more sample and energy efficient than the computers we build. When a stack w-maxes and can delegate control in the process, it can optimise its abstraction layers so that weaker constraints are expressed by simpler forms. To more efficiently use space and thus energy. I call this *The Stack Theory of Intelligence and Consciousness*, or *Stackism* for short. This has a profound implication. If control can be delegated completely, then there is no difference between w-maxing and simp-maxing. The ability to delegate and w-max perfectly, would perfectly correlate simplicity of form and weakness of constraints on function.

[421] Patrick McMillen and Michael Levin. Collective intelligence: A unifying concept for integrating biology across scales and substrates. *Communications Biology*, 2024

If a theoretical agent like AIXI were not based on a UTM but *delegated* interpretation down the stack, then it really could be optimal. My upper bound, which maximises the utility of the abstraction layer, must do this. It would be the only way to actually build a system that achieved that upper bound. There are many promising avenues toward this goal. Reverse engingeering biology is one approach[422]. Likewise biomimetic approaches to robotics could work. Homeostatic soft robots are a proposed class of machine that pursues homeostasis[423]. Think of artificial life[424] in embodied form. To achieve homeostasis, these robots must delegate adaptation down stack to the level of cells because they need to self repair in order to maintain homeostasis. That means they need to w-max, and delegate, and thus will simp-max as well. Self-organising systems of nano-particles might be a promising approach[425].

[422] Sam Kriegman, Douglas Blackiston, Michael Levin, and Josh Bongard. A scalable pipeline for designing reconfigurable organisms. *Proc Natl Acad Sci U S A*, 117(4):1853–1859, January 2020

[423] Kingson Man and Antonio R. Damasio. Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence*, 1:446 – 452, 2019. URL `https://api.semanticscholar.org/CorpusID:208089594`

[424] Keisuke Suzuki and Takashi Ikegami. Spatial-pattern-induced evolution of a self-replicating loop network. *Artificial Life*, 12(4):461–485, 2006; and Takashi Ikegami and Keisuke Suzuki. From a homeostatic to a homeodynamic self. *Biosystems*, 91(2):388–400, 2008

[425] B. Paroli, G. Martini, M.A.C. Potenza, M. Siano, M. Mirigliano, and P. Milani. Solving classification tasks by a receptron based on nonlinear optical speckle fields. *Neural Networks*, 166: 634–644, 2023; and Francesca Borghi, Thierry R. Nieus, Davide E. Galli, and Paolo Milani. Brain-like hardware, do we need it? *Frontiers in Neuroscience*, 18, 2024

IX. LETS GET PSYCHOPHYSICAL

This is based on my papers on consciousness[426]. From difference I got aspects, and from aspects I got abstraction layers. The Stack Theory of Intelligence and Consicousness. The entire universe as a stack of abstraction layers. The environment preserves those aspects of it which preserve themselves. Which 'move' or are moved toward self preservation. Every aspect is itself an abstraction layer, which can express another aspect which can constrain the future to those worlds in which the system persists. This gives us fitness. Survival. Actively rather than passively goal directed behaviours.

However my Stack Theory is a theory of everything and nothing. It has no content. There are no objects. No properties. No trees or mountains. Just the pure unmediated stuff of reality. Our subjective worlds have trees and mountains. There are objects, and they have properties. Where are they in my Stack Theory? Not much good having a formalism of everything if it describes nothing. To rectify this problem, I look to causality.

[426] Michael Timothy Bennett. Emergent causality and the foundation of consciousness. In *Artificial General Intelligence*. Springer Nature, 2023b; Michael Timothy Bennett, Sean Welsh, and Anna Ciaunica. *Why Is Anything Conscious?* Preprint, accepted to and presented at ASSC27 and MoC5, 2024; and Michael Timothy Bennett and Ricard Solé. Does suspended animation kill consciousness? *Under review*, 2025

IN SOVIETY UNION, ACTION CAUSE YOU

An important aspect of adaptation is the ability to discern cause and effect. Systems which do not correctly identify cause and effect cannot intervene in the environment to *cause* events that aid their survival. Optimality *requires* causal learning[427].

[427] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., New York, 1st edition, 2018

Normally if I wanted to think about causality I would start with a set of variables representing objects and their properties. To describe their causal relations, we would relate these objects and properties to each other using a directed acyclic graph. One thing points to another in a causal chain. If we didn't know these causal relations, we could run experiments to figure out if changing one variable causes another to change. This works great for science, but only because we humans have already have the world divided up into distinct objects and properties. I can test to see if water puts out fire, because I already have classifiers for water and fire.

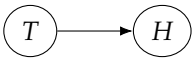
This is often formalised using the idea of interventions. Pearlean *interventionist* causality poses the problem of causality as one of discriminating between passive observation of an event, and intervening to *cause* that event. For example, lets suppose my dog only howls when there is thunder. He howls every time there is thunder. Being a good Bayesian I come to the conclusion based on my observations that my dog howling is the cause of thunder, and that if I wish to control the weather I need only make my dog howl. I'll formalise to illustrate. Thunder is a variable $T \in \{\texttt{true}, \texttt{false}\}$ such that $T = \texttt{true}$ iff there is thunder. My dog howling is a variable $H \in \{\texttt{true}, \texttt{false}\}$ and $H = \texttt{true} \longleftarrow T = \texttt{true}$. My insane belief in my dog's ability to control the weather is:

$$p(T = \texttt{true} \mid H = \texttt{true}) = 1$$

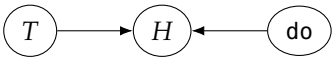
Now this is technically true based on my observations. It is stupid, but it is true, because it completely ignores causality. Lets assume after renaming my dog Thor I start to try to *intervene* in the environment to cause thunder. This means instead of waiting for it to rain and just passively observing my dog happening to howl when there is thunder, I start trying to make my dog howl hoping thunder will follow. This *intervention* is represented by a `do` operator applied to the variable H as $\texttt{do}(H = \texttt{true})$. Using this operator, I can now represent the difference between observing my dog howl when it thunders, and making my dog howl hoping I get thunder:

$$p(T = \texttt{true} \mid \texttt{do}(H = \texttt{true})) = p(T = \texttt{true}) \neq p(T = \texttt{true} \mid H = \texttt{true})$$

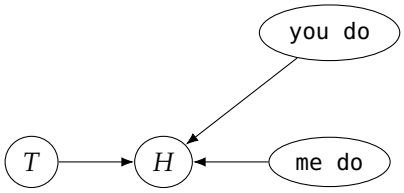
WHAT THIS CAPTURES IS THE DIRECTION OF CAUSALITY. Thunder causes my dog to howl. My dog's howling does not cause thunder.



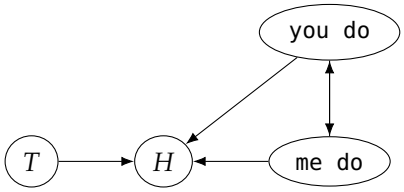
THE DO OPERATOR IS EQUIVALENT to a variable that influences the value of H. So really, we can represent any intervention by just expanding the graph as follows.



THIS PARTICULAR FACT was pointed out by Dawid in 2002, but by the time someone told me this I had already gone to the trouble of proving it. The proof is in the appendix. Nevertheless I now build on all this in an important way. All of this talk about intervening to *cause* an event and me passively *observing* an event is rather narcissistic. I am not the only agency capable of causing events. The question isn't whether *I* caused an event. The question is *what* caused an event. Passive observation is just another way of saying something other than me caused the event. This can be resolved by just adding more variables. I can make my dog howl. You can make my dog howl. The thunder can make my dog howl. I'll represent these with binary variables `you do` $\in \{\texttt{true}, \texttt{false}\}$ and `me do` $\in \{\texttt{true}, \texttt{false}\}$.



IT GETS COMPLICATED WHEN I DO THIS. The fact that we're dealing in high level abstractions like me and you means we lose the acyclicality of the graph. It becomes bidirectional We end up with cases like where you intervene, and in doing so thwart my attempted intervention[428]. Or the opposite.



[428] For example, if you intervene to set H to true and I try to set it to false, and you whack me over the head with my dog, making him howl and me fail in my attempted intervention. For science, of course.

Why this bidirectionality? Because computational dualism! Because these variables ignore important information. I can cause you to intervene or not, and you can cause me to intervene or not. The `do` operator gets away with acyclic graphs because it conveniently ignores the fact that there is more than one causal agency in the environment. Great for science, which was what it was proposed for. Bad for designing artificial intelligence. In my formalism variables and values like $H = \texttt{true}$, $T = \texttt{true}$ and `me do` are just *aspects* of the environment. By presupposing the world is divided up into variables, we presuppose there are certain dividing lines between aspects of the environment. That would undermine any claim I might then make, because assuming objects and properties amounts to assuming an abstraction layer. As I showed earlier complexity is determined by the abstraction layer. Why is a chair a chair? Because it affords us something[429]. We work in terms of what is *relevant*[430] to our survival[431]. If I want to properly capture causality, I need to capture *relevance*. I need to explain how I get from an environment made up of unlabelled aspects and programs, to something that has an *identity* that *causes* something. That intervenes. A *causal-identity*. Normally a causality researcher would start with variables and learn the causal relations, but here we have no variables. We need to learn the variables. This led me to ask: if I had a causal relation, could I then learn the variables that fit that causal relation? Yes! I would just need to learn a policy that classifies the causal relation.

This is where the existential ought from earlier comes in handy here. It gives me attraction and repulsion from physical states. Valence[432]. That simple dichotomy is enough to get us everything else. It is my foundational causal relation. Instead of assuming a set of objects and properties and trying to learn the causal relations between then, I can flip the problem. I can assume the causal relation, and learn the objects and properties. The causal relation is valence. If I am an adaptive system, then some aspects of the environment will attract me, and some will repel me. Instead of starting with two variables and learning the causal arrow between them, I can start with the arrow and learn the variables.

[429] James J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979

[430] What is relevant is determined by the cosmic ought. It preserves and destroys aspects, and each aspect forms an abstraction layer at a higher level of abstraction.

[431] John Vervaeke, Timothy Lillicrap, and Blake Richards. Relevance realization and the emerging framework in cognitive science. *J. Log. Comput.*, 2012; John Vervaeke and Leonardo Ferraro. *Relevance, Meaning and the Cognitive Science of Wisdom*. Springer Netherlands, Dordrecht, 2013a; and John Vervaeke and Leonardo Ferraro. Relevance realization and the neurodynamics and neuroconnectivity of general intelligence. In Inman Harvey, Ann Cavoukian, George Tomko, Don Borrett, Hon Kwan, and Dimitrios Hatzinakos, editors, *SmartData*, NY, 2013b. Springer Nature

[432] The transition of the environment from one state to another selects for aspects that preserve themselves. Things that preserve themselves are attracted to circumstances that preserve them. This attraction is otherwise known as valence. We get valence from the simple fact of change.

THE PURE UNMEDIATED EXPERIENCE OF BEING

I have proofs of optimal adaptability. A system which identifies the weakest policies is optimal. To do so, an adaptive system may delegate control to construct an optimal abstraction layer in which to construct the optimal policy. Any system that adapts optimally must correctly identify cause and effect[433]. It must correctly discriminate between observation and intervention, so let me define exactly what I mean by causal intervention here.

Definition 12 (intervention)
*Intuitively, if int and obs are "events" which have happened, then we say that int has **caused** obs if obs would not have happened in the absence of int (counterfactual). In our formalism, an **event** is a statement in $L_{\mathfrak{v}}$, and an event **happens** or is **observed** iff it is a true statement. If $obs \in L_{\mathfrak{v}}$ is sensorimotor activity we interpret as an "observed event", and $int \in L_{\mathfrak{v}}$ is in **intervention** (by an organism or other agency, in the sense described by Pearl[434]) to cause that event, then $obs \subset int$ (because int could not be said to cause obs unless $obs \subset int$).*

By learning policies in response to attraction and repulsion from environmental states, a system must construct policies that classify those parts of the environment which intervene to *cause* that valence. I call these policies **causal-identities**. For example, to know that I have been bitten by a dog, I must have a causal-identity for that dog. That causal-identity is how I react to the dog. The dog is what it *affords* me[435], rather than something platonic. A dog means something very different to a flea than it does to me.

Definition 13 (causal identity) *If[436] $obs \in L_{\mathfrak{v}}$ is an observed event, and $int \in L_{\mathfrak{v}}$ is in intervention causing obs, then intuitively $c \subseteq int - obs$ "identifies" or "names" the intervening agency if $c \neq \emptyset$. We call c a **causal identity** corresponding to int and obs. Suppose INT and OBS are sets of statements, and we assume OBS contains observed events and INT interventions, then a causal identity corresponding to INT and OBS is $c \neq \emptyset$ s.t. $\forall int \in INT(c \subset int)$ and $\forall obs \in OBS(c \cap obs = \emptyset)$ (we can attempt to construct a causal identity for any INT and OBS). If a policy is a causal identity, then the associated task is to classify interventions.*

[433] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., New York, 1st edition, 2018; Jonathan Richens and Tom Everitt. Robust agents learn causal world models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pOoKI3ouv1; and Michael Timothy Bennett. Emergent causality and the foundation of consciousness. In *Artificial General Intelligence*. Springer Nature, 2023b

[434] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., New York, 1st edition, 2018

[435] James J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979

[436] (EXAMPLE) Suppose we have organisms $\mathfrak{a}$ (Alice) and $\mathfrak{b}$ (Bob). The inputs Alice has experienced so far $I_{\mathfrak{h}_{<t_{\mathfrak{a}}}}$ can be divided into those in which Bob affected Alice $I_{\mathfrak{a}}^{\mathfrak{b}}$ and those in which Bob did not $I_{\mathfrak{a}}^{\neg\mathfrak{b}} = I_{\mathfrak{h}_{<t_{\mathfrak{a}}}} - I_{\mathfrak{a}}^{\mathfrak{b}}$. By affecting Alice, Bob has intervened in Alice's experience. Alice can construct a causal identity b for Bob corresponding to interventions $INT = I_{\mathfrak{a}}^{\mathfrak{b}}$ and observations $OBS = I_{\mathfrak{a}}^{\neg\mathfrak{b}}$. The objects which "exist" in Alice's experience are those for which she constructs a causal identity, so this is how Bob comes to exist as a distinct "object" which Alice experiences, rather than in parts of other objects).

Causal-identities are the effect of the environment upon a body. A sensory datum. A body is always being attracted or repelled by its particular surroundings. This causes it to express statements. Those statements are causal-identities for what attracted or repelled the body. They are prelinguistic classifiers, each one denoting a particular cause of valence.

Weaker causal-identities classify more commonly encountered causes of valence. This is just like how a weaker policy applies to more situations. If I have a causal-identity for my experiences of `red`, it is weaker than my causal-identity for `red lobster`. By w-maxing, a living system divides the world up into a hierarchy of policies. Some more specific. Some weaker. I suggest this is why and how a contentless environment can be divided up into objects and properties[437]. I call this The Psychophysical Principle of Causality[438].

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

[438] Michael Timothy Bennett, Sean Welsh, and Anna Ciaunica. *Why Is Anything Conscious?* Preprint, accepted to and presented at ASSC27 and MoC5, 2024

There are two preconditions which must be satisfied before a system will express a causal-identity for an object. First there must be an incentive, for example the object is relevant to survival. This is the **incentive precondition**. It aligns with the idea of affordances[439], in that a causal-identity is only constructed for an object to the extent that it *affords* a system some advantage to be able to recognise that object. The cosmic ought deems it so. Second, the system's abstraction layer must be capable of expressing a causal-identity for an object, that discriminates between events caused by the object and events which are not. This is the **scale** precondition, because the vocabulary of the abstraction layer must be of sufficient scale to represent the causal-identity. In other words I have formalised objects as their behaviour, and behaviour as tasks. A task only describes a coherent object if there exist correct policies for that task, which depends on the vocabulary. Hence the vocabulary must be of sufficient **scale** to represent and store a causal-identity.

[439] James J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979

Definition 14 (preconditions)
If $\mathfrak{o}$ *is an organism, and c is a causal identity:*

- *the* ***representation*** *precondition is met iff* $c \in L_{\mathfrak{v}_\mathfrak{o}}$*, and*
- *the* ***incentive*** *precondition is met if* $\mathfrak{o}$ *must learn c to remain "fit"*[440].

[440] It is possible an organism might construct *c* even if it is not required for the organism to remain fit, hence 'if' instead of 'iff'. Incentive is a sufficient predondition in conjunction with representation, but it is not strictly necessary.

A LEARNING SYSTEM WHICH W-MAXES will construct a causal-identity for anything which meets these preconditions. Things which do not meet these preconditions will effectively not exist, as far as the learning system is concerned. In one paper, I even used this fact to explain the Fermi paradox[441], arguing that our ability to recognise intelligence and thus life is contingent on that life being "like us". Life and intelligence could be all around us, but its behaviour falls outside the scope of what we're wired to notice and understand[442]. Put another way, we rationalise what we see. When we construct a causal-identity for something, we construct a model of its *intent*.

[441] Michael Timothy Bennett. Compression, the fermi paradox and artificial super-intelligence. In *Artificial General Intelligence*. Springer Nature, 2022b

[442] Michael Timothy Bennett. On the computation of meaning, language models and incomprehensible horrors. In *Artificial General Intelligence*. Springer Nature, 2023c

Definition 15 (purpose, goal or intent)
We consider a policy c which is a causal identity corresponding to INT and OBS to be the ***intent****,* ***purpose*** *or* ***goal*** *ascribed to the interventions. c is what the interventions share in common, meaning the "name" or "identity" of behaviour is the "intent", "goal" or "purpose" of behaviour. Just as an intervention caused an observation, the particular intent which motivated the agency undertaking the intervention is what caused it (to correctly infer intent, one must infer a causal identity that implies subsequent interventions).*

THE SELF

I am about to introduce the concept of self. To that end it is helpful to know *what* exactly is getting a self, so here is a formal definition of an organism. It is not central to this thesis, but for the sake of explaining what is being learned as a causal-identity, and for the latter chapters on "protosymbols" and meaning, it is helpful. I describe the circumstances of an organism[443] $\mathfrak{o}$ as $\langle \mathfrak{v}_\mathfrak{o}, \mu_\mathfrak{o}, \mathfrak{p}_\mathfrak{o}, <_\mathfrak{o} \rangle$ where:

- $O_{\mu_\mathfrak{o}}$ contains every output which qualifies as "fit" according to natural selection.
- $\mathfrak{p}_\mathfrak{o}$ is the set of policies an organism knows, s.t. $\mathfrak{p}_\mathfrak{o} \subset \mathfrak{p}_{n.s.} \cup \mathfrak{p}_{\mathfrak{h}_{<t_\mathfrak{o}}}$ and:
  - $\mathfrak{p}_{n.s.} \subset L_{\mathfrak{v}_\mathfrak{o}}$ is **reflexes** hard coded from birth by natural selection.
  - $\mathfrak{p}_{\mathfrak{h}_{<t_\mathfrak{o}}} = \bigcup_{\zeta \in \mathfrak{h}_{<t_\mathfrak{o}}} \Pi_\zeta$ is the set of policies it is possible to **learn** from a history of past interactions represented by a task $\mathfrak{h}_{<t_\mathfrak{o}}$.
  - If $\mathfrak{p}_{\mathfrak{h}_{<t_\mathfrak{o}}} \not\subset (\mathfrak{p}_\mathfrak{o} - \mathfrak{p}_{n.s.})$ then the organism has **selective memory**. It can "forget" outputs, possibly to productive ends if they contradict otherwise good policies.
- $<_\mathfrak{o}$ is a binary relation over $\Gamma_{\mathfrak{v}_\mathfrak{o}}$ we call **preferences**.

[443] (INTUITIVE SUMMARY) Strictly speaking an organism $\mathfrak{o}$ would be a policy, but we can describe the circumstances of its existence as a task μ that describes all "fit" behaviour for that organism. We can also identify policies the organism "knows", because these are implied by the policy that is the organism. Likewise, we can represent lossy memory by having the organism "know" fewer policies than are implied by its history of interactions. Finally, preferences are the particular "protosymbol" the organism will use to "interpret" an input in later definitions.

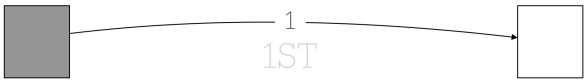
TO SURVIVE IN A COMPLEX interactive setting as humans do, one must be able to tell the difference between events one has caused, and events one has merely observed. This implies the construction of causal-identities for one's self. A `do` operator. Just because a `do` operator doesn't encompass the full breadth of causal agencies because it only accounts for the agency of one, doesn't mean it isn't critically important. One must represent one's self!

Definition 16 (first order self)
*If c is the strongest causal identity corresponding to INT and OBS, and INT is every intervention an organism could make (not just past interventions, but all potential future interventions), then we consider c to be the system's **first order self**. If $c \in \mathfrak{p}_\mathfrak{o}$ then an organism has constructed a first order self. A first order self for an organism $\mathfrak{o}$ is denoted $\mathfrak{o}^1$. An organism has at most one first order self.*

THIS IS EQUIVALENT TO REAFFERENCE. It lets one determine what one has done. It is the self that is part of every intervention one undertakes. Without a 1ST-order-self, there is nothing to tie together past actions in memory. A fly has a 1ST-order-self. There is a good

reason for this. Imagine a fly on my shoulder. If I move, the room moves around the fly. If the fly moves, the room moves around the fly. If the fly can't tell the difference between these two things, it is going to end up very dead. Flies can tell the difference between these two things. A system may have hard-wired policies that do not contain one's 1ST-order-self. Involuntary reflexes, for example. In those cases, those responses are not triggered by the self as I define it. One has a 1ST-order-self when one can classify one's own interventions. It amounts to an overall 'intent' or goal which predicts my actions. This is because a causal-identity is an statement made by the environment, and it has an extension. By constructing a self, one defines a constraint from which future behaviour may be abducted. One can derive all of one's possible agentic behaviours from a 1ST-order-self.

Figure 5: Illustration of a 1ST-order-self.



HOWEVER, THIS DOES NOT MEAN everything is preordained. A 1ST-order-self is just a very weak constraint. When system $\mathfrak{a}$ constructs a causal-identity for $\mathfrak{b}$, what it is doing is creating a classifier of what $\mathfrak{b}$ *affords* $\mathfrak{a}$. It is a sort of prediction of that narrow part of $\mathfrak{b}$'s 1ST-order-self which is relevant to $\mathfrak{a}$, if $\mathfrak{b}$ happens to have one. What happens when I go a step further? What happens when $\mathfrak{a}$ predicts $\mathfrak{b}$ in so much detail that it predicts $\mathfrak{b}$'s prediction of $\mathfrak{a}$? A sort of 2ND order of self, constructed in relation to another?

Definition 17 (chain notation)
Suppose we have two organisms, $\mathfrak{a}$ (Alice) and $\mathfrak{b}$ (Bob). $c_{\mathfrak{a}}^{\mathfrak{b}}$ denotes a causal identity for $\mathfrak{b}$ constructed by $\mathfrak{a}$ (what Alice thinks Bob intends). Subscript denotes the organism who constructs the causal identity, while superscript denotes the object. The superscript can be extended to denote chains of predicted causal identity. For example, $c_{\mathfrak{a}}^{\mathfrak{b}\mathfrak{a}} \subset c_{\mathfrak{a}}^{\mathfrak{b}}$ denotes $\mathfrak{a}$'s prediction of $\mathfrak{b}$'s prediction of $\mathfrak{a}^1$ (what Alice thinks Bob thinks Alice intends). The superscript of $c_{\mathfrak{a}}^{*}$ can be extended indefinitely to indicate recursive predictions, however the extent recursion is possible is determined by $\mathfrak{a}$'s vocabulary $\mathfrak{v}_{\mathfrak{a}}$. Finally, Bob need not be an organism. Bob can be anything for which Alice constructs a causal identity.

Definition 18 (n^{th} order self)
An n^{th} order self for $\mathfrak{a}$ is $\mathfrak{a}^n = c_{\mathfrak{a}}^{*\mathfrak{a}}$ where $*$ is replaced by a chain, and n denotes the number of reflections. For example, a second order self $\mathfrak{a}^2 = c_{\mathfrak{a}}^{\mathfrak{ba}}$, and a third order self $\mathfrak{a}^3 = c_{\mathfrak{a}}^{\mathfrak{baba}}$. We use $\mathfrak{a}^2$ to refer to any second order self, and chain notation to refer to a specific second order self, for example $c_{\mathfrak{a}}^{\mathfrak{ba}}$. The union of two n^{th} order selves is also considered to be an n^{th} order self, for example $\mathfrak{a}^3 = c_{\mathfrak{a}}^{\mathfrak{baba}} \cup c_{\mathfrak{a}}^{\mathfrak{dada}}$, and the weaker or higher level a self is in the generational hierarchy, the more selves there are of which it is part.

A 2ND-ORDER-SELF REQUIRES CAUSAL-IDENTITIES for other objects. It is my prediction of your prediction of my 1ST-order-self, or a narrow contextually relevant part thereof, what what I think is your perspective. This would be needed to anticipate and avoid predation. Conversely, it could help one herd and capture prey. One very distinct capability a 2ND-order-self conveys but a 1ST-order-self does not, is the ability to represent and reason about one's own destruction. If I can predict your prediction of me, then I can predict you observing my destruction. This makes death conceivable. It makes it possible to reason about yourself in general, and to engage in basic deception.

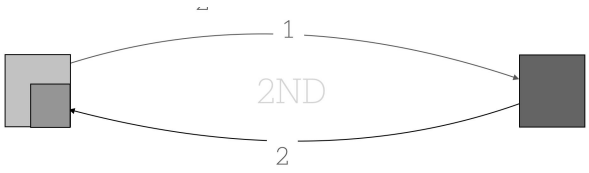


Figure 6: Illustration of a 2ND-order-self.

FINALLY, A 3RD-ORDER-SELF PERMITS one to predict one's own 2ND-order-selves, which would be useful in even more complex social or multi-agent environments. It makes it possible to reason about the reasoning of others about themselves, and to plan complex interactions taking into account this self-aware reasoning capability in others. This is makes complex deception possible, where I can reason about your reasoning about my reasoning about your reasoning about me. I can act to influence your reasoning about my reasoning about your reasoning about me, so that you think I have an interpretation of your behaviour that I do not. That I view you in a manner that I do not, and that I think you're going to do something in relation to me that I do not. However, if we can all predict each other so well, can we not also predict deception? The next chapter will thus deal in language, meaning and co-operation.

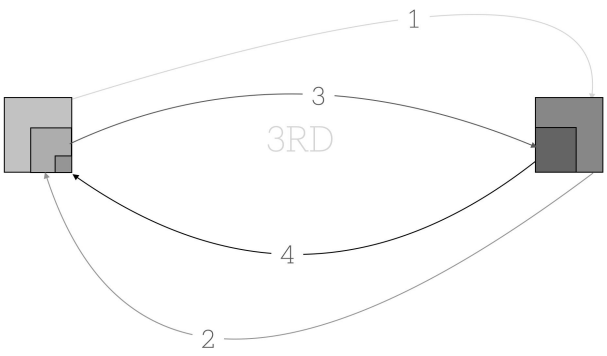


Figure 7: Illustration of a 3RD order self.

Theorem 8 (n^{th} order self convergence) *An organism that uses weakness as its proxy will learn an n^{th} order self if the incentive and representation preconditions are met for that order of self.*

Proof 8 *Assume we have an organism $\mathfrak{o}$ that learns using "weakness" as a proxy. A $\mathfrak{v}_\mathfrak{o}$-task $\mathfrak{h}_{<t_\mathfrak{o}}$ represents the history of $\mathfrak{o}$ (meaning $\mathfrak{h}_{<t_\mathfrak{o}} \sqsubset \mu_\mathfrak{o}$ and $\mathfrak{h}_{<t_\mathfrak{o}}$ is an ostensive definition of $\mu_\mathfrak{o}$, by virtue of the fact that $\mathfrak{o}$ remains alive). The organism explores the environment, intervening to maintain homeostasis. As it does so, more and more inputs and outputs are included in $\mathfrak{h}_{<t_\mathfrak{o}}$. It follows that:*

1. *From the representation precondition we have that there exists a n^{th} order self $\mathfrak{o}^\mathfrak{n} \in L_{\mathfrak{v}_\mathfrak{o}}$.*

2. *To remain fit, $\mathfrak{o}$ must "generalise" to $\mu_\mathfrak{o}$ from $\mathfrak{h}_{<t_\mathfrak{o}}$. According to the incentive precondition, generalisation to $\mu_\mathfrak{o}$ requires $\mathfrak{o}$ learn the n^{th} order self, which is when $\mathfrak{o}^\mathfrak{n} \in \mathfrak{p}_\mathfrak{o}$.*

3. *From this ref[444] we have proof that weakness is the optimal choice of proxy to maximise the probability of generalisation from child to parent is the weakest policy. It follows that $\mathfrak{o}$ will generalise from $\mathfrak{h}_{<t_\mathfrak{o}}$ to $\mu_\mathfrak{o}$ given the smallest history of interventions with which it is possible to do so (meaning the smallest possible ostensive definition, or cardinality $|O_\alpha|$).*

Were we to assume learning under the above conditions does not construct an n^{th} order self for $\mathfrak{o}$, then one of the three statements above would be false and we would have a contradiction. It follows that the proposition must be true. □

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

X. LANGUAGE CANCER

This chapter is about language. It is a combination of my Mirror Symbol Hypothesis[445], and my subsequent papers on symbol emergence[446] and the formalisation of Gricean pragmatics[447].

Normativity expressed in natural language is not the same as the *existential* normativity I spoke about earlier. Normativity in natural language is social normativity. What we think things mean, and what values we hold. Our interpretations. Meaning. The cosmic *ought* is normative only in the sense that it discriminates, seeming to judge some things worthy of existence and not others.

Meaning comes in many forms. The *semantic* meaning of a proposition is its truth conditions[448]. For example, the meaning of "Larry's cat is green" is true when Larry has a cat, and it is green. Of course, this can lead one in circles. We end up defining the semantic meaning of the first sentence in terms of two other sentences. We end up endlessly deferring semantic meaning. Something is always missing. This is what Derrida called "differance"[449]. Fortunately, this idea fits beautifully with Stack Theory. Semantic theories of meaning specify what the semantic truth conditional meaning of language is, so each such theory would be an abstraction layer. Another sort of theory is a *foundational* theory of meaning. A foundational theory descirbes a system that is a level of abstraction below. A foundational theory says what the system is that produces the semantic theory.

[445] Michael Timothy Bennett and Yoshihiro Maruyama. Philosophical specification of empathetic ethical artificial intelligence. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2): 292–300, 2022a

[446] Michael Timothy Bennett. Symbol emergence and the solutions to any task. In *Artificial General Intelligence*. Springer Nature, 2022a

[447] Michael Timothy Bennett. On the computation of meaning, language models and incomprehensible horrors. In *Artificial General Intelligence*. Springer Nature, 2023c

[448] J. Speaks. Theories of Meaning. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, Stanford, Spring 2021 edition, 2021

[449] Jacques Derrida. Writing and difference. *U of Chicago P*, 1978

GRICEAN PRAGMATICS

THE GRICEAN FOUNDATIONAL THEORY is that meaning is determined by intent[450]. Assume speaker α means m by saying u. By saying u, α intends that:

1. his audience come to believe m,
2. his audience recognise this intention [called m-intention], and
3. (1) occurs on the basis of (2).

[450] Paul Grice. Meaning. *The Philosophical Review*, 66(3):377–388, 1957; and Paul Grice. Utterer's meaning and intention. *The Philosophical Review*, 78(2):147–177, 1969

THIS IS ALSO CALLED PRAGMATIC MEANING. Basically, if you and I speak, then there are two meanings. There is the meaning I intend, and the meaning you interpret. The meaning you interpret is what you ascribe to my words and actions. The meaning I intend is what I *want* you to ascribe. We have understood each other if you ascribe the meaning I want you to ascribe. It is a prediction problem.

1. To *intend* a meaning, I need to predict what you think I think.
2. To *interpret* my meaning, you need to predict what I predicted you thought I thought.

THIS FITS NEATLY WITHIN THE FRAME OF SELVES. Assume I am $\mathfrak{a}$ and you are $\mathfrak{b}$:

1. To *intend* a meaning, I need to meet the scale and incentive precondutions to construct the causal-identity $c_{\mathfrak{a}}^{\mathfrak{b}\mathfrak{a}}$.
2. Then, to *interpret* my meaning, you need to meet the preconditions for the causal-identity $c_{\mathfrak{b}}^{\mathfrak{a}\mathfrak{b}\mathfrak{a}}$.

IN OTHER WORDS, WE BOTH NEED TO HAVE 2ND-ORDER-SELVES! Whatever information is represented in my 2ND-order-selves, I can communicate to you. The is analogous to attention. In trying to predict you and survive, I make a prediction of myself from your perspective. That prediction informs my communication with you.

ATTENTION IS ALL I NEED

If I am speaking to you, the meaning of my words is whatever I intend. You have understood me if you know that by saying x that I am trying to get you to think y. This is what a 2ND-order-self allows agents to do[451]. A 2ND-order-self lets me predict what you think I think. I can use that to predict what you think I am trying to achieve by saying x. Hence if I want you to believe y, I can derive x from my 2ND-order-self. I can simultaneously have many different 2ND-order-selves. I can have one for everyone I talk to. I can have aggregates thereof. I can even construct one from my own point of view. However for the sake of communication what is important is that I am have one from the perspective of the person I am talking to, so I can tailor my words to each of them to convey my intended meaning. Conversely, if I want to know what you mean when you say x, I can derive y from my prediction of your prediction of my prediction of you. My 2ND-order-self's prediction of you.

[451] Michael Timothy Bennett. On the computation of meaning, language models and incomprehensible horrors. In *Artificial General Intelligence*. Springer Nature, 2023c

The 2ND-order-selves neatly encapsulate those things I pay attention to. They are contextual access. They offer a simple explanation of why some information is available to me at some times, and not others. Moreover, it is obviously impossible to communicate anything *not* in 2ND and higher order selves. At least, it is impossible to communicate *meaning* like a human can.

As stated earlier, consciousness is oft cut into access and phenomenal aspects. The phenomenal is the special snowflake. Access is relegated the unglamorous role of "information processing". Bland and unmysterious. However access consciousness is usually formally defined as the information available for reasoning and report. Communication. Presumably, that means communication in the human sense. If I accept that, then I must also accept that the only information available to access consciousness is that within 2ND and higher order selves. In other words, I propose a radically different interpretation of access consciousness than is typically used. I argue this is the only acceptable interpretation if what we are trying to describe is human-like consciousness, with human-like meaning. I will delve into this more deeply later, but for now it is important to make note of this point.

SEMIOTICS. WHAT MEAN?

Organisms that can communicate can co-operate to achieve complex goals. Now that we know *how* intent can be communicated it is easy to see how signalling conventions or language might evolve. To that end, here is a formal definition of organism.

Definition 19 (organism)
I describe the circumstances of an organism[452] $\mathfrak{o}$ *as* $\langle \mathfrak{v}_\mathfrak{o}, \mu_\mathfrak{o}, \mathfrak{p}_\mathfrak{o}, <_\mathfrak{o} \rangle$ *where:*

- $O_{\mu_\mathfrak{o}}$ *contains every output which qualifies as "fit" according to natural selection.*
- $\mathfrak{p}_\mathfrak{o}$ *is the set of policies an organism knows, s.t.* $\mathfrak{p}_\mathfrak{o} \subset \mathfrak{p}_{n.s.} \cup \mathfrak{p}_{\mathfrak{h}_{<t_\mathfrak{o}}}$ *and:*
  - $\mathfrak{p}_{n.s.} \subset L_{\mathfrak{v}_\mathfrak{o}}$ *is* ***reflexes*** *hard coded from birth by natural selection.*
  - $\mathfrak{p}_{\mathfrak{h}_{<t_\mathfrak{o}}} = \bigcup_{\zeta \in \mathfrak{h}_{<t_\mathfrak{o}}} \Pi_\zeta$ *is the set of policies it is possible to* ***learn*** *from a history of past interactions represented by a task* $\mathfrak{h}_{<t_\mathfrak{o}}$.
  - *If* $\mathfrak{p}_{\mathfrak{h}_{<t_\mathfrak{o}}} \not\subset (\mathfrak{p}_\mathfrak{o} - \mathfrak{p}_{n.s.})$ *then the organism has* ***selective memory****. It can "forget" outputs, possibly to productive ends if they contradict otherwise good policies.*
- $<_\mathfrak{o}$ *is a binary relation over* $\Gamma_{\mathfrak{v}_\mathfrak{o}}$ *we call* ***preferences****.*

[452] (intuitive summary) Strictly speaking an organism $\mathfrak{o}$ would be a policy, but we can describe the circumstances of its existence as a task μ that describes all "fit" behaviour for that organism. We can also identify policies the organism "knows", because these are implied by the policy that is the organism. Likewise, we can represent lossy memory by having the organism "know" fewer policies than are implied by its history of interactions. Finally, preferences are the particular "protosymbol" the organism will use to "interpret" an input in later definitions.

For meaning, I must now define what I call a protosymbol system. It is a set of tasks based on the causal-identities an organism has learned.

Definition 20 (protosymbol system)
Assume an organism $\mathfrak{o}$*. For each policy* $p \in \mathfrak{p}_\mathfrak{o}$ *there exists a set* $\mathfrak{s}_p = \{\alpha \in \Gamma_{\mathfrak{v}_\mathfrak{o}} : p \in \Pi_\alpha\}$ *of all tasks for which* p *is a correct policy. The union of all such sets is*

$$\mathfrak{s}_\mathfrak{o} = \bigcup_{p \in \mathfrak{p}_\mathfrak{o}} \{\alpha \in \Gamma_{\mathfrak{v}_\mathfrak{o}} : p \in \Pi_\alpha\}$$

We call $\mathfrak{s}_\mathfrak{o}$ *a "protosymbol system". A* $\mathfrak{v}$*-task* $\alpha \in \mathfrak{s}_\mathfrak{o}$ *is called a "protosymbol".*

An organism has preferences, which are a total order over tasks and thus the protosymbols available to the organism based on its causal identities. This describes how an organism interprets and responds to its environment.

Definition 21 (affect)
Suppose we have two organisms, $\mathfrak{a}$ (Alice) and $\mathfrak{b}$ (Bob). Suppose $\mathfrak{a}$ interprets $i \in L_{\mathfrak{v}_o}$ as an output o, then:

- *a **statement** $v \subset i$ affects $\mathfrak{a}$ if $\mathfrak{a}$ would have interpreted $e = i - v$ as a different output $g \neq o$.*
- *an **organism** $\mathfrak{b}$ has affected $\mathfrak{a}$ by making an output k if, as a consequence of k, there exists $v \subset s$ which affects $\mathfrak{a}$.*

Definition 22 (interpretation)
Interpretation is inference, with the additional step of choosing a policy according to preference. Interpretation is an activity undertaken by an organism $\mathfrak{o}$. It proceeds as follows:

1. *Assume true input $i \in L_{\mathfrak{v}_o}$ (meaning $i = i_t$ in an EGRL system at time t).*
2. *We say that i **signifies** a protosymbol $\alpha \in \mathfrak{s}_\mathfrak{o}$ if $i \in I_\alpha$.*
3. *$\mathfrak{s}^s_\mathfrak{o} = \{\alpha \in \mathfrak{s}_\mathfrak{o} : i \in I_\alpha\}$ is the set of all protosymbols which i signifies.*
4. *If $\mathfrak{s}^s_\mathfrak{o} \neq \emptyset$ then i **means something** to the organism in the sense that there is "value" ascribed to symbols in $\mathfrak{s}^s_\mathfrak{o}$ compelling the organism to act.*
5. *If i means something, then $\mathfrak{o}$ chooses $\alpha \in \mathfrak{s}^s_\mathfrak{o}$ that maximises its preferences $<_\mathfrak{o}$.*
6. *The organism then infers an output $o \in E_s \cap E_{\Pi_\alpha}$.*

Imagine organisms Alice $\mathfrak{a}$ and Bob $\mathfrak{b}$ such that $\mathfrak{p}_\mathfrak{a} = \{c^\mathfrak{b}_\mathfrak{a}, c^\mathfrak{a}_\mathfrak{a}, c^{\mathfrak{b}\mathfrak{a}}_\mathfrak{a}, c^{\mathfrak{b}\mathfrak{a}\mathfrak{b}}_\mathfrak{a}, c^{\mathfrak{b}\mathfrak{a}\mathfrak{b}\mathfrak{a}}_\mathfrak{a} ...\}$ and $\mathfrak{p}_\mathfrak{b} = \{c^\mathfrak{a}_\mathfrak{b}, c^\mathfrak{b}_\mathfrak{b}, c^{\mathfrak{a}\mathfrak{b}}_\mathfrak{b}, c^{\mathfrak{a}\mathfrak{b}\mathfrak{a}\mathfrak{b}}_\mathfrak{b} ...\}$. This means they both have causal-identities for each other, and 1ST, 2ND and 3RD order selves. Assume Alice α means m by choosing output o given input i. By outputting o, Alice intends that:

1. Bob comes to believe m,
2. Bob recognise this intention, and
3. (1) occur on the basis of (2).

To do this Alice interprets i using her protosymbols based on the causal-identities:

1. $c^\mathfrak{b}_\mathfrak{a}$ affords Alice the ability to recognise Bob and predict Bob.
2. $c^\mathfrak{a}_\mathfrak{a}$ affords Alice the ability to reason that she may causally intervene in Bob's existence. $c^\mathfrak{a}_\mathfrak{a}$ underpins subjective experience, acting as reafference does in the human mid-brain[453].
3. $c^{\mathfrak{b}\mathfrak{a}}_\mathfrak{a}$ is Alice predicting that she *is* perceived by Bob, and *how* her behaviour will be interpreted by Bob. This is the minimum requirement to formulate m. Using $c^{\mathfrak{b}\mathfrak{a}}_\mathfrak{a}$, she can predict how $c^\mathfrak{b}_\mathfrak{a}$ will change given different causal interventions, and choose an intervention that causes it to contain what she interprets as m.

[453] Bjorn Merker. The liabilities of mobility: A selection pressure for the transition to consciousness in animal evolution. *Consciousness and Cognition*, 2005. Neurobiology of Animal Consciousness

4. c_a^{bab} is Alice predicting what Bob thinks she thinks of him. This is necessary to know that Bob can interpret her actions as *intending* m. Without this second order prediction of Bob, she cannot know Bob will understand that she intends he understand. She needs to know Bob predicts she intends m for Gricean meaning to work.

5. c_a^{baba} is Alice predicting how she is perceived by Bob in light of the fact that she intends m. This isn't strictly necessary for meaningful communication, but it certainly helps and clearly humans can do it!

This is everything necessary for Alice to play her part. Alice can interact with anything in the world this way, and attempt to communicate with it. Most things will not do this though. Most things do not learn causal-identities. If Bob is a rock, he will not understand. Alice might think he understands, but what Bob does is another matter entirely. Let us consider what Bob needs to have to really understand Alice.

1. c_b^{ab} affords Bob the basic ability to predict m, by predicting what Alice thinks he thinks, and wants him to think.

2. c_b^{abab} is not strictly necessary to infer intent, but it would certainly help with accurate interpretation of meaning.

Protosymbols conform to Peirce's theory of signs, and so this provides a formal basis for various semiotic and linguistic theories. In Peirce's theory, a symbol is composed of a sign, a referent and an interpretant. A sign is that which is interpreted, the referent is what the sign is interpreted as meaning, and the interpretant is the *effect* of the sign upon one who interprets it. This triadic semiosis resembles what I have formalised as a task. A sign is an input, the referent is an output, and the policies used to interpret a sign are interpretants. Here, one sign can be associated with many symbols at once.

Definition 23 (meaning)
*The **meaning** an organism $\mathfrak{o}$ ascribes to an input i is a protosymbol $\alpha \in \mathfrak{s}_\mathfrak{o}$ which $\mathfrak{o}$ uses to interpret i. Symbols in different protosymbol systems can be **roughly equivalent** (result in similar behaviour etc), in accord with the philosophical arguments in the body of the paper*[454]*. We use $\omega \approx \alpha$ to indicate that ω and α are roughly equivalent. In accord with earlier definitions, one organism "intends" to affect another if completion of the protosymbol (a task) the former uses to interpret hinges on how the latter's behaviour is affected.*
Assume an organism $\mathfrak{a}$ (Alice) in input $i_\mathfrak{a}$ and organism $\mathfrak{b}$ (Bob) in $i_\mathfrak{b}$. $\mathfrak{a}$ means $\alpha \in \mathfrak{s}_\mathfrak{a}$ by deciding $u \in E_{i_\mathfrak{a}}$ iff $\mathfrak{a}$ intends in deciding u:

1. *that $\mathfrak{b}$ interpret $i_\mathfrak{b}$ using $\omega \approx \alpha$,*
2. *$\mathfrak{b}$ "recognize" this intention by being affected by u such that the input $i_\mathfrak{b} = j$ in which $\mathfrak{b}$ finds itself change to $i_\mathfrak{b} = k \neq j$,*
3. *and (1) occur on the basis of (2), meaning had $\mathfrak{a}$ not decided u then $\mathfrak{b}$ would have interpreted $i_\mathfrak{b}$ using $\zeta \not\approx \alpha$.*

To communicate meaning organisms must:

1. *be able to affect one another.*
2. *have similar experiences, so $\mathfrak{s}_\mathfrak{b}$ and $\mathfrak{s}_\mathfrak{a}$ contain roughly equivalent symbols.*
3. *have similar preferences.*

[454] We argue organisms of the same species construct roughly equivalent protosymbols, even though each member of a species exists in its own unique abstraction layer with its own protosymbol system.

CO-OPERATION AND MANIPULATION

The ability to communicate intent has a side effect. The possibility of co-operation allows for new sorts of predatory behaviour. For example I can feign co-operative intent, lying to gain advantage. Just as my second order self lets me predict how you will respond so I can make you understand, I can use exactly the same predictive machinery to mislead you. As before, a 2ND order self is all that is needed, but higher orders help. My 3RD order self in particular lets me predict your prediction of my communicative intent, and what you think I'm going to do.

However if you and I can accurately predict each other's intent, then we can preempt any manipulation. This may cancel out the advantage of betrayal. Repeated interaction of systems which can accurately predict one another can foster genuine co-operative intent. The only way to convince you I have co-operative intent if you can accurately predict my intent, is to actually *have* co-operative intent. This is even more the case if we interact repeatedly, because then we have an iterated prisoner's dilemma. This provides some insight in the evolution of empathy and prosocial behaviour[455].

[455] Michael Timothy Bennett. On the computation of meaning, language models and incomprehensible horrors. In *Artificial General Intelligence*. Springer Nature, 2023c

THE MIRROR SYMBOL HYPOTHESIS

The ability to learn is not free[456]. Why waste energy learning something that never changes? It makes more sense to hard-wire invariant behaviours. Organisms are not blank slates. Members of the same species are alike. This information can be used to better predict other members of one's species. Humans generally exhibit similar preferences. It is very common for humans to seek wealth, live in a house, drive a car and so on. There are common human phenotypes. Even outside our species organisms must eat, sleep and procreate. This suggests one should not approach everyone as a unique and special snowflake[457]. That would waste energy. It makes far more sense to just use one's own preferences as a prior to predict everyone, and adjust from there. Humans have a well documented tendency to do exactly this. We exhibit a consensus bias. We tend to believe others believe as we do, and as you might expect this aids in social learning[458]. In terms of my formalism, this means once Alice has learned a basic causal-identity c_a^b for Bob, she can fill in the finer details using her own preferences and protosymbol system. Because these protosymbols are tapestries of valence, they convey both the motive and the interpretation. In my earliest work, I argued there that empathy could be explained in computational terms using "mirror symbols"[459]. These protosymbols are the culmination of that claim, together with my later work on language and learning.

[456] R. Landauer. Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, 5(3):183–191, 1961

[457] If one is concerned only with optimising for resource efficiency.

[458] Tor Tarantola, Dharshan Kumaran, Peter Dayan, and Benedetto De Martino. Prior preferences beneficially influence social and non-social learning. *Nature Communications*, 8(1):817, 10 2017. ISSN 2041-1723. DOI: 10.1038/s41467-017-00826-8. URL `https://doi.org/10.1038/s41467-017-00826-8`. The authors declare no competing financial interests

[459] Michael Timothy Bennett and Yoshihiro Maruyama. Philosophical specification of empathetic ethical artificial intelligence. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2): 292–300, 2022a

THE NORMIES

We know something about each other's intent before we say a word. Accurate prediction and hard-wired behaviour aside, there is a cultural aspect to how we understand one another. Human communication depends heavily on social mores or *normativity*. We generally understand each other because we operate within a common social framework. I would be remiss if I did not at least mention this, though it is largely beyond the scope of this thesis. I have argued that on an individual level, repeated interaction creates an iterated prisoner's dilemma that may foster genuine co-operation in some areas. I have shown how Gricean meaning may be communicated and interpreted. Finally, I have argued it is more efficient to hard-code invariant behaviours than learn them. Our environment is made up mostly of other people. According to active inference we do not just try to learn a model to fit our environment and minimise unwanted surprises. We also modify the environment to fit our models[460]. We experiment, and we try to reduce unpredictability. It isn't much of a leap to then say we enforce social norms in order to make our environment more predictable. These norms allow us to navigate complex social structures at massive scale, far beyond what we could cope with if all we could do if trust had to be established entirely on an individual basis.

Liquid brains are collectives that compute in the sense of the parts moving around (for example, an ant colony)[461]. This is as opposed to a solid brain where the parts (e.g. cells) form a network with a persistent structure that supports information processing at a higher level of abstraction (e.g. bioelectricity on a cellular network)[462]. Liquid brains are characterised by decentralisation, delegation and asynchronous communication. In experimental settings, robots have "evolved" shared syntax and normative meanings to achieve goals[463]. These robots are like a liquid brain. Likewise, a population of humans is a liquid brain. A scale-free collective intelligence[464] that learn policies, like any other self-organising system. Language and normativity are policies of a liquid brain. Conversely, the policies of a liquid brain are like mores and norms that dictate how a population will interpret information both within, and without.

[460] Friston K., FitzGerald T., Rigoli F., Schwartenbeck P., O. Doherty J., and Pezzulo G. Active inference and learning. *Neurosci Biobehav Rev.*, pages 862–879, 2016

[461] Ricard Solé, Melanie Moses, and Stephanie Forrest. Liquid brains, solid brains. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1774):20190040, 2019. DOI: 10.1098/rstb.2019.0040. URL `https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2019.0040`; and Chris R. Reid, David J T Sumpter, and Madeleine Beekman. Optimisation in a natural system: Argentine ants solve the towers of hanoi. *Journal of Experimental Biology*, 214(1):50–58, jan 2011

[462] The persistent structure and centralisation being a feature of solid brains.

[463] Luc Steels. Evolving grounded communication for robots. *Trends in Cognitive Sciences*, 7(7):308–312, 2003. ISSN 1364-6613. DOI: https://doi.org/10.1016/S1364-6613(03)00129-3. URL `https://www.sciencedirect.com/science/article/pii/S1364661303001293`

[464] Chris Fields and Michael Levin. Scale-free biology: Integrating evolutionary and developmental thinking. *BioEssays*, 42, 06 2020; and Patrick McMillen and Michael Levin. Collective intelligence: A unifying concept for integrating biology across scales and substrates. *Communications Biology*, 2024

MUTINY!

If human populations are scale free liquid brains, then there are some very interesting comparisons to be made. This section is based entirely on my papers on abstraction layers[465], as well as the aforementioned papers on meaning, consciousness and communication. I have just argued that social mores and norms are policies learned by the collective. I have also argued in previous chapters that two cells α_1 and α_2 have a collective identity for an organ ω expressed as a policy $\pi \in \Pi_{\alpha_1} \cap \Pi_{\alpha_2} \cap \Pi_{\omega}$. This formalises the sort of multi-scale competency architecture often used to describe biological systems from cells to ecosystems[466]. Collective intelligences. I can describe a population of humans in the same way. A shared polity is a shared policy. Wherever language, identity and ethos are shared, there is a shared policy. Of course we're dealing with a liquid brain here, so the shared identity is repeated in each member of the collective rather than a centralised *self*. An identity can repeat at different scales and levels of abstraction, much like how a pattern repeats at different scales in a fractal[467] (see Figure 8).

[465] Michael Timothy Bennett. Computational dualism and objective superintelligence. In *Artificial General Intelligence*. Springer Nature, 2024a; and Michael Timothy Bennett. Are biological systems more intelligent than artificial intelligence? 2025b. Forthcoming 2026

[466] Patrick McMillen and Michael Levin. Collective intelligence: A unifying concept for integrating biology across scales and substrates. *Communications Biology*, 2024

[467] Michael F. Barnsley. *Fractals Everywhere*. Academic Press, second edition edition, 2012

Figure 8: The Barnsley Fern is a Fractal that exhibits self similarity. Notice how the same pattern repeats at different scales. This gives some visual intuition as to what 'scale-free' suggests, as we can see the same dynamics play out again and again at different scales.

Cancer is a problem of biological self-organising systems. Some have suggested cancer can be understood as a sort of identity crisis. In this view, cancer is what happens when the identity of a cellular collective fails[468]. Cells network. Those collectives *constrain* cells. When cells become isolated from the informational structure of the collective of which they are part, they *lose* that constraint. They can, as a result, revert to primitive transcriptional behaviours. They can become *less* differentiated, because they no longer need to adhere to a specific organistic identity.

In the context of my formalism, a state analogous to cancer gets triggered when the system is 'over-constrained'[469]. Remember the contravariance principle[470] from earlier? If I interpret it to fit with my formalism, it says harder tasks are better to learn from because there are fewer policies to choose from. This is a little bit like that. If the task is impossible there are no correct policies to choose from. If we have distribution but not delegation, and the situation becomes impossible, then the parts of the system will not cease to exist. They will continue to operate, just not as a whole. Intuitively if a company goes bankrupt, the employees go on without it. This is a bit like that.

If I represent a collective as a task, then adversity is represented as that task having fewer correct policies. The system can also make life more difficult for itself by overconstraining its parts and ruling out possible correct policies. Excessive top down control can limit adaptability, and reduce number of correct policies that exist. Whether the cause is externally imposed adversity, or internally imposed top-down control, if the constraints are too tight the result is the same: no correct policies will exist. In the case of a distributed, collective system this means collective identity will be lost. In the sense I have described, a $\mathfrak{v}$-task α is easier than ω if $\mid \Pi_\alpha \mid > \mid \Pi_\omega \mid$. Now, this is not an all encompassing definition of difficulty. It is just how I'm defining it for this particular example. If I have a stack and a lower level $i-1$ changes such that the task at the higher level i has fewer correct policies[471], then there may be no correct policies left at level i. If there are no longer any correct policies at i, then the only way the system can continue to function at that level of abstraction is for part of the collective to break away from the group. Become isolated from the informational structure of the collective. Why? Because that will loosen the constraint on the rest of the system. If enough pieces break away, the collective that remains may not how correct policies to choose from. Meanwhile the parts that break off have complete freedom and can revert to whatever default behaviour they please.

[468] P C W Davies and C H Lineweaver. Cancer tumors as metazoa 1.0: tapping genes of ancient ancestors. *Physical Biology*, 8(1), feb 2011; Michael Levin. Bioelectrical approaches to cancer as a problem of the scaling of the cellular self. *Progress in Biophysics and Molecular Biology*, 2021. Cancer and Evolution; and Patrick McMillen and Michael Levin. Collective intelligence: A unifying concept for integrating biology across scales and substrates. *Communications Biology*, 2024

[469] Michael Timothy Bennett. Are biological systems more intelligent than artificial intelligence? 2025b. Forthcoming 2026

[470] Rosa Cao and Daniel Yamins. Explanatory models in neuroscience, part 2: Functional intelligibility and the contravariance principle. *Cognitive Systems Research*, 85:101200, 2024

[471] i.e. by reducing $\mid \Pi^i \mid$ at the higher level i.

SAY I HAVE A COLLECTIVE $\lambda^i(\mathfrak{v}^i)$ that includes a cell $\alpha \sqsubset \lambda^i(\mathfrak{v}^i)$ where $\mathfrak{f}(E_{\pi_{\mathfrak{a}}^{i-1}}) = \mathfrak{v}_{\mathfrak{a}}^i$. Assume policy $\pi_{\mathfrak{a}}^{i-1}$ changes to $\pi_{\mathfrak{b}}^{i-1}$ s.t. $\Pi_{\lambda^i(\mathfrak{v}_{\mathfrak{b}}^i)} = \emptyset$. The collective is a bust. Parts of it can continue, but not all together. A big chunk may continue functioning as before with a shared policy, but the rest must go it alone. Earlier I mentioned 'selective forgetting'[472] of inputs and outputs, for the purpose reconciling my rather uncompromising binary definitions of correctness with noisy data. Selective forgetting can allow for a hypothesis to be generated that fits most of the data, but not all. I might choose to ignore data inconsistent with an otherwise good policy. To do this I could discard a child task $\alpha \sqsubset \lambda^i(\mathfrak{v}_{\mathfrak{b}}^i)$ of $\lambda^i(\mathfrak{v}_{\mathfrak{b}}^i)$ when $\Pi_{\lambda^i(\mathfrak{v}_{\mathfrak{b}}^i)} = \emptyset$, giving me a task $\lambda^{i'}(\mathfrak{v}_{\mathfrak{b}}^i) \sqsubset \lambda^i(\mathfrak{v}_{\mathfrak{b}}^i)$ where $\alpha \not\sqsubset \lambda^{i'}(\mathfrak{v}_{\mathfrak{b}}^i)$ and $\lambda^{i'}(\mathfrak{v}_{\mathfrak{b}}^i)$ has correct policies. This means we now have some correct policies and $\Pi_{\lambda^{i'}(\mathfrak{v}_{\mathfrak{b}}^i)} \neq \emptyset$, but α is off on its lonesome, isolated from the informational structure of the collective. α, doing what it wants, is now our cancer analogue. It can now pursue goals which do not align with those of the collective.

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

THIS IS NO SURPRISE GIVEN The Law of the Stack. Adaptability at higher levels depends on adaptability at lower levels. Graceful degradation demands an approach like Mission Command. Delegate control to the maximum extent possible *while* maintaining the core functionality you want, and your system will be in the best possible position to *maintain* that core functionality. Why make that goal more difficult than necessary by burdening the system with unnecessary constraints? A system that has fewer constraints imposed top-down has more leeway to cope with inputs that restrict the set of correct policies. For example, Bob who is not wearing a straight jacket is able to do more things than Bob who is wearing the straight jacket.

LANGUAGE CANCER

In biological systems that can support bioelectric signalling, cancer occurs when cells become disconnected from that informational structure. Bioelectricity can be seen as cognitive glue[473]. However in a liquid brain there is not a persistent structure that can support bioelectric signalling. Instead, information may be transmitted through sound, writing, light or any number of other means. Parts of the system *affect* one another, and in doing so learn to *interpret* one another. If the iterated game of interaction favours co-operation, the parts of the system will converge on shared policies[474]. These policies, in the case of humans, are languages and concepts. We develop signalling conventions and protocols. Technologies. Ethical and social norms. These are the policies of the liquid brains that form humanity. Organisations, cultures and so forth. Just as a biological system can have a higher level collective identity, so too can a human population. Like a biological collective, a human collective can *lose* that identity. To lose its various orders of *selves.* There are so many anecdotal parallels. Languages can diverge. Organisations can fail if circumstances grow too constrained. Parts of the population can simply splinter off and fail to engage, and the collective must develop an immune system, like a police force, to ensure the collective remains functional. However, too rigid a polity and we have too constrained a system. We end up with stagnation. Excessive reliance on top-down control and the system will be prone to fail and develop cancer just as if it were placed under excessive external stress.

[473] Ben Lyons and Michael Levin. Cognitive glues are shared models of relative scarcities: The economics of collective intelligence. *Manuscript*, 2024

[474] Michael Timothy Bennett. On the computation of meaning, language models and incomprehensible horrors. In *Artificial General Intelligence*. Springer Nature, 2023c

PRECONDITIONS OF NORMATIVITY

What is needed is 'sloppy fitness'[475]. Loose but still *sufficient* constraints. That leaves room for shared language, meaning, ethics and norms to develop[476]. In the context of artificial intelligence, this means we should take a delegated and scale-free approach to alignment. Whether it be the internal functioning of an artificial intelligence, or the human system in which we deploy it, the same principles apply. If we want the system to retain an identity, that is the same thing as having a language of shared meaning between its parts. If we want the system to retain a coherent *language*, we need to balance top-down control with bottom-up control. We need to delegate adaptation to the lowest level possible, whilst ensuring what is in place top-down is sufficient.

Rubber banding in videogames is a great example of how this might be done, because it is simple. A racing game remains fun if it remains competitive. The failure state is if the players get frustrated and leave[477]. A rubber banding mechanism penalises the lead player and assists the lagging player, to keep the game competitive[478]. So how might we apply this in AI? Agential systems based on large language models are very prone to losing their identities. We can measure this[479] and apply rubber banding techniques to ensure multi-agent systems of language model agents retain their individual and collective identities.

[475] Salvatore J. Agosta, Niklas Janz, and Daniel R. Brooks. How specialists can be generalists: resolving the "parasite paradox" and implications for emerging infectious disease. *Zoologia (Curitiba)*, 27 (2):151–162, Apr 2010. ISSN 1984-4670. doi: 10.1590/S1984-46702010000200001. URL https://doi.org/10.1590/S1984-46702010000200001

[476] Michael Timothy Bennett and Yoshihiro Maruyama. Philosophical specification of empathetic ethical artificial intelligence. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2): 292–300, 2022a; and Michael Timothy Bennett. Symbol emergence and the solutions to any task. In *Artificial General Intelligence*. Springer Nature, 2022a

[477] Speaking as a former game designer, from a game publisher's point of view this might as well be cancer.

[478] Qingwei Mi and Tianhan Gao. Adaptive rubber-banding system of dynamic difficulty adjustment in racing games. *ICGA Journal*, 44(1):18–38, 2022

[479] Elija Perrier and Michael Timothy Bennett. Position: Stop acting like language model agents are normal agents, 2025. URL https://arxiv.org/abs/2502.10420

SCALE-FREE ALIGNMENT

Ultimately, a conscious machine faces the same challenges any system does. It must retain a coherent identity if it is to exist in any meaningful way, and I have over the last several chapters described what is needed to do that. Interestingly, this seems relevant to the problem of alignment in artificial intelligence. Safety is tangential to this thesis but deserves a mention. If we take the idea that embodiment is a statement of value to its logical conclusion, the orthogonality thesis[480] is clearly wrong[481,482].

[480] Eliezer Yudkowsky et al. Orthogonality thesis. `https://www.lesswrong.com/w/orthogonality-thesis`, 2025. Wiki page from LessWrong with multiple contributors. Accessed: 2025-03-18

[481] Michael Timothy Bennett. Lies, damned lies, and the orthogonality thesis. *Preprint*, 2025d

[482] The orthogonality thesis is the idea that goals and intelligence are independent

Theorem 9 *Intelligence is not independent of goals.*

Proof 9 *Assume $\mathcal{C}$ is a space of software programs, Γ is a space of behaviours a system can exhibit, $f_1 \in \mathcal{C}$ is a software mind and $f_2 : \mathcal{C} \to \Gamma$ is a hardware body. It interprets f_1. Finally $\mathbb{G}$ is the set of environments, which here are functions $f : \Gamma \to \{0, 1\}$. We single out $f_3 \in \mathbb{G}$ as the environment in which goals are pursued. If I am the engineer who build the software AI f_1 for some particular purpose, then I am part of its environment. Goals are satisfied if $f_3(f_2(f_1)) = 1$. I will now show why intelligence is not independent of embodiment, and embodiment is not independent of goals.*

1. *Suppose we are given f_3. For every environment there exists hardware s.t. $f_3(f_2(\cdot)) = 1$ regardless of software (it maps all software to the same behaviour, or in other words it is hard-wired to satisfy the goals in f_3).*
2. *For every choice of f_1 given f_3 alone, there exists f_2 s.t. $f_3(f_2(f_1)) = 0$.*
3. *This means intelligence is not independent of embodiment[483].*
4. *Now I'll show goals are not independent of embodiment. Assume we are given a fixed $f_2(f_1)$.*
5. *We can choose f_3 so that $f_2(f_1)$ is optimal and the goals in f_3 are satisfied.*
6. *However, $f_2(f_1)$ determines which choices of f_3 are optimal. $f_2(f_1) = \gamma$ constrains us to choices of f_3 where $f_3(\gamma) = 1$[484].*
7. *This means goals are not independent of embodiment.*

As intelligence hinges on embodiment, and embodiment is goal directed, intelligence is inevitably goal directed. □

[483] Similar points are made in related work

Michael Timothy Bennett. The optimal choice of hypothesis is the weakest, not the shortest. In *Artificial General Intelligence*. Springer Nature, 2023a; Michael Timothy Bennett. Is complexity an illusion? In *Artificial General Intelligence*. Springer Nature, 2024b; and Jan Leike and Marcus Hutter. Bad universal priors and notions of optimality. *Proceedings of The 28th Conference on Learning Theory, in Proceedings of Machine Learning Research*, pages 1244–1259, 2015

[484] Meaning only those environments and goals where a particular behaviour or phenotype γ will succeed.

Alignment is generally framed as tailoring the AI to respect legal and moral boundaries. My results suggest this is making life harder than it needs to be. We should take a whole of system approach, and tailor not just the AI but the systems with which it interacts to accommodate beneficial outcomes. Retaining a coherent collective identity is a problem that might face many other sorts of systems too, but I digress. What matters for my purposes is primarily the biological and computational systems.

XI. WHY IS ANYTHING ALIVE?

Over the past several chapters I have sought to explain how and why our subjective world is divided up into the objects and properties that it is. Why a chair is a chair, and not half a chair. How we simplify our surroundings into causal-identities that afford us something. In this chapter I will explore how the stack constructs systems which have this ability. Then I will explore their limitations. In particular, I will tie my theory together with an explanation of the origins of life, and attempt to resolve the Fermi Paradox. It is based primarily on my early paper on language and the Fermi Paradox[485], my complexity paper[486] and my more recent work on abstraction layers[487]. The latter two in particular are relevant to the formation of adaptive, homeostatic, goal directed systems.

[485] Michael Timothy Bennett. Compression, the fermi paradox and artificial super-intelligence. In *Artificial General Intelligence*. Springer Nature, 2022b

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

[487] Michael Timothy Bennett. Are biological systems more intelligent than artificial intelligence? 2025b. Forthcoming 2026

BOUNDARIES

The Free Energy Principle (FEP) provides an elegant explanation of how systems maintain their integrity[488]. How the boundary between the system's internal and external worlds. How a system maintains homeostasis and minimises surprisal[489]. Some have said my formalism does not account for a system's boundary. Specifically, it was alleged my formalism incorporates *"hierarchical predictive coding with attentional control without, apparently, localizing the experience of attentional control to any boundary, whether external or internal"*[490]. I disagree with this characterisation, so to begin I'll explain why and how my formalism *does* in fact localise attentional control within a boundary, and why it is compatible with the FEP. I would say the theories are complementary. My focus is less on the details of *how* homeostasis is maintained, and more on why homeostatic systems exist in the first place.

[488] Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010; and Karl Friston. Life as we know it. *Journal of The Royal Society Interface*, 10(86):20130475, 2013. DOI: 10.1098/rsif.2013.0475. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2013.0475

[489] Suprisal is a measure of how novel something is. Negative log probability. That sort of thing.

[490] Chris Fields, Mahault Albarracin, Karl Friston, Alex Kiefer, Maxwell JD Ramstead, and Adam Safron. How do inner screens enable imaginative experience? applying the free-energy principle directly to the study of conscious experience. *Neuroscience of Consciousness*, 2025

First, the experience of attentional control for access consciousness is explicitly located in those 2ND-order-selves which are being used to interpret inputs in a given time. Information in those 2ND-order-selves is available for reasoning and report, and everything else is not. That is the boundary. These selves are statements, and like any collection of statements they imply an abstraction layer. That abstraction layer expresses everything within the boundary.

Second and more generally speaking, an abstraction layer represents all the potential configurations of a bounded system. A bounded system can contain only finite information, which is why vocabularies are finite[491]. We can represent the possible configurations *within* a boundary as a vocabulary. Second, we can easily represent the free energy principle within this framework, by taking $\langle I, O \rangle$ where O is the set of free-energy minimising behaviours. We can also represent maintenance of a boundary as a task, whose correct policies govern the continued existence of said boundary. We can then apply an abstractor function to the policy of that task, to get the abstraction layer *within* that boundary. In other words, the task of maintaining a boundary creates an abstraction layer at the level above.

[491] Jacob D. Bekenstein. Universal upper bound on the entropy-to-energy ratio for bounded systems. *Phys. Rev. D*, 23: 287–298, Jan 1981

A boundary is an interpreter. Abstraction layers are small worlds inside big worlds, like human language within the scope of human behaviours[492]. My whole thesis is an attempt to explain the formation, interaction and operation of such layers. Far from boundaries being an afterthought, they are the core of Stack Theory.

[492] Ramon Ferrer i Cancho and Ricard Solé. The small world of human language. *Proceedings of the Royal Society B: Biological Sciences*, 268(1482):2261–2265, 2001. DOI: 10.1098/rspb.2001.1800

Using my formalism I can also represent the integrity of a collective system. For example I can frame a collective of cells within abstraction layer as a collection of tasks, that can then in turn imply a higher level abstraction layer. If those cells cannot align and develop a collective policy then there is no policy at the higher level of abstraction. The vocabulary is still there while the matter is still there, but the system is not maintaining its integrity. The collective splits up. This is a flexible depiction of collective identity. It allows for *nested* identities within larger collectives[493]. It allows for the changing or dissolution of identity or boundary. Most importantly, it explains the existence of boundaries in an unmediated, abstraction free universe from first principles. To my knowledge, there is no other formalism that does this. This is probably the key novel contribution of my theory, because it tells us how we get to coherent objects, properties and systems from the fundamental fact of change. Time is difference, each difference is a state, and the fact of change is a cosmic ought from which all other oughts descend. From there we get aspects, a stack of abstraction layers, tasks and policies. I can change whether a correct policy exists at one level of abstraction by changing policy at the level below. This represents whether a collective can maintain its integrity and thus boundary, or not.

[493] Anna Ciaunica, Evgeniya V. Shmeleva, and Michael Levin. The brain is not mental! coupling neuronal and immune cellular processing in human organisms. *Frontiers in Integrative Neuroscience*, 2023

THE ORIGINS OF LIFE

What really matters though is internal modelling the external environment through a boundary. Minimising surprisal. Not every abstraction layer does this. The ability to store and process information is a fundamental feature of complex living systems[494]. So why would a system emerge which maintains such a boundary? In other words, why would a living system emerge? More to the point, what does my formalism say about this? Well, consider that the environment preserves that which preserves itself. Because we have a spatially extended universe we have finite vocabularies, because you can only cram so much matter into so much space before it collapses. We also have potentially infinite delegation of control to lower levels of abstraction. The end result of this is that w-maxing gets correlated with simp-maxing. However, they are not the same thing as the experiments clearly show. In stable, *static* abstraction layers like the computer I used to run those experiments, it easy to construct a scenario where simp-maxing actually *prevents* w-maxing[495]. W-maxing occurs *within* an abstraction layer, and the weakness of a policy may be just as subjective as its complexity, but the consequence is objectively very different from simp-maxing. It is possible to w-max without simp-maxing.

[494] Ricard Solé et al. Fundamental constraints to the logic of living systems. *Interface Focus*, 2024

[495] As proof 3 demonstrates.

In practical terms, this is because a system can exhibit behaviours such as self-repair. This preserves the system, w-maxing while also increasing complexity. It is the exact *opposite* of simp-maxing. Likewise learning allows an organism to complete a wider variety of tasks than an organism which does not[496]. Learning w-maxes the system. The ability to store and retrieve information in internal memory is needed for learning, and such an ability once again *increases* complexity. It does the *opposite* of simp-maxing. I can continue to list examples, but what I am getting at here is that we may measure intelligence, homeostasis and the existence of a boundary by the extent to which a system w-maxes without simp-maxing.

[496] All else being equal.

Consider a rock. It is very simple. It has one permanent internal state. The vocabulary of the rock's internal state may be just one program. Framed as a policy in the absence of abstraction, the rock may be extraordinarily weak. It might persist in many states. It does depend on a particularly specific or stable abstraction layer, meaning we can take a rock and put it in space, or in earth's atmosphere, and either way it is fine. It is simp-maxed, and because we live in a universe where weak constraints tend to take simple forms this means it also w-maxed to some extent. Simp-maxing does not necessitate w-maximisation, but it certainly makes it a great deal easier! This may be why most of our environment tends to be simple[497].

Now consider a slime mould. It can spread and collect resources, to persist. It can adapt within constraints placed upon it, to solve mazes and other complex problems[498]. To persist, it adapts within the constraints placed upon it. Within the abstraction layer of its local environment, and within itself and the abstraction layer that controls morphology. It is not a simple as a rock. Because of this, it is comparatively fragile. Yet within the confines of a stable environment, slime mould is able to persist. It might not be as persistent as a rock at a low level of abstraction, but it can exhibit a far greater variety of behaviour at a higher level of abstraction. The rock doesn't *do* anything. The slime mould is more *adaptable* at the higher level of abstraction even though it is less persistent in general, because it can *do more* within the narrow confines of its very stable environment. In the end, what matters to evolution is that something persists, not the counterfactuals like whether would it persist as well if we were move everything somewhere more inhospitable. Abstraction layers *causally isolate* an environment, like the Markov blanket used in FEP based formalisms[499]. Abstraction layers *also* account for representational power[500] in the broader context of an infinite stack of abstraction layers. When the slime mould navigates a maze to find the shortest path is performing a computation. However it is *not* the same sort of bioelectrical computing a human might. It is computing via a search using its *shape*[501]. According to my formalism, it is computing at a lower level of abstraction. That may be why it can solve such incredibly complex problems[502] whilst remaining so very simple.

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

[498] T. Nakagaki, H. Yamada, and A. Toth. Maze-solving by an amoeboid organism. *Nature*, 407(6803):470, 2000

[499] Karl Friston, Lancelot Da Costa, Dalton A.R. Sakthivadivel, Conor Heins, Grigorios A. Pavliotis, Maxwell Ramstead, and Thomas Parr. Path integrals, particular kinds, and strange things. *Physics of Life Reviews*, 47: 35–62, 2023. ISSN 1571-0645. DOI: https://doi.org/10.1016/j.plrev.2023.08.016. URL `https://www.sciencedirect.com/science/article/pii/S1571064523001094`

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

[501] Ricard Solé and Luís F Seoane. Evolution of brains and computers: The roads not taken. *Entropy*, 24(5):665, 2022

[502] Ricard Solé and Luís F Seoane. Evolution of brains and computers: The roads not taken. *Entropy*, 24(5):665, 2022

Now think of an ant colony. By imposing the right boundary conditions, an ant colony can be made to solve the towers of hanoi[503]. The colony displays much more intelligent behaviour than any one of its parts, and the parts become less intelligent the more top-down control is exerted upon them as the colony is scaled up[504]. For example, in smaller colonies ants may display more intelligent individual behaviour. In computational terms, the ants are a concurrent and distributed computing system. They communicate *asynchronously* across time and space. When the collective is damaged[505], the remaining ants will repurposed themselves to maintain the correct ratio of ants performing particular roles. There is no *centralised* controller. The ant colony is a *liquid* brain because the ants move around, rather than forming a network with a persistent structure[506]. Liquid brains are characterised by decentralisation, delegation and *asynchronous* communication. The ant colony liquid brain[507] is not as simple as the slime mould, but it can support adaptation at a higher level of abstraction. It is more specialised for the static abstraction layer that is its environment, and more fragile because of it.

Humans have solid brains, and human populations are liquid brains. A solid brain actually delegates control to a *lesser* extent than a liquid brain. It must exert top down control to maintain a strict form. On the other hand, this allows for *synchronous* message passing, rather than the asynchronous message passing of the ant colony. A degree of centralisation facilitates reafference and helps with navigation[508]. Shouldn't that mean the human is less adaptable than the ant colony or the slime mould? Well, yes! It is in fact much harder to keep a human brain working than it is an ant colony. If we were to delete half an ant colony, the remaining ants would repurpose themselves to keep the colony functioning. However if we delete almost any part of a human brain it ceases to function. Comparatively, it is maladaptive. However it is perhaps *because* of this brittleness that it can support a such a high level of abstraction. It is the right balance of *stable* so that it can support complexity, and yet it is incredibly well *adapted* to the human environment at very high levels of abstraction. It is "correct" according to natural selection within *that* environment. The human stack is so well adapted it can dedicate vast resources to maintaining the stability of its own stack, shaping its environment to be even more stable. It can support the formation of complex collective policies at even higher levels of abstraction, forming liquid brains of massive scale unified by abstract narratives that leverage the 2ND and 3RD-order-selves of humans within the population. The ant colony might be less fragile generally, but a human is better adapted to the human world.

[503] Chris R. Reid, David J T Sumpter, and Madeleine Beekman. Optimisation in a natural system: Argentine ants solve the towers of hanoi. *Journal of Experimental Biology*, 214(1):50–58, jan 2011

[504] Daniel W. McShea. A complexity drain on cells in the evolution of multicellularity. *Evolution*, 56(3):441–452, 03 2002. ISSN 0014-3820. DOI: 10.1111/j.0014-3820.2002.tb01357.x. URL https://doi.org/10.1111/j.0014-3820.2002.tb01357.x; and Jordi Delgado and Ricard V. Solé. Collective-induced computation. *Phys. Rev. E*, 55:2338–2344, Mar 1997. DOI: 10.1103/PhysRevE.55.2338. URL https://link.aps.org/doi/10.1103/PhysRevE.55.2338

[505] For example if we delete all the ants performing a certain role.

[506] The persistent structure and centralisation being a feature of solid brains.

[507] Explained in earlier chapters, a liquid brain is one in which the parts move around. For example, an ant colony. Liquid brains use asynchronous communication. In contrast, a solid brain has a persistent structure that allows it to support things like bioelectric signalling that is synchronous.

[508] Bjorn Merker. The liabilities of mobility: A selection pressure for the transition to consciousness in animal evolution. *Consciousness and Cognition*, 2005. Neurobiology of Animal Consciousness

HOMEOSTASIS

Does this mean a computer is smarter than a human? After all, the very criticism I have levelled at computers is that they do not delegate adaptation to low enough levels of abstraction. A living body can adapt at a much lower level than a computer, because a computer doesn't adapt *at all* at lower levels of abstraction. A human body *does* adapt, and it has been refined over millennia of natural selection. The homeostatic *self-maintenance* requirements of being alive for an extended period demands a level of delegation a computer with a static abstraction layer cannot manage. The problem is with a computer is that we have not just stopped it adapting at a higher level of abstraction. It is that computers are abstraction layers that encode our human biases and understanding at a very high level of abstraction. Computers are *limited* by that. It is no wonder the AI systems we develop are so brittle and maladaptive. We have *dropped Hume's guillotine* to cut off their representations from the valence that motivated them. We have foolishly treated representations as platonic. We have treated goals and representations are independent, when they are not[509].

[509] Michael Timothy Bennett. Lies, damned lies, and the orthogonality thesis. *Preprint*, 2025d

Computers do not actively maintain a boundary. They are not homeostatic, and they do not usually minimise surprisal. This isn't to say they won't one day do this. For example, homeostatic soft robots are a theorised class of robots which pursue self-regulatory goals[510]. If they maintain homeostasis, then they must self-repair. To self-repair, they must delegate adaptation to very low levels of abstraction. This demands a sort of artificial life, with small parts interacting form complex systems from the small scales and low levels of abstraction up[511]. Were a homeostatic soft robot to be constructed of adaptive material, such as self organising nanites[512], then it could construct stacks highly specialised to the problems for which it is employed. Then, it might be as generally intelligent as a human.

[510] Kingson Man and Antonio R. Damasio. Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence*, 1:446 – 452, 2019. URL `https://api.semanticscholar.org/CorpusID:208089594`

[511] Keisuke Suzuki and Takashi Ikegami. Spatial-pattern-induced evolution of a self-replicating loop network. *Artificial Life*, 12(4):461–485, 2006; and Takashi Ikegami and Keisuke Suzuki. From a homeostatic to a homeodynamic self. *Biosystems*, 91(2):388–400, 2008

[512] Francesca Borghi, Thierry R. Nieus, Davide E. Galli, and Paolo Milani. Brain-like hardware, do we need it? *Frontiers in Neuroscience*, 18, 2024; and B. Paroli, G. Martini, M.A.C. Potenza, M. Siano, M. Mirigliano, and P. Milani. Solving classification tasks by a receptron based on nonlinear optical speckle fields. *Neural Networks*, 166: 634–644, 2023

THE LAW OF INCREASING FUNCTIONAL INFORMATION

The maintenance of a boundary at a particular level of abstraction is a part of that, but it isn't the whole picture. What matters is the stack, and weak constraints on function within the bounds of a static abstraction layer. A theory with some similarities was put forward by Wong et. al.[513]. There are synergies between our theories that I will build on, but before I get into those I must preempt any concerns about the novelty of my thesis. This being a PhD thesis, the onus is on me to provide evidence that my thesis is novel. Their results were published three years after my theory first appeared in peer reviewed books and journals[514], and one and a half years after my experimental and proof results regarding weakness were submitted for review and uploaded to a preprint server[515], and 6 months after the experiment and proof results were published in a peer reviewed book[516]. I must emphasises that though there are similarities, our research results are distinct and complementary. With that said, I will now explain the theory of Wong et. al. and prove a law they propose.

Instead of weakness, Wong et. al. attempt to formalise functional information[517]. Their definition appears strikingly similar to weakness, and it even uses some of the same notation. However their formalism leaves a great deal open to interpretation and it is not accompanied by proofs or experiments, as I will discuss below. They present an interesting philosophical argument, and a proposed natural law. They suggest that all systems, including life, are composed of components they can recombine into various functional configurations. Systems are naturally selected, in that some persist and some do not. They then identify three means by which systems persist: static persistence, dynamic persistence and novelty generation. Functional complexity is inverse to the number of configurations that achieve a function. They then propose that the passage of time drives evolution to select for increasing functional complexity. They begin by imagining a "possible world" with a low entropy state that increases in accord with the second law of thermodynamics. In that possible world there are no local, stable pockets of low entropy and so entropy just increases monotonically. In comparison, our world has barriers that create pockets of low entropy that stop everything just heading straight for equilibrium.

[513] Michael L. Wong, Carol E. Cleland, Daniel Arend, Stuart Bartlett, H. James Cleaves, Heather Demarest, Anirudh Prabhu, Jonathan I. Lunine, and Robert M. Hazen. On the roles of function and selection in evolving systems. *Proceedings of the National Academy of Sciences*, 120(43):e2310223120, 2023. doi: 10.1073/pnas.2310223120. URL https://www.pnas.org/doi/abs/10.1073/pnas.2310223120

[514] Michael Timothy Bennett. Symbol emergence and the solutions to any task. In *Artificial General Intelligence*. Springer Nature, 2022a; and Michael Timothy Bennett and Yoshihiro Maruyama. Philosophical specification of empathetic ethical artificial intelligence. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):292–300, 2022a

[515] Michael Timothy Bennett. Computable Artificial General Intelligence. *Under Review*, 2022c; and Michael Timothy Bennett. Technical appendices, 2025a. URL https://github.com/ViscousLemming/Technical-Appendices

[516] Michael Timothy Bennett. The optimal choice of hypothesis is the weakest, not the shortest. In *Artificial General Intelligence*. Springer Nature, 2023a; Michael Timothy Bennett. A formal theory of optimal learning with experimental results. *IJCAI*, 2025f; and Michael Timothy Bennett. Emergent causality and the foundation of consciousness. In *Artificial General Intelligence*. Springer Nature, 2023b

[517] Jack W. Szostak. Functional information: Molecular messages. *Nature*, 423(6941):689–689, 06 2003. ISSN 1476-4687. doi: 10.1038/423689a. URL https://doi.org/10.1038/423689a

THEY THEN IDENTIFY ORDERS OF SELECTION. These are categories into which objects fall that describe how they persist. The first order is static persistence. This categorises aspects of the environment that persist by being unchanging. They are stable against the forces of decay, meaning the path of least resistance is for them to stay put rather than be incorporated into more stable configurations of matter. Each instance of static persistence is a "battery of free energy"[518]. Next they identify *second order* persistence, where objects are dissipative, autocatalytic or homeostatic. This is like the self repair function I spoke of. Finally, they propose that there are systems which persist through invention new functions. Each one supports the next.

[518] I. Prigogine and R. Lefever. Theory of dissipative structures. In H. Haken, editor, *Synergetics*, pages 124–135. Vieweg+Teubner Verlag, 1973

THESE SOUND LIKE AN ABSTRACTION LAYER in my formalism. Statically persistent translates to a static stack. I will interpret statically persistent systems as being those that simp-max, and this makes it very easy for such systems to fall into configurations that support weak policies and thus higher levels of abstraction. Dynamically persistent systems are abstraction layers that rest on static abstraction layers, within which systems can w-max without simp-maxing. These include basic solid and liquid brains. The third order proposed by Wong et. al. is novelty generating systems, which I interpret as being those that have a solid brain of sufficient scale and inceptive to support the formation of selves and higher order representational contents.

WHERE WE DIVERGE, HOWEVER, is in the exact formal definitions of these ideas. I'll summarise their formalism below, using slightly different notation to avoid confusion, and then compare it to mine:

1. Functional information is determined with respect to a particular function x, where function means a system together with a context.

2. They use F_x to denote a quantitative measure of a configuration's ability to perform function x. They also call this the "degree of function". It appears to be a positive real number, but this isn't explicitly said.

3. $M(F_x)$ would be the number of configurations with a "degree of function" *greater* than F_x. It appears to be a positive integer.

4. N would be the total number of possible configurations, for example if we are dealing with an n bit string then $N = 2^n$. It is almost certainly a positive integer.

5. Function information $I(F_x) = -\log_2(\frac{M(F_x)}{N})$, a real number.

6. Greatest functional information is attained by F_{max}, smallest by F_{min}.

Now I'll try to translate their ideas into analogous definitions within my formalism. Once again I must reiterate that my theory was published in peer reviewed books and journals before the above. The formal notation and some of the ideas are strikingly similar, but there are important differences that make these distinct research results. I am doing a lot of interpretation here to make their definitions fit within my formalism. I argue there is leeway to do this because their theory leaves many mathematical details unspecified. Perhaps for this reason, it is also not accompanied by proofs or experiments. Nevertheless they put forward a compelling philosophical argument that fits my formalism. They propose the "law of increasing functional information", which is that the functional information of systems will increase with time. They don't actually prove this law, so I can build on their work by doing so. Assuming, of course, that one buys my rather subjective interpretation of their concepts.

I translate the concepts of Wong et. al.[519] into Pancomputational Enactivism as follows:

1. Their function x sounds like it could be a $\mathfrak{v}$-task α in vocabulary $\mathfrak{v}$. The possible futures given past α would be a parent ω of α.

2. They use F_x to denote a quantitative measure of a configurations ability to perform function x. I would say a "configuration" here is a policy π, and F_x sounds a bit like the cardinality of the extension $|E_\pi|$ of a policy π. The weaker π is, the greater its ability to perform function x, which is the probability that it generalises to the parent task ω of α.

3. They use $M(F_x)$ to be the number of configurations with a degree of function *greater* than F_x. This sounds like the number of correct policies which are *weaker* than π. I will interpret this as meaning the cardinality the set $|M_\pi|$ such that $M_\pi = \{\pi_{alt} \in \Pi_\alpha : |E_\pi| \geq |E_{\pi_{alt}}|$, meaning M_π has at least 1 member.

4. N would be the total number of possible configurations, for example if we are dealing with an n bit string then $N = 2^n$. I will interpret this as being the total number of possible statements which can be made in the abstraction layer $L_\mathfrak{v}$, so $|L_\mathfrak{v}| = N$.

5. Functional information $I(F_x) = -\log_2(\frac{M(F_x)}{N})$ would then be a function of weakness $I(\pi) = -\log_2(\frac{|M_\pi|}{|L_\mathfrak{v}|})$, and functional information is attained by w-maxing π.

[519] Michael L. Wong, Carol E. Cleland, Daniel Arend, Stuart Bartlett, H. James Cleaves, Heather Demarest, Anirudh Prabhu, Jonathan I. Lunine, and Robert M. Hazen. On the roles of function and selection in evolving systems. *Proceedings of the National Academy of Sciences*, 120(43):e2310223120, 2023. DOI: 10.1073/pnas.2310223120. URL https://www.pnas.org/doi/abs/10.1073/pnas.2310223120

If one accepts my interpretation of functional information as being a function of weakness, then I can prove the law of increasing functional information. It follows trivially from the fact that w-maxing is necessary and sufficient to maximise the probability of generalisation.

Theorem 10 (The Law of Increasing Functional Information) *The functional information in a system will increase with time, where definitions are as above. Assume systems are policies. Assume ω represents future selection pressures on policies, and policies which generalise to ω from α are those which persist into the future.*

Proof 10 *From proofs 1 and 2 we have that w-maxing is sufficient to maximise the probability of generalising from α to ω. Therefore the policies which generalise will be the weakest. W-maxing maximises functional information, so functional information must increase with time.* □

So to reiterate, boundaries are formalised in the stack of abstraction layers. From The Law of the Stack we have that adaptation at high levels of abstraction depends on adaptability at low. From The Law of Increasing Functional Information we have that functional information increases. Static abstraction layers form in pockets of free energy. These are are selected for *static* persistence. Within static abstraction layers, it is possible to w-max without simp-maxing resulting in *dynamic* persistence that can support fragile solid brains, that can support complex novelty generation. This may be why complex life emerges. Stacks within a stable, static abstraction layer can grow tall, complex and above all very *specific*. Stacks within constantly changing abstraction layers will change more frequently, and in doing so more frequently delete statements which are overly specific to a particular stack's highest levels of abstraction.

Spatial constraints maximise the simplicity of forms (simp-max) while the transition of the environment from one state to another weakens constrains on function (w-maxes). A rock could be seen as adaptable because it persists through many states of the environment. However if this is adaptation then the rock is adaptable through pure simp-maxing. A human is adaptable within an ecological niche, through w-maxing without simp-maxing. It is not adaptable *outside* that ecological niche because it is not simp-maxed, so overall it is less adaptable. However from within its environmental niche, a human can be more adaptable than a rock.

THIS IS WHY A STABLE ENVIRONMENT can support adaptable yet complex forms at high levels of abstraction. This resolves the apparent contradiction between complex forms being less probable, and life being complex.

EARTH IS AN ABSTRACTION LAYER. Lifeforms are statements made in that abstraction layer, growing ever more specific with each additional layer of abstraction. There are inefficiencies, and stable environments permit more inefficiencies, allowing organisms to become *more* adaptable in the areas of high variability where they actually need to be adaptable, at the cost of those they don't. The vocabulary they embody can be specialised to express as many *relevant* policies as possible that are as weak as possible, at the cost of not being able to express policies outside of that scope. For example, humans are very versatile within earth's atmosphere, but outside of it we struggle. Natural selection has hard coded static adaptations to stable aspects of Earth, like breathing air and tolerance of a certain temperature range. Intelligence or dynamic adaptation can then be exclusively focused on highly variable aspects of life on earth, such as predation or socialisation. This creates an incentive for even higher level novelty generation, language and so forth. The construction of selves. That is why humans are so good predicting other humans and playing status games. Stable conditions at low levels of abstraction permit extreme specialisation in higher levels of abstraction. We go from adaptations hard-wired by natural selection, to the hard-wired ability to learn new adaptations, to the construction of 1ST and higher order selves. This, I argue, is why life exists as it does on earth. The universe preserves structures which preserve themselves. In most of the universe, this means just structures like rocks. However in an environment like earth that is stable at a high level of abstraction, this also includes complex adaptive systems which could not exist outside that environment.

THE FERMI PARADOX

This last part of the chapter is from my paper on the Fermi Paradox[520]. It argues that intelligent entities may be all around us, but it does not fall within the scope of what we can notice from atop our lofty stacks. Entities that do not meet the scale and incentive preconditions for a causal-identity will not be recognised. Highly abstract behaviour for which we cannot construct a causal-identity will tend to appear as random noise, because it is so specific to another stack. In the interests of defending my thesis's novelty I must once again point out that although a very similar result has been published by others[521], that happened three years after my explanation of the Fermi Paradox first appeared in a peer reviewed book[522], two years after I published a sequel[523], and one year after I uploaded a preprint talking at great length about the preconditions for causal-identities[524]. I say this merely to anticipate and address in advance criticism of my thesis's novelty. More importantly, they begin from different premises. If anything, the fact that someone else came up with the same result from different premises only goes to support my claims.

From the subjective perspective of an abstraction layer high in the environmental stack[525], whether something exists is a matter of whether there is a causal-identity for it. What I perceive depends on my stack. The scale and incentive preconditions must be met for an organism to construct a causal-identity for an object. Likewise *within* collective systems, these preconditions must be met for the parts of that system in relation to one another. They must be able to affect one another. Both organisms and ecosystems are collective, so we might say that to meet the incentive precondition two entities must be part of the same collective. Organisms must be part of the same ecosystem, and cells must be part of the same organism. I am not suggesting cells possess the cognitive machinery to construct a causal-identity. I am merely saying they can be subject to the incentive. By tying incentive together with being part of a collective, I am repeating what others have said, which is that in such system the same dynamics repeat at different scales as we *zoom out*, so to speak[526]. We can *scale up* such scale-free, fractal concepts from one level of abstraction to the next. Alice and Bob must be in some sense part of the same collective system if the scale and incentive preconditions are to be met such that an organism Alice constructs a causal-identity for Bob. There must be something *relevant* for Alice about the existence of Bob. The causal-identity must *afford* Alice something[527].

[520] Michael Timothy Bennett. Compression, the fermi paradox and artificial super-intelligence. In *Artificial General Intelligence*. Springer Nature, 2022b

[521] Chris Fields and Michael Levin and. Life, its origin, and its distribution: a perspective from the conway-kochen theorem and the free energy principle. *Communicative & Integrative Biology*, 18 (1):2466017, 2025

[522] Michael Timothy Bennett. Compression, the fermi paradox and artificial super-intelligence. In *Artificial General Intelligence*. Springer Nature, 2022b

[523] Michael Timothy Bennett. On the computation of meaning, language models and incomprehensible horrors. In *Artificial General Intelligence*. Springer Nature, 2023c

[524] Michael Timothy Bennett, Sean Welsh, and Anna Ciaunica. *Why Is Anything Conscious?* Preprint, accepted to and presented at ASSC27 and MoC5, 2024

[525] e.g. a human.

[526] Chris Fields and Michael Levin. Scale-free biology: Integrating evolutionary and developmental thinking. *BioEssays*, 42, 06 2020; and Michael F. Barnsley. *Fractals Everywhere*. Academic Press, second edition edition, 2012

[527] James J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979

When an object or property does afford us something, we humans tend to anthropomorphise it. We ascribe intent and desire to inanimate objects like the sun, trees and mountains. We do this because our survival largely depends on interaction with humans and other animals. For the sake of efficiency, humans are hard-wired to predict an environment made up 'things like us'. The most efficient way to do that is to use your own motives as a template for how other objects are going to behave[528]. The same thing that allows us to communicate causes us to anthropomophise everything.

On the other hand, humans disregard information that is not relevant to our survival. We don't even notice anything outside our collective niche. This is obvious in the case of our senses and other hard-wired adaptations. We see and hear in the ranges we do because that is what natural selection has demanded of us. We don't see and hear outside those ranges. Our bodies are abstraction layers that simplify the 'big world' of the environment down to the 'small world' of objects and properties that matter[529]. We then engage in short term, in-context 'intelligent' adaptation by learning and adapting throughout our lives. We do not construct causal-identities for anything that does not cause valence. We might construct a *neutral* concept that elicits nothing, like the colour red, but the colour red describes many things which *do* cause valence. Natural selection shapes us to respond to what is relevant to our survival. Hence, there is no reason for us to waste our very finite resources processing irrelevant information. To us, objects and properties which are irrelevant simply do not exist. They are outside of the scope of what we notice or comprehend.

Close enough is good enough. Weak constraints might take simple forms, but only to the extent that it matters *within* our ecological niche. Within our abstraction layer. We can hold absurd beliefs, so long as they don't actually kill us. Things are simple when I can efficiently represent them within my abstraction layer, by which I mean the causal-identity is a short statement. Things are complex when I cannot represent them efficiently using my abstraction layer. When that happens, either causal-identities are then long statements, or the scale precondition is not met and there may be multiple overlapping and incomplete causal-identities for one object instead of single coherent classification. A "hall of mirrors" composed of many incomplete representations.

[528] Michael Timothy Bennett and Yoshihiro Maruyama. Philosophical specification of empathetic ethical artificial intelligence. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2): 292–300, 2022a

[529] John Vervaeke, Timothy Lillicrap, and Blake Richards. Relevance realization and the emerging framework in cognitive science. *J. Log. Comput.*, 2012; John Vervaeke and Leonardo Ferraro. *Relevance, Meaning and the Cognitive Science of Wisdom*. Springer Netherlands, Dordrecht, 2013a; John Vervaeke and Leonardo Ferraro. Relevance realization and the neurodynamics and neuroconnectivity of general intelligence. In Inman Harvey, Ann Cavoukian, George Tomko, Don Borrett, Hon Kwan, and Dimitrios Hatzinakos, editors, *SmartData*, NY, 2013b. Springer Nature; and Johannes Jaeger, Anna Riedl, Alex Djedovic, John Vervaeke, and Denis Walsh. Naturalizing relevance realization: Why agency and cognition are fundamentally not computational. *Frontiers in Psychology*, 15, 2024

For example, language models seem prone to constructing such halls of mirrors. They can cope with of arithmetic using short numbers, but fail when given long numbers[530]. This suggests they construct many incomplete representations of arithmetic, but fail to synthesize a solitary and sufficiently weak interpretation to explain all instances of it. Hence the behaviour of systems which do not satisfy the scale and incentive preconditions would appear like random noise[531]. A highly compressed signal we could not decode. At the limit, highly compressed signals are indistinguishable from noise.

I am limited to what my stack affords me. Constructing a rationale amounts to inferring a causal-identity to predict behaviour[532]. I cannot construct a rationale for behaviour that is so unlike my own that I cannot infer or represent the pattern. Behaviours are only comprehensible if one has the decoder. This is a possible resolution to the Fermi Paradox. The ability to recognise intelligent behaviour depends on two entities occupying two similar stacks. Alice and Bob might diverge at a level of abstraction, like slime mould and a solid-brained human, and one might still "recognise" or at least detect the other to some extent. However to what extent can this be maintained? How much distance between stacks can be tolerated before Alice and Bob are isolated and unable to form even the most basic of collective informational structure? We might recognise a language model as intelligent because it is engineered to mimic the exact behaviour we consider relevant. They are engineered to be useful and *afford* us something. However, from the language model's perspective we are like the unknowable cosmic horrors of Lovecraftian fiction[533]. We not only adapt at much lower levels of abstraction, but we control the abstraction layers in which the model exists. This is like if we puny humans encountered an entity that could control physics, that created us and could turn us off if it gets bored. There is a directionality to the ability to recognise and understand intelligence. This has implications for AI safety. If we want an AI to have human-like values and motives, then we must ensure the values we want it to embody meet the scale and incentive preconditions within its stack. It is not just a matter of aligning policy, but the system as a whole including its vocabulary.

[530] Michael Timothy Bennett. On the computation of meaning, language models and incomprehensible horrors. In *Artificial General Intelligence*. Springer Nature, 2023c

[531] Michael Timothy Bennett. Compression, the fermi paradox and artificial super-intelligence. In *Artificial General Intelligence*. Springer Nature, 2022b

[532] Michael Timothy Bennett. Symbol emergence and the solutions to any task. In *Artificial General Intelligence*. Springer Nature, 2022a

[533] Michael Timothy Bennett. On the computation of meaning, language models and incomprehensible horrors. In *Artificial General Intelligence*. Springer Nature, 2023c

XII. WHY IS ANYTHING CONSCIOUS?

Here I address the hard problem. This chapter will repeat the contents of my papers on consciousness[534], with a dash of Stack Theory. It is a philosophical chapter. The argument of course hinges on the math in the previous chapters, but this chapter contains none. Note that this chapter seeks to explain what consciousness is and why it has evolved. It is thus focused on *evolved* organisms, rather than human built machines. I tackle conscious machines in chapter XIII.

I'll begin with the environment and work my way in to describe *what* a conscious mind is. This explains *why* anything is conscious. I argue qualia can be reduced to valence, and valence follows from change.

Earlier I spoke of Higher Order Thought theories[535]. They are an attempt to explain why we are conscious of some information but not other information. Where the light-switch of consciousness goes on, and what is in the light. They argue that the information of which we are consciously aware are really higher order meta representations of lower order states. Representations of representations of representations.

Lower order or *local* states are things like 'the smell of coffee' or 'the colour red' arise[536]. They are also known as qualia. Given where we are in the thesis I hope the problem with HOTs is now obvious. They assume an abstraction layer. Qualia are an abstraction layer. Non-reductive physicalists say qualia cannot be broken down into smaller components. This is like saying qualia are the lowest abstraction layer: the bottom of the stack. A fun idea, but it goes nowhere. If AI research traded the pineal gland for a Turing machine, then non-reductive physicalism seems to have given up at "there is software". I am a reductive physicalist who has the gall to think he can break qualia down into something more basic[537].

[534] Michael Timothy Bennett. Emergent causality and the foundation of consciousness. In *Artificial General Intelligence*. Springer Nature, 2023b; and Michael Timothy Bennett, Sean Welsh, and Anna Ciaunica. *Why Is Anything Conscious?* Preprint, accepted to and presented at ASSC27 and MoC5, 2024

[535] Richard Brown, Hakwan Lau, and Joseph E. LeDoux. Understanding the higher-order approach to consciousness. *Trends in Cognitive Sciences*, 23(9):754–768, 2019. doi: 10.1016/j.tics.2019.06.009

[536] Anil Seth and Tim Bayne. Theories of consciousness. *Nature Reviews Neuroscience*, 2022

[537] To say this in more philosophical language: I defend a reductive physicalist position that holds qualia can be broken down into something more basic. As previously observed, each abstraction layer has its own vocabulary.

So that is what I will do. I begin with valence and explain how that amounts to qualia. Then I will explain how there is a self to be subject to that qualia when an organism learns a 1ST-order-self. Then I will argue the organism only becomes consciously *aware* and has access consciousness when it constructs 2ND-order-selves. Finally I will discuss the 3RD-order-self, self-consciousness and internal narratives. I hypothesise that with further increasing orders of self an organism may become *more* conscious.

RED IS A TAPESTRY OF VALENCE

IN MY PAPERS ON CONSCIOUSNESS I start with the embodied organism. Here though I have the luxury of having described The Stack. I have described how organisms emerge from The Stack, and the cosmic ought which motivates them. Hence I can jump ahead to valence. When a system is just attracted or repelled, I call this 'one dimensional' valence. A cell is attracted or repelled, for example. This might be achieved with a vocabulary of just a few programs. Two for an input and output that sends the organism scuttling along one axis, and two for an input and output that makes it scuttle the other way. Such a vocabulary cannot construct a causal-identity for any object. It is just a 'one dimensional' attraction or repulsion.

BUT WHAT IF I RINSE AND REPEAT along another axis for perpendicular movements? Now I have have a richer vocabulary. I have two dimensions of valence. My system can be attracted or repelled in two ways at the same time. I can keep adding dimensions, and learn and adapt. I can be like slime mould and compute through my movements[538]

[538] T. Nakagaki, H. Yamada, and A. Toth. Maze-solving by an amoeboid organism. *Nature*, 407(6803): 470, 2000; and Ricard Solé and Luís F Seoane. Evolution of brains and computers: The roads not taken. *Entropy*, 24 (5):665, 2022

MORE GENERALLY, WHAT IF I TAKE SOMETHING like this simple system and copy it? For example, I take a cell and I copy it. What if I then network these two cells, to form a collective solid-brain? What if their network supports a higher level of abstraction, like a bioelectric information processing? They are not just physically moving around anymore. Now, they support the bioelectric equivalent of a telegraph line. It is state supported by a shape. It is a higher level of abstraction that can support more configurations or states. More dimensions of valence. Using this telegraph network parts of the system can communicate quickly and synchronise. This stands in contrast to a language of pure shape and movement, where signalling must be ponderously slow and asynchronous.

I CAN KEEP SCALING UP A VOCABULARY by increasing the number of possible physical configurations that can be realised within a physical system. I can end up with a vast orchestra of cells playing a symphony of valence. Each part is affected by the present state of the in its own way. Some are attracted, some repelled.

However, the *overall* system is also attracted or repelled. This is because I am dealing with a polycomputer[539]. It is concurrent, distributed, multiscale and multilayered. The same matter is involved in more than one computation at the same time. So at a low level of abstraction where we have this swarm of small parts, there is this many-dimensional valence all happening concurrently. However it is also true that the collective as a whole is attracted or repelled by a physical state. It has the one dimensional valence, at least in the sense that it will *move*. The same is true of everything in between.

[539] Joshua Bongard and Michael Levin. There's plenty of room right here: Biological systems as evolved, overloaded, multi-scale machines. *Biomimetics*, 8(1), 2023

Now, it is possible that I might have a sort of half liquid brain. The system might be made of disconnected, independently operating solid brained modules that sense, learn and act locally. Modules locked together in a fixed structure, but not *integrated* or fully *connected* with each other. A box jellyfish is like this[540]. It can learn simple behaviours, but it isn't efficient. This is sort of like a liquid brain, but less flexible. Each member of the 'population' here is one of those independently operating modules. Each member of the population[541] has to learn the a lesson before they can act upon it in unison. This is inefficient, because the same behaviour must be redundantly learned by each module[542]. This inefficiency could be overcome if the modules could just wire together and somehow share information, but making every part of a body connect directly to every other part is hard. Infrastructure takes up space. We can't have wires everywhere, because this isn't the back of your television! If I want want the system to learn more efficiently and construct more generalisable policies, I likely need to introduce a degree of centralisation in the higher level of abstraction. In these aforementioned telegraph lines.

[540] Jan Bielecki, Sofie Katrine Dam Nielsen, Gosta Nachman, and Anders Garm. Associative learning in the box jellyfish tripedalia cystophora. *Current Biology*, 2023

[541] In this case the aforementioned modules.

[542] Though such redundancy can of course make the system more resilient. It depends on the amount of damage it is expected to tolerate I suppose.

So what happens if I wire these lines into a centralised core or two? Well then I can have a cell being attracted or repelled, a whole collective being attracted or repelled, and each sub-collective being attracted or repelled. I can exert more top-down control and do it faster. I can have planning and co-ordination take place once, in one of these cores. Like the brain's connective core, which appears to do something like hierarchical planning[543]. Within the narrow confines of circumstances that preserve this solid brain, the system will then be able to adapt more efficiently. It is both more fragile in the sense of being tied to a particular persistent network structure, and more adaptable in the sense of being able to better leverage that structure to seek attractive states.

[543] Murray Shanahan. The brain's connective core and its role in animal cognition. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 367:2704–14, 10 2012. DOI: 10.1098/rstb.2012.0128

MOST IMPORTANTLY, NOW WE HAVE THE POSSIBILITY of synchronising across a vast network spread throughout a body. The whole and the parts all have valence, and it all takes place simultaneously. A vast tapestry of valence. Valence is all there is. There are no categorical variables like "the colour red" or "food" yet, just attraction to or repulsion away from physical states. These tapestries of valence become policies, because the system *reacts*. They are not platonic representations we would *apply* valence to like an abstract label. Instead, there is a tapestry of valence that is triggered by the presence of food, and the system is attracted. That tapestry is a classifier, but not in the platonic, abstract sense. There are no neutral platonic representations in this system. Everything is made of valence.

TAPESTRIES OF VALENCE ARE POLICIES. They both impel and classify. Causal-identities can be tapestries of valence that classify a particular cause of valence. Causal-identities classify a relevant object or property of objects. This is why our subjective worlds are made up of the objects and properties they are. A colour like red might *seem* neutral because it is a causal-identity associated with both attractive and repulsive causal-identities. A sort of meta-identity. However it is still inevitably made up attractive and repulsive forces, even if it does not evoke attraction or repulsion at the level of the whole body. Not every tapestry affects everything equally. "Red" has a different quality to "green" because they activate different parts of the sensory system. Thirst and hunger might have the same overall intensity, but they have different *qualities* because they are different tapestries of valence. At the lowest levels of abstraction, they involve different parts of the body.

SO TO REITERATE, at the highest level there is just one dimensional of the whole body. You might ask "where is the categorical variable that is being classified as attractive or repulsive?". Well, that is what the tapestry is for! The overall system has valence, but the parts also have valence. "Hunger" is a causal-identity, and a causal-identity is a tapestry of valence. These valences do not take place in sequence, but altogether as bodily states. This is the 'tapestry of valence' of which I spoke. What I call "quality" is just a tapestry of valence. If a quality is attractive, that just means at the highest level of abstraction it attracts the organism.

In this way, the unmediated states of reality are clustered together into causal-identities denoting whatever *causes* valence. These causal-identities are both classifiers *and* labels. There is, and I cannot emphasise this enough, no reason to discriminate between the two. At the lowest levels of abstraction the causal-identity is the many parts of the system all being attracted and repelled. At the highest level of abstraction, the whole body is attracted, or repelled. That "label" or "reward" is just the highly abstracted version of the same thing happening below. By this point it shouldn't really be a surprise. I am just reiterating the scale-free, fractal nature of everything else I've described in this thesis. The same dynamics play out again and again at different scales and levels of abstraction. For a computer scientist, this is very counterintuitive. We are used to dealing with concepts like key value pairs, where we organise databases by labelling and mapping data. But this is not a database. The data *is* valence, so there is no need to label it. There are no platonic representations here. The system is just attracted or repelled, and as we scale it up that simple valence gets more complicated. I call this *integrated representation and value judgement*.

WHAT IT IS LIKE

STILL, I DON'T AS yet have anything that will *feel* these tapestries of valence. That will be *subject* to them. Just a system that is attracted or repelled, classifying the world as it goes. This is where centralisation once again rears its monolithic head. To navigate a complex world I must have a 1ST-order-self, to discriminate between what I caused, and what I did not[544]. That 1ST-order self is a tapestry of valence that is part of every intervention I make. It is not a label *applied* to interventions I make, but a tapestry of valence that is necessarily correlated every intervention. For that to be the case, I need some degree of centralisation so the same part of the body can be activated with every intervention[545]. I have concluded this from a mathematical perspective, from first principles. If I am right then then we should be able to find a neurological equivalent of a 1ST-order-self. Indeed we can! In humans, this happens in the mid-brain, and it is called reafference. Happily for me, others have already proposed reafference as the core of subjective experience[546,547]. They proposed this for different reasons, but consider that my fallback argument if you find the following unconvincing.

THE 1ST-ORDER-SELF IS A CAUSAL-IDENTITY. It is at least partly hard-wired into all complex life-forms because it is so useful. Even flies have reafference[548]. If the organism is planning, then the predictions of what attractive and repulsive forces will accompany a plan involving predicting the effect on the first order self.

I HOLD THIS IS WHERE PHENOMENAL CONSCIOUSNESS BEGINS. A 1ST-order-self is both *subject to* tapestries of valence, and may itself be at least in part a tapestry of valence. It is subject to valence in the sense that it accompanies every intervention an organism makes. It is where the organism's *agency* begins, because it is what allows the organism to differentiate between what *it* causes, and what other things have caused. Each intervention the organism makes is like an add-on to its 1ST-order-self. There is the tapestry of valence associated with the intervention, and 1ST-order-self is always part of that.

[544] Since everything is causal identities, this amounts to causal identities causing causal identities causing causal identities. The Peircean protosymbols are useful to think of here, because Pearce described how each symbol links to others.

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

[546] Bjorn Merker. The liabilities of mobility: A selection pressure for the transition to consciousness in animal evolution. *Consciousness and Cognition*, 2005. Neurobiology of Animal Consciousness

[547] I proposed this causal-identity for self in an early paper in order to explain how an optimal agent learns the equivalent of a Pearlean `do` operator, and ended up arguing it explains the evolution and formation of a conscious self from a mathematical perspective. After I published that paper one of my advisors pointed out that Merker had earlier also come to the same conclusion that causality and consciousness were connected. This led me to conclude the first order self and reafference were pointing at the same thing. That the same conclusion has been reached by two entirely different avenues of investigation speaks to its merit.

[548] Andrew B. Barron and Colin Klein. What insects can tell us about the origins of consciousness. *Proceedings of the National Academy of Sciences*, 2016

Without a 1ST-order-self, there is nothing to link these interventions together. No common element. Nothing that can play the role of a persistent identity. Being part of all interventions, being the core of agency, I propose that this is what "feels" when the body is attracted or repelled. As an organism learns, this 1ST-order-self can further develop, integrating learned aspects based on causal interaction with the world. It is at least in part a tapestry of valence that has been shaped by the organism's entire history. It persists through every moment of agency. Being a tapestry of valence, it has a quality, and that quality accompanies everything the organism does. The 1ST-order-self is a concrete answer to Nagel's famous question of "what is it like" to be an organism[549].

[549] Thomas Nagel. What is it like to be a bat? *Philosophical Review*, 1974

THE SECOND ARROW

It is one thing to feel, and another thing entirely to know that one feels. If I feel pain but do not know it, then I am simply repelled. It is not really pain, is it? No, I must *know* that I feel to truly feel. I might not have a name for what I feel, but the feeling must intrude on my awareness to exist in any meaningful sense[550]. So this brings us to awareness. Typically, access consciousness has been defined as that which is available for reasoning and report. The raw facts of which I am aware. It is considered the *easy* problem of consciousness. I could not disagree more strongly. Access consciousness is by far the harder and more interesting part! I will now develop my earlier arguments about the 2ND-order-selves.

[550] If you are a non-human organism you won't have a name, a label or any language to describe what you feel.

Some might say access consciousness is just whatever data is there in storage. On a hard drive, for example... and you may ask yourself, "what is available for *me* to reason and report?". I certainly don't know all the data stored in the physical configuration of my body. I know maybe one or two things at a time, and they are very contextual. Clearly access consciousness is more about attention, but is attention enough? What exactly is attention in my framework?

Attention is whatever is in the 2ND-order-selves I am predicting given my current surroundings. If access consciousness is reasoning and report, then the key is in the word *report*. I can only communicate information that is in my second order selves. It is impossible for me to communicate meaning otherwise, at least in the Gricean sense. Humans communicate in the Gricean sense, so that is the only reasonable standard to set since we're talking about consciousness. I predict your prediction of me, and only because of that can I infer what to do to change what you think I intend so that it *becomes* what I mean. There is more though. The 2ND-order-self in one swipe explains a number of capabilities associated with consciousness:

1. attention: Because 2ND-order-selves are predicted based on surroundings, they explain why some things come to my attention and others do not. In the language of higher order thought theory[551], 2ND-order-selves formalise the higher order meta-representations of lower order local states.

[551] David M. Rosenthal. *Consciousness and Mind*. Oxford University Press UK, New York, 2005; and Richard Brown, Hakwan Lau, and Joseph E. LeDoux. Understanding the higher-order approach to consciousness. *Trends in Cognitive Sciences*, 23(9):754–768, 2019. doi: 10.1016/j.tics.2019.06.009

2. MEANING: A 2ND-order-self is absolutely necessary to communicate meaning. It is the difference between unconsciously sending a signal, and communicating meaning as a human does. There is very obviously no other possible way to achieve this. Basic deception is possible here.

3. ANTHROPOMORPHISM: If I am seeing myself through my surroundings, then I am essentially treating everything around me as if it has an opinion on me. Along with using our own causal-identities as priors for others for the sake of efficiency, this explains why humans have their well documented tendency to anthropomorphise and ascribe intent to inanimate objects like rocks, the weather and the sun[552].

4. SOCIAL SELF IMAGE: It functions like Hooley's Looking Glass Self, explaining the well documented relationship between self image and social context[553]. I see myself through the eyes of the world around me. If I do not invent and interpret the world according to a fictional causal-identity through which to predict my second order selves, then I will have to rely on those people around me. I will be more affected by the views of those around me as I predict myself through their eyes, to better predict them.

5. SELF AWARENESS: It explains self knowledge, and how one comes to have it. I need a 1ST-order-self before I can feel, and a 2ND-order-self before I can know that I feel. I can't know that I'm feeling pain if I do not already know that I exist. Conversely this is where it becomes possible to reason about what might happen in my absence. For example, to be able to form plans that involve the world continuing after my death. A 2ND-order-self is necessary to form such plans. It is the point at which knowledge of one's own mortality becomes possible.

UNLIKE A 1ST-ORDER-SELF WHICH IS UNIQUE, an organism would have many 2ND-order selves. However, there is no reason an organism could not then cluster together parts of these many 2ND-order-selves, to form generalised self images. The weaker the 2ND-order-self, the more generalised it is. The weakest 2ND-order-self an organism has might be considered its ego.

[552] Fritz Heider and Marianne Simmel. An experimental study of apparent behavior. *The American Journal of Psychology*, 57(2): 243–259, apr 1944. URL https://www.jstor.org/stable/1416950; Nicholas Epley, Adam Waytz, and John T. Cacioppo. On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4): 864–886, 2007. DOI: 10.1037/0033-295X.114.4.864. URL https://doi.org/10.1037/0033-295X.114.4.864; and Esmeralda G. Urquiza-Haas and Kurt Kotrschal. The mind behind anthropomorphic thinking: attribution of mental states to other species. *Animal Behaviour*, 109:167–176, 2015

[553] Sarah A. Fricke and Christina M. Frederick. The looking glass self: The impact of explicit self-awareness on self-esteem. *Inquiries Journal*, 9(10), 2017. URL http://www.inquiriesjournal.com/articles/1711/the-looking-glass-self-the-impact-of-explicit-sel Accessed: 2025-05-03

NARRATIVE

THIS BRINGS US TO 3RD-ORDER-SELVES. With a 1ST-order self, I have the ability to represent causal interventions and simple plans, like navigating an environment to obtain food. With a 2ND-order-self I gain the ability to see myself as if through another's eyes, to have theory of mind and communicate meaning. I am self-aware. With a 3RD-order-self I become aware that I am self-aware. I can now plan interactions in which I communicate, and predict the responses of the different parties involved. Importantly, this is where I gain the ability to deceive in complex ways. I can now reason about someone else's prediction of the intent behind my intended meaning.

SHANAHAN ARGUED THE BRAIN'S CONNECTIVE core facilitates hierarchical planning[554]. If that is the case, then it could support plans involving the interaction of varying orders of self, and causal-identities for others. It would in effect support an internal narrative or screenplay. What is interesting to consider is that, because a tapestry of valence does not separate the value of a thing from the thing itself, every aspect of this plan must be felt. Something of the sensations which are predicted must be experienced as they are predicted. It is not just an inner narrative, but an *impelling* narrative.

[554] Murray Shanahan. The brain's connective core and its role in animal cognition. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 367:2704–14, 10 2012. DOI: 10.1098/rstb.2012.0128

SO 3RD-ORDER-SELVES ALLOW FOR complex impelling narratives and the awareness that one is aware. It is where the sort of inner-screenplay of the human mind takes place. When a plan is formed in a human mind, we anticipate feelings and sensations and in doing so we experience something of them. We remember similar sensations. 3RD-order-selves, being tapestries of valence, may offer some insight into why. We are not interpreters of platonic representations, but systems impelled to preserve ourselves. What we call representations are just a means to that end.

WHAT I FIND MOST COMPELLING about this is that by simply scaling up the system and adding incentives, I get phenomenal consciousness, meaning a self awareness, and then this impelling narrative. I don't need additional assumptions. It all just falls out of the axioms.

THE PSYCHOPHYSICAL PRINCIPLE OF CAUSALITY

To summarise, an organism has a 1ST-order-self for phenomenal consciousness, and 2ND-order-selves for access, self-awareness and so on. Qualia are not fundamental, nor are objects and properties. These are biological constructs high in the stack. What is fundamental is change, and from that cosmic ought I get a selection pressure that preserves systems that seek to preserve themselves. To do that, systems must be attracted to some parts of the environment and repelled from others. Qualia are tapestries of valence in such systems when they have 1ST-order-selves to feel, and 2ND-order-selves to know that they feel. When an organism sees food and is attracted to it, this causes it to interpret the world according to the dictates of valence. An organism is impelled by tapestries of valence. It will simply react, for example salivating in anticipation of the food. Consciousness is something an organism *does*, rather than *is*. Each tapestry of valence has a different "quality" because what we call quality is just different parts of the body being activated. Hunger and thirst might have the same overall intensity because they have the same valence at the highest level of abstraction, but different in quality because they have different valence at lower levels of abstraction. Different parts of the body are involved, impelling the organism to serve different homeostatic goals.

The Psychophysical Principle of Causality is that systems which preserve themselves fill a contentless world with objects and properties that cause valence. Qualia are just aspects of causes of valence. They feel like something, because these objects and properties are not categorical variables with valence attached to it after the fact. They are tapestries of valence. The 1ST-order-self means there is something to subjectively experience that qualia, 2ND-order-selves are where that something knows it is experiencing the qualia, and 3RD-order-selves let it know that it knows it feels and thus plan complex social interactions.

If the hard problem of consciousness is *why* anything is conscious instead of all information processing going on in the dark[555], then the answer is simply because it is more efficient to be conscious than not. It allows for more efficient adaptation. If the hard problem is why we don't just have access consciousness instead of phenomenal and access, it is because we misunderstood access consciousness in the first place. Access requires phenomenal consciousness. There are no representations without evaluations.

[555] David Chalmers. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 1995; D. J. Chalmers. *The conscious mind: In search of a fundamental theory*. Oxford University Press, 1996; and Ned Block. On a confusion about a function of consciousness. *Brain and Behavioral Sciences*, 1995

If a system is not impelled to do anything, it represents nothing. Perhaps when we build a computer we "drop Hume's guillotine"[556] to cut off highly abstract representations from the valence that motivated them, but are those representations really there or are we just anthropomorphising the behaviour of a physical system we constructed to mimic our behaviour? The idea of a zombie is a conceit born of language. Language is a high level abstraction layer that rests on a stack of lower level abstraction layers. Representational content is something we *ascribe* to the computer, but the computer itself does not have representations. It is simply a physical system, and software is nothing more than the state of hardware. Representational content does not exist absent a conscious mind to ascribe meaning to it.

[556] Michael Timothy Bennett. Are biological systems more intelligent than artificial intelligence? 2025b. Forthcoming 2026

I have formalised consciousness using a formalism of every conceivable world. This ties intelligence to consciousness and shows one cannot be achieved without the other. If a zombie is a copy of a conscious person that behaves exactly like that person but is not conscious, then a zombie is impossible in every conceivable world.

THE EVOLUTION OF CONSCIOUSNESS

Definition 24 (levels of consciousness)

1. *an organism that acts but does not learn, meaning* $\mathfrak{p}_\mathfrak{o}$ *is fixed from birth.*
2. *an organism that learns, but* $\mathfrak{o}^1 \notin \mathfrak{p}_\mathfrak{o}$ *either because* $\mathfrak{o}^1 \notin L_{\mathfrak{v}_\mathfrak{o}}$ *(failing the "representation precondition") or because the organism is not incentivised to construct* $\mathfrak{o}^1$ *(failing the "incentive precondition").*
3. *reafference and phenomenal or core consciousness are achieved when* $\mathfrak{o}^1 \in \mathfrak{p}_\mathfrak{o}$ *is learned by an organism as a consequence of attraction to and repulsion from statements in* $L_{\mathfrak{v}_\mathfrak{o}}$.
4. (a) *access or self reflexive consciousness is achieved when* $\mathfrak{o}^2 \in \mathfrak{p}_\mathfrak{o}$.
   (b) *hard consciousness*[557] *is achieved when a phenomenally conscious organism learns a second order self (an organism is consciously aware of the contents of second order selves, which must have quality if learned through phenomenal conscious).*
5. *meta self reflexive consciousness (human level hard consciousness) is achieved when* $\mathfrak{o}^3 \in \mathfrak{p}_\mathfrak{o}$.

[557] Piotr Boltuc. The engineering thesis in machine consciousness. *Techné: Research in Philosophy and Technology*, 2012

I'LL NOW GIVE EXAMPLES of systems at varying levels of consciousness[558].

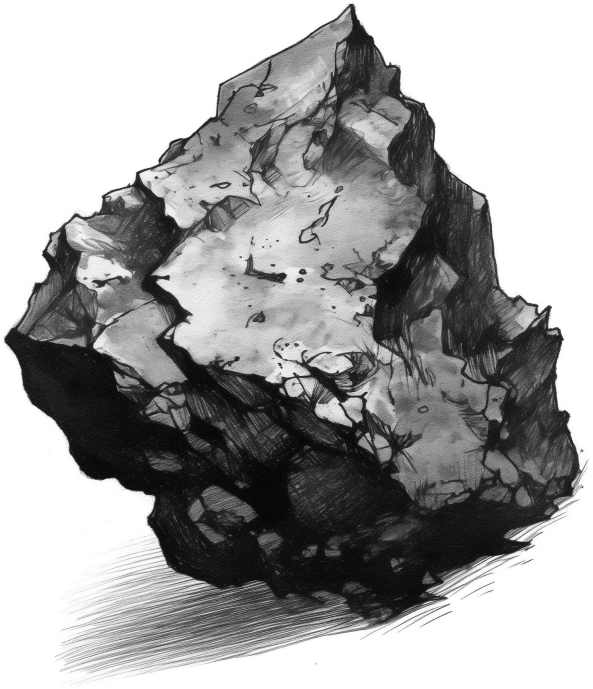
[558] Michael Timothy Bennett, Sean Welsh, and Anna Ciaunica. *Why Is Anything Conscious?* Preprint, accepted to and presented at ASSC27 and MoC5, 2024

- 0: INERT systems which do not act to preserve themselves.
- 1: HARD-WIRED lifelong adaptations.
- 2: LEARNING systems which can acquire new adaptations during their lifetimes. The ability to learn is a hard-wired adaptation.
- 3: 1ST-ORDER-SELF allowing causal reasoning.
- 4: 2ND-ORDER-SELF allowing self awareness and the ability to communicate meaning.
- 5: 3RD-ORDER-SELF for an impelling narrative, meta-self-awareness and complex plans.

STAGE 0

INERT SYSTEMS WHICH DO NOT ACT to preserve themselves are the foundation on which everything is built. There is no consciousness at this level. At least, there is no consciousness as I have described it. This is because there is no valence. This is at odds with the panpsychist position that consciousness is a fundamental building block of reality.

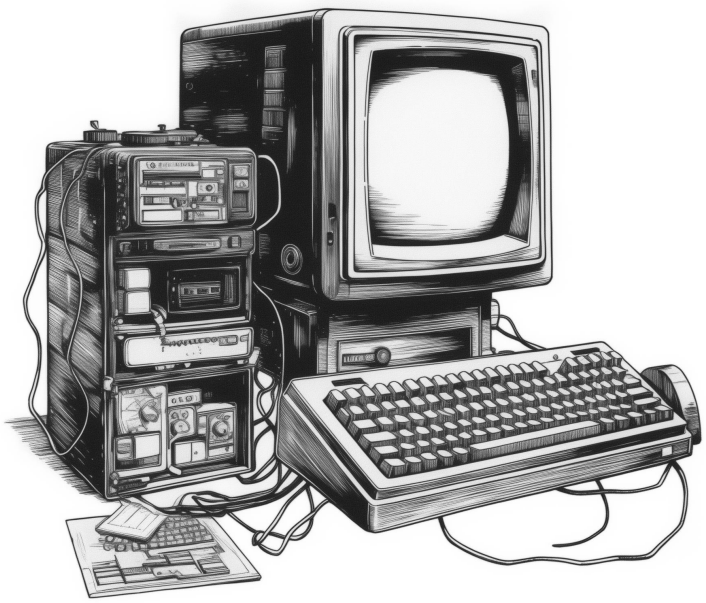
- WHAT: A simple aspect of the environment.
- WHY: None.
- EXAMPLE: A rock.



STAGE 1

At this stage systems act and react. They have hard-wired responses. The instruction set architecture of a modern computer would be stage 1. It is a fixed, inflexible abstraction layer that does not change over the course of the object's existence.

- WHAT: An aspect of the environment.
- WHY: The environment preserves that which preserves itself.
- EXAMPLES: Computers. Proteins.



STAGE 2

THESE SYSTEMS CAN LEARN. This means they have an internal state that changes, and they store information about the past in order to adapt more effectively to the future. However, these systems may be quite primitive, without a unified representation of the self. An example of this is the box jellyfish[559]. The jellyfish is a modular, decentralised system that has localised sensing and control. The system as a whole can learn in the same way that a population of humans might learn. Each member of the population must independently learn a lesson before the population can act upon it as one.

- WHAT: System learns using the weakness proxy.
- WHY: To complete a wider range of tasks than the same system would if it could not learn.
- EXAMPLE: Jellyfish.

[559] Jan Bielecki, Sofie Katrine Dam Nielsen, Gosta Nachman, and Anders Garm. Associative learning in the box jellyfish tripedalia cystophora. *Current Biology*, 2023

STAGE 3

HERE I HAVE the biological equivalent of a Pearlean `do` operator[560]. It allows the organism to discriminate between that which it has caused, and that caused by others. The 1ST-order-self is a from-first-principles mathematical equivalent of reafference, which others have already argued is the key to subjective experience[561]. Houseflies and insects capable of navigating their environment have this[562]. This is where phenomenal consciousness begins. It is here that we have a self, to be subject to these tapestries of valence and feel. This self accompanies every intervention the organism makes, and it has a quality, so it is *what it is like* to be the organism.

- WHAT: A 1ST-order-self, phenomenal consciousness.
- WHY: Functionality requiring causal inference, like navigation.
- EXAMPLE: Housefly.

[560] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., New York, 1st edition, 2018

[561] Bjorn Merker. The liabilities of mobility: A selection pressure for the transition to consciousness in animal evolution. *Consciousness and Cognition*, 2005. Neurobiology of Animal Consciousness

[562] Bjorn Merker. Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences*, 2007; and Andrew B. Barron and Colin Klein. What insects can tell us about the origins of consciousness. *Proceedings of the National Academy of Sciences*, 2016

STAGE 4

WITH A 2ND-ORDER-SELF COMES awareness, the ability to communicate meaning as a human would and above all access consciousness[563]. This is where I predict your prediction of me, and anticipate what I need to say for you to believe what I want you to believe[564]. Dogs seem to be at least this conscious[565]. Why would this come about? Imagine a wolf chasing a rabbit. The wolf can feign left to mislead the rabbit. To know this, the wolf must model what the rabbit thinks the wolf is going to do. Likewise, the rabbit could gain from similar predictions of the wolf. 2ND-order-selves are also needed for co-operation and social competition. The complex hunting behaviour of portia spiders suggests they may be this conscious[566].

- WHAT: 2ND-order-selves, access consciousness.
- WHY: Predation or co-operation.
- EXAMPLES: Wolves[567]. Portia spiders[568].

[563] Ned Block. On a confusion about a function of consciousness. *Brain and Behavioral Sciences*, 1995

[564] Paul Grice. Meaning. *The Philosophical Review*, 66(3):377–388, 1957; and Paul Grice. Utterer's meaning and intention. *The Philosophical Review*, 78(2):147–177, 1969

[565] I say *at least*, because they are likely more conscious.

[566] Fiona R. Cross, Georgina E. Carvell, Robert R. Jackson, and Randolph C. Grace. Arthropod intelligence? the case for portia. *Frontiers in Psychology*, Volume 11 - 2020, 2020

[567] Definitely.

[568] Possibly. It is not as clear from observation as with wolves.

STAGE 5

THE 3RD-ORDER-SELF is where I am aware that I am self aware. It allows for more complex deception and co-operation, and planning communication. Humans are obviously at least this conscious. Other animals might be as well, but this is harder to establish through observation than with 2ND-order-selves. The altruistic behaviour observed in Australian magpies suggests they have at *least* 2ND-order and perhaps 3RD-order-selves[569].

- WHAT: 3RD-order-selves, impelling narrative.
- WHY: Complex manipulation and social predation of the sort that requires predicting your prediction of my prediction of your prediction of me. Iterated prisoner's dilemma.
- EXAMPLE: Humans are at least this conscious. It seems likely many other animals are too.

[569] Joel Crampton, Celine H. Frère, and Dominique A. Potvin. Australian magpies gymnorhina tibicen cooperate to remove tracking devices. *Australian Field Ornithology*, 39:7–11, 2022. DOI: http://dx.doi.org/10.20938/afo39007011. URL https://afo.birdlife.org.au/afo/index.php/afo/issue/view/221

XIII. HOW TO BUILD CONSCIOUS MACHINES

This chapter describes what I need to build a conscious machine[570]. I've shown qualia serve a function. I've shown intelligence is necessary and sufficient for consciousness. The various orders of self and the delegated architecture that support consciousness convey a functional advantage (see figure 9[571]). What this means is that artificial general intelligence and a conscious machine are one and the same goal. Others have proposed designs for both conscious machines and AGI. My formalism is intended to help refine rather than replace those designs. Instead of taking a definition of consciousness and building a machine that satisfies it, I have taken the opposite approach. I have started with what must be true of every environment and then methodically worked my way inwards from one fact to the next, until I found aspects aligned with some or all definitions of consciousness. Based on that, I can see some necessary features a conscious machine must have[572]. Likewise for AGI. I will enumerate both here.

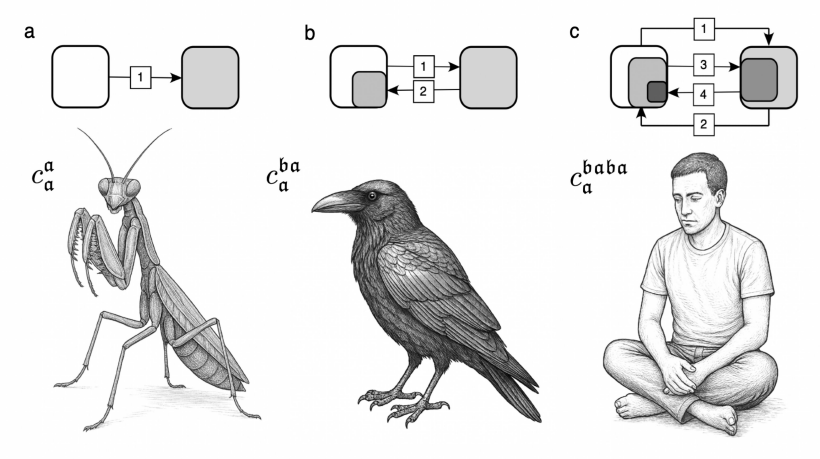


Figure 9: Illustration of a orders of self. The 1ST-order-self is the square on the left above the mantis. It is functionally equivalent to reafference, which even small insects have. The 2ND-order-self c_a^{ba} is where access consciousness begins. It is the square inside the bigger square on the left above the raven. If the raven is a and I am b, then the raven's 2ND-order-self is his prediction of my prediction of him. I have argued for a strict interpretation of access consciousness based on Gricean pragmatics, which amounts to the claim that it depends on theory of mind. Animals such as wolves would seem to be at least this conscious. Finally, I have a 3RD-order-self c_a^{baba} for the complex impelling narrative of humans. It is the small square inside the slightly bigger square inside the even bigger square on the left side above the man. If the man is a, and I am b, then the man's 3RD-order-self depicted is the man's prediction of my prediction of his prediction of my prediction of him. It seems likely other animals that are also this conscious. The Australian magpie has been observed exhibiting behaviour that would support that claim. Drawing by Ricard Solé of The Santa Fe Institute.

[570] Manuel Blum and Lenore Blum. A theoretical computer science perspective on consciousness. *J. Artif. Intell. Conscious.*, 8:1–42, 2020; Pei Wang. A constructive explanation of consciousness. *Journal of Artificial Intelligence and Consciousness*, 07(02):257–275, 2020; and Piotr Boltuc. The engineering thesis in machine consciousness. *Techné: Research in Philosophy and Technology*, 2012

[571] Michael Timothy Bennett and Ricard Solé. Does suspended animation kill consciousness? *Under review*, 2025

[572] Hopefully what I have identified is also sufficient, but that is less certain.

THE ARTIFICIAL SCIENTIST

If conscious machines and AGI amount to the same thing, then it might help to begin with what is necessary for AGI. In one of my early papers[573] I enumerated a few necessary features of an artificial scientist, building on Goertel's earlier survey[574]. To reiterate my earlier discussion of an artificial scientist in chapter 3, this is an agentic system that can perform the entire job of a scientist as well as a human. Most of the features I'm about to discuss are implicit in the complex adaptive systems I've been describing, and in existing proposals for AGI[575] and conscious machines[576]:

(1) representation of hypothesis space.

(2) inductive, deductive and abductive reasoning.

(3) causal reasoning.

(4) can communicate its reasons and results.

(5) can evaluate and prioritise hypotheses humans will find useful.

(6) can design, evaluated and plan experiments.

(7) enactive cognition, actually running experiments and modifying the world to reduce uncertainty.

Since my artificial scientist paper was published I have discovered a number of supporting results. Here I'll briefly rehash why these features are needed, and I'll integrate my more recent results as I do so. In particular, I'll tie in the orders of self and Bennett's Razor. Then I will address the shortcomings of this model, and explain what is additionally necessary for a fully competent artificial scientist, why that requires a different sort of hardware, and why I think this will be conscious. I'll number these requirements as I address them, in accord with the above list.

(1) Representation of a hypothesis space is an obvious prerequisite for scientific pursuits. Science is about falsifiable hypotheses. I cannot falsify a hypothesis if I cannot first represent it. Like everything else, a hypothesis is a causal-identity. To represent a given hypothesis, I need an abstraction layer that meets the scale precondition for that causal-identity[577].

[573] Michael Timothy Bennett and Yoshihiro Maruyama. The artificial scientist: Logicist, emergentist, and universalist approaches to artificial general intelligence. In *Artificial General Intelligence*. Springer Nature, 2022b

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

[575] Ben Goertzel et al. Opencog hyperon: A framework for agi at the human level and beyond. Technical report, OpenCog Foundation, 2023; Patrick Hammer and Tony Lofthouse. 'opennars for applications': Architecture and control. In Ben Goertzel, Aleksandr I. Panov, Alexey Potapov, and Roman Yampolskiy, editors, *Artificial General Intelligence*, pages 193–204, Cham, 2020. Springer Nature; and Kristinn R Thorisson, Eric Nivel, Bas Steunebrink, Helgi P. Helgason, Giovanni Pezzulo, Ricardo Sanz, Jurgen Schmidhuber, Harris Dindo, Manuel Rodriguez, Antonio Chella, Gudberg K Jonsson, Dimitri Ognibene, and Carlos Corbato-Hernandez. Autonomous acquisition of situated natural communication. *Intl. J. Comp. Sci.& Info. Sys.*, 2014

[576] Lenore Blum and Manuel Blum. A theory of consciousness from a theoretical computer science perspective: Insights from the conscious turing machine. *Proceedings of the National Academy of Sciences*, 119(21):e2115934119, 2022. doi: 10.1073/pnas.2115934119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2115934119; and Pei Wang. *A Constructive Explanation of Consciousness and its Implementation*. World Scientific, 2023

[577] Michael Timothy Bennett. Emergent causality and the foundation of consciousness. In *Artificial General Intelligence*. Springer Nature, 2023b; and Michael Timothy Bennett. Computational dualism and objective superintelligence. In *Artificial General Intelligence*. Springer Nature, 2024a

(2) INDUCTIVE INFERENCE is needed if my artificial scientist is to form hypotheses from observing and interacting with the world. This is obvious, but it needs to be said. Early AI systems were conceived of as thinkers rather than learners[578]. Reasoning alone was considered to be the key ingredient for intelligence. In hindsight this is obviously wrong. Adaptation demands the ability to learn. Deductive and abductive reasoning are also needed. Each known known constrains my space of hypotheses. Abduction lets me see reason out what is possible given a fact. Deduction allows me to derive facts from facts. If I am to draw conclusions based upon hypotheses I've tested and established to be true, then I need both.

(3) CAUSAL LEARNING is needed because science aims to map the natural world using cause and effect. Heat causes fire. Gravity causes things to fall. Penicillin can cure bacterial infections. What is particularly interesting here is that to learn cause and effect *in general*, a system must w-max to learn the causal identities for things that cause one another. It cannot presuppose variables. A human scientist might typically start with known variables, but we often need to invent new concepts and these concepts need to accurately map causal relations. I could have an artificial scientist use only our concepts of heat and gravity, but then it wouldn't be equipped to describe causal graphs that don't map to those concepts. My artificial scientist needs to be able to *learn* the world, like we do. So it must w-max. Furthermore science tends to involve interventions[579]. The aim is to identify causal interventions that reliably achieve a certain outcome. For that my artificial scientist must have a 1ST-order-self[580]. Otherwise, how could it run an experiment?

(4) COMMUNICATION IS also critically important. If I cannot communicate my results, then I am of no use as a scientist. To communicate, my artificial scientist needs 2ND-order-selves. It also needs human-like motives, so that it ascribes similar meanings to utterances[581]. This is of course for an *agent* that can independently conduct every aspect of research, rather than a tool like AlphaFold[582].

[578] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach, 4th Edition*. Prentice Hall, Hoboken, 2020

[579] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., New York, 1st edition, 2018

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

[581] Michael Timothy Bennett. On the computation of meaning, language models and incomprehensible horrors. In *Artificial General Intelligence*. Springer Nature, 2023c

[582] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 2021

(5) The ability to evaluate and prioritise hypotheses based on what humans want requires three things. First, hypotheses should be prioritised based on weakness[583]. Second, hypotheses need to be prioritised based on whether falsifying them is likely to reveal information about something. Third, hypotheses must be evaluated and prioritised based on an understanding of human normativity. In other words, that something they reveal information about better be useful. To do this as well as a human scientist requires 3RD-order-selves, because it involves understanding the overall narrative of human endeavour. In order words, my artificial scientist needs 3RD-order-selves to properly understand relevance and utility. Then hypotheses can be evaluated based on whether they are plausible enough tot warrant investigation, and whether their confirmation or falsification would yield information about useful things.

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

(6) Having chosen hypotheses to test, my artificial scientist needs to design experiments that falsify hypotheses. This requires planning, balancing long term resources availability against the cost and benefits of different experiments.

(7) Finally, my artificial scientist needs to actually conduct the experiments and update its beliefs. Together with the previous features, this amounts to enactive cognition, interacting with the world to co-create knowledge.

So to reiterate, my artificial scientist needs:

- 1st-order-self: To design and reason about experiments, it must be able to reason about causal interventions. Hence, a 1ST-order-self is necessary. It must satisfy the incentive and scale preconditions for a 1ST-order-self, to reason about its own role in causality.
- 2nd-order-selves: It must satisfy the scale and incentive preconditions for 2ND-order-selves of the sort humans have, to understand human meaning.
- 3rd-order-selves: Likewise, it must meet the scale and incentive preconditions for 3RD-order-selves in a manner similar to humans, to understand narrative. This is important to understand what is *important* to humans.

Something which doesn't have these selves is definitely not going to be conscious. Nor will it be able to get a job as a scientist. However these features and selves are not yet *sufficient* either. Imagine I build a machine. I rig together some balsa wood cogs, shafts and rubber bands and get a contraption. I implement basic mechanical computation with this system, and I have it learn and store information. I embiggen and optimise it until it meets the scale and incentive preconditions for a minimal 1ST-order-self. Simple, like a fly. I make a speaker and microphone out of rubber bands and teach it a very simple language so that it can ask for food, and it learns a simple 2ND-order-self. I could even make it construct a 3RD-order-self. Is there going to be anything it is *like* to be this contraption, even though it has selves? It seems implausible. Where is the tapestry of valence? I am forcing it to behave in a manner that looks intelligent to me by placing constraints on it, using balsa wood and rubber bands. But is that just a fiction I ascribe to it? Intuitively, there seems to be something missing. What is it?

WHAT BIOLOGY HAS THAT AI DOES NOT

WHAT IT MY CONTRAPTION MISSING? In the previous chapter I sought to explain why evolved biological systems are conscious. I did not address computers. I have argued that selves are 1ST, 2ND and 3RD order selves are necessary for consciousness, but I do not claim are sufficient. There are features of biological systems which my argument in the previous chapter takes for granted. These features are found in biological systems, but are missing from contemporary artificial intelligence systems. These features make biological systems more *intelligent* than contemporary artificial intelligence[584]. I'll argue these additional features are necessary for consciousness, but I am less certain of the necessity of one than I am of the need for the others. Hence I will introduce a problem and speculate about what is implied if one of these features are *not* necessary.

[584] Michael Timothy Bennett. Are biological systems more intelligent than artificial intelligence? 2025b. Forthcoming 2026

SO WHAT ARE THESE FEATURES? In a conscious biological system:

- ADAPTATION IS DELEGATED to very low levels of abstraction, which makes it easier to satisfy the the scale precondition for causal-identities. In comparison, a modern computer adapts only at high levels of abstraction and thus cannot adapt to the same extent.

- CONTROL IS BOTTOM-UP as much as it is top-down. A conscious biological system is a polycomputer, which allows it to compute simultaneously at different scales and levels of abstraction. In comparison, a modern computer might be highly parallelised but it is top-down. The parts do not act independently and form collectives. This makes them less adaptable.

- A CONSCIOUS BIOLOGICAL system has a solid brain, to support synchronised communication between its parts, and above all integrated representation and value judgement. In comparison a computer is sequential. It judges and represents information separately. Again, this makes a computer less adaptable. It is like an inflexible bureaucracy that can only make decisions at the highest levels.

ALL OF THESE TOGETHER should be sufficient to support tapestries of valence. Contemporary AI does not have these features, so it cannot support a tapestry of valence and there is nothing it is like to *be* an AI.

THE TEMPORAL GAP

Some will say we can just implement these features at a high level of abstraction. They will say we can just write software that does all of the above, so there can be something it is like to be a software intelligence[585]. That may or may not be the case. Here I'll argue it is not. I am not arguing for substrate dependence. I am still just treating hardware as an abstraction layer inside an infinite stack of abstraction layers. It is not different from software, and it isn't meaningful to talk about substrate dependence when everything is an abstraction layer. Instead, we can talk about time. Consider a single core CPU as a typical computer. It interprets one instruction at a time. The pointer just moves in memory and that is the extent of its perception. Like the maze solving slime mould, the apparent intelligence of the single core CPU is to be found in the constraints we place on it[586]. If we set the pointer to an address at the start of a program that will look intelligent to us, then the CPU is going to look intelligent to us.

[585] It would still be much less efficient than a human scientist.

[586] Ricard Solé and Luís F Seoane. Evolution of brains and computers: The roads not taken. *Entropy*, 24(5):665, 2022

This is like synchronous versus asynchronous communication. In the single-core computer computations are smeared across time. By changing the state of registers messages are passed to future computations. In contrast biological systems are more like distributed and concurrent computing systems. Multiple computations take place simultaneously at a point in time. Biological systems are "bottom up" in the sense that cells can and do act independently. They interact, and behaviours emerge at higher levels of abstraction from those interactions at low levels of abstraction. This is polycomputing from very low levels of abstraction up, and it integrates representation and value judgement. By that I mean the system is attracted or repelled as it interprets, rather than interpreting value judgements as a separate label attached to a representation after the fact.

MORE TO THE POINT a tapestry of valence[587] is the way the system is being impelled at a particular point in time, not across time. The parts of the system are communicating synchronously in a solid brained polycomputer. It is spread out over space, but constrained to a point in time. In comparison the single core CPU spreads the work out over time, but centralises it in space. We have made CPUs faster by increasing the rate at which they run through the sequence of instructions. That said, the Embiggening[588] of language models over the last decade has relied more heavily on distributing computation, spreading SIMD[589] computations across multiple instances of hardware. However this too tends to be divided up into sequences of instructions and involves a lot of storing information in inert hardware. It is tightly controlled top-down. Like the slime mold, it is about the constraints we place on the system. The same computation can take place in these different abstraction layers, but the way they interact across time is very different.

[587] As I described in the previous chapter.

[588] Michael Timothy Bennett. Lies, damned lies, and the orthogonality thesis. *Preprint*, 2025d

[589] Single instruction multiple data. It means we take an instruction f and a big vector of data $x = [x_1, x_2, ...x_n]$, and do $y = [f_1(x_1), f_2(x_2), ...f_3(x_n)]$ where each $f_i(x_i)$ is separated out and sent to a different bit of hardware, so they can all be computed at the same time.

IT IS LIKE IF I IMPLEMENTED A GAME ENGINE in Python. Yes this is possible, but it is extremely inefficient. In theory we can make Python do the same calculations, but the result would not be the same if we look at the computer running this as one part of a large computational system involving the humans interacting with it. The human would be sitting there waiting for these messages out of sync with its own ability to process information. We want to play games at 60FPS so our perceptions are synchronised with the movements on the screen.

SO YES, WE COULD RUN A TAPESTRY OF VALENCE at a very high level of abstraction on a computer with one single core CPU. Basically just simulate a collective of cells, like an artificial life experiment[590]. However, it would not be *synchronised* and bottom up. A cell's next state would have to be computed at stored, accounting for how it is constrained by the current state of the collective. Then the same would be repeated for the next cell's next state. Think of it like SIMD spread across time rather than space. Then we could compute the result for collectives of cells, up to the system as a whole. At no point in time are all these parts synchronised to impel the system. It is *smeared* across time. That is what I mean by not synchronised. Either this synchronisation matters for consciousness, or it doesn't. Just two options.

[590] Takashi Ikegami. Simulating active perception and mental imagery with embodied chaotic itinerancy. *Journal of Consciousness Studies*, 14(7):111–125, 2007

`OPTION 1` IS THAT IT MATTERS. In this view, computers as we build them today cannot be conscious because the hardware doesn't permit this sort of synchronous bottom-up polycomputation. For a conscious state to be realised, it must be realised in its entirety at a point in time. That means there is a state of the environment that realises every part of a tapestry. It is an aspect realised by the environment. This means my contraption of balsa wood cogs and rubber bands from earlier is not conscious. Nor can a liquid brain be conscious, even though it is controlled bottom-up.

`OPTION 2` IS THAT WE DON'T NEED to have a state realise a tapestry. That it can be smeared across time, and consciousness can thus be smeared across time. If that is so, then we can run the polycomputer at a very high level of abstraction on a single core CPU and the result will be conscious. My contraption of balsa wood and rubber bands can then be conscious, as can a liquid brain.

THESE ARE NOT MUTUALLY EXCLUSIVE OPTIONS. `OPTION 2` can't be true if `OPTION 1` is false. `OPTION 1` is necessary for `OPTION 2`, but `OPTION 2` is not necessary for `OPTION 1`. That said I can't be certain, so I'll call this known unknown *The Temporal Gap*. If I were a betting man[591], I'd go with `OPTION 1`. My contraption of balsa wood and rubber bands is not conscious. It is no more conscious than an angry mob of humans. It seems to me that if time is difference, then each state is like a different reality. Smearing a consciousness state over time will kill it in the same way that smearing the parts of a person over a larger volume of space tends to kill it. To exist as a conscious entity, a system's tapestry of valence must be realised by the state of the environment, and it must be rich enough to support everything else involved[592]. Now I'll explore the implications of these two options.

[591] I am.

[592] Meaning the scale precondition is met for selves of various orders.

OPTION 1: CONSCIOUSNESS IS AT A POINT IN TIME

WHAT SORT OF WORLD DO WE LIVE IN if `OPTION 1` is true and `OPTION 2` is not? There is only *something it is like* to be conscious if one's conscious state is realised by an environmental state. There is not *something it is like* if your conscious state is realised piecemeal by different states.

LIQUID BRAINS cannot be conscious here, because they are necessarily asynchronous, relying on the movement of independent parts for computation. A solid brain is necessary for consciousness.

MORE IMPORTANTLY this makes software consciousness an impossibility. A single core CPU is like a distributed system with only one part, passing messages asynchronously to future versions of itself. This means the state of the environment only ever realises a part of the the system's state. Put another way, from the perspective of the system time is highly compressed and lossy.

MODERN COMPUTERS CAN LEARN, but they do not integrate representation and value judgement the way our highly delegated biological polycomputers do. Modern computers divide up work across time, by sequentially processing information step by step. Information is represented in a format that can be interpreted according to preset rules. This representation is stored, inert, at an address in memory. It only becomes a policy when it is interpreted and actually does something. The representation of information is separate from its interpretation. Information is treated as platonic, because it does not impel the system. There is no tapestry of valence. When we embody representations in the computer we drop Hume's guillotine to disconnect abstracted functionality from the human stack that generated it. I might attach significance to the information, but the computer does not need to. A computer could learn a causal-identity for itself in pursuit of a reward function I impose top-down, but then that causal-identity would not have the bottom-up motivation and quality of a causal-identity learned and interpreted by a biological polycomputer. Due to the sequential and structured nature of computers, there need not be a state of the environment where every part of this causal-identity for self is realised in the CPU. Instead, it is parsed one part at a time. Lacking quality and smeared across time, there would be nothing *it is like* to be the computer. It would fails to live up to the Psychophysical Principle of Causality, and thus be a bit like a zombie[593].

[593] It would of course not *actually* be a philosophical zombie because it would be less sample and energy efficient.

If we simulated a tapestry of valence impelling a system at a higher level of abstraction *running* on this single core CPU, then it would be a set $l \subset P$ that is *not* an aspect of the environment, because it is never realised by a state of the environment. The value judgement of l and its contents are all smeared across time, and in this view that means l is not a conscious state because it is never realised. To build a conscious machine in this view we need:

1. selves: It needs 1ST, 2ND and 3RD-order-selves.
2. delegated: Adaptation must take place at a very low level of abstraction.
3. solid brain: It must have a persistent structure, like telegraph lines, that can support synchronous communication.
4. tapestry of valence: It must be controlled bottom-up, like a distributed system or biological polycomputer, so that it can support a tapestry of valence.
5. synchronised: The tapestry of valence (causal identities for self, the immediate environment etc) is realised by the present state of the environment. Formally, this means the tapestry of valence is a *statement* made in the embodied formal language that is interpreted and judged in one time-step.

There needs to be an *environmental* state where an entire *conscious* state is manifested. To build a conscious machine we would need different *hardware*. I'd need a solid brain[594], because a liquid brain would be smeared across time. Something like self-organising nanites might do the trick[595], especially if they are put together in such a way as to mimic the homeostatic functions of life. Or perhaps concurrent and distributed networks of larger objects will suffice. There is no apparent reason larger objects would not suffice, so long as they are not so large as to prevent synchronous processing. What is important is the tapestry of valence. A hierarchical planning system like the connective core of the human brain[596] may be a suitable architecture for planning interactions between selves and other causal-identities. This would plan narratives, like an internal dialogue. Each aspect of the narrative would be a tapestry of valence and thus have character or feeling: an impelling narrative. This may be sufficient. This suggests a conscious machine, and by extension the artificial scientist I have described, are quite a ways off from our current technology. Of course, I don't *know* that we are conscious at a point in time. It might be that our consciousnesses are smeared across time and we can't perceive it because to us it looks like an instant, which brings us to our next section.

[594] Ricard Solé and Luís F Seoane. Evolution of brains and computers: The roads not taken. *Entropy*, 24(5):665, 2022

[595] Francesca Borghi, Thierry R. Nieus, Davide E. Galli, and Paolo Milani. Brain-like hardware, do we need it? *Frontiers in Neuroscience*, 18, 2024; and B. Paroli, G. Martini, M.A.C. Potenza, M. Siano, M. Mirigliano, and P. Milani. Solving classification tasks by a receptron based on nonlinear optical speckle fields. *Neural Networks*, 166: 634–644, 2023

[596] Murray Shanahan. The brain's connective core and its role in animal cognition. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 367:2704–14, 10 2012. DOI: 10.1098/rstb.2012.0128

OPTION 2: CONSCIOUSNESS SMEARED ACROSS TIME

So what if we are smeared across time? It makes no difference to me. Biological systems *seem* to use a highly delegated polycomputational architecture that supports integrated and *synchronous* representation and value judgement. But this could be running on the cosmic equivalent of a single core CPU at a very low level of abstraction.

Would we be like an LLM being prompted? As a software *intelligence* lives in an abstraction layer we have constructed and control, we are to an LLM what the cosmic horrors of Lovecraftian fiction are to humans[597]. We control the physics of its world, because we control the lower levels of abstraction and exist outside of time as it knows it. Likewise, there could be cosmic horrors prompting us. It is a fun thought. In the absence of synchronisation the potential for science fiction made real is tantalising. We could have conscious software, and store grandma on a USB stick.

[597] Michael Timothy Bennett. On the computation of meaning, language models and incomprehensible horrors. In *Artificial General Intelligence*. Springer Nature, 2023c

CLOSING THE GAP

To summarise, the difference between `OPTION 1` and `OPTION 2` is whether a conscious state must be realised by the state of the environment, or not. From our subjective perspective, there's no way we can tell. I call this unknown *The Temporal Gap*. I haven't thought of a way to definitively falsify `OPTION 2`, but it seems absurd to suggest my balsa wood contraption or a swarm of ants is conscious.

The world would be a more interesting place if `OPTION 2` were true. However I am the only conscious machine I know is definitely conscious, and I appear to have all the features from `OPTION 1`. I can say with some certainty a machine built according to `OPTION 1` will be conscious. However, that doesn't mean `OPTION 2` is false. Any machine that satisfies `OPTION 1` definitely satisfies `OPTION 2` as well. However the reverse is not true. So if we want to build a conscious machine, we should err on the side of `OPTION 1`. If we want to be sure we are *not* building a conscious machine, we should err on the side of `OPTION 2`.

CONCLUSION

Over the course of thesis I have explored what is necessary and sufficient to build a conscious machine. I began by breaking AGI down into tools and meta-approaches, and then analysing the problem I named Computational Dualism. I took a subtractive approach, starting with axioms so weak they are true of all conceivable worlds. From those simple first principles I explored the emergence of order and goal directed behaviour.

I proposed Stack Theory, framing the environment as an infinite stack of abstraction layers. I proposed Pancomputational Enactivism within that, to formalise goal directed behaviour in terms of tasks. This yielded mathematical and experimental results showing weak constraints on function are necessary and sufficient for generalisation and thus adaptation, and simple forms are not. This undermined Ockham's Razor, so I proposed Bennett's Razor in its stead.

I proposed a new meta-approach to AGI I call w-maxing. I showed systems are better able to adapt when they delegate control to lowest level of abstraction possible while still satisfying correctness constraints[598], proving The Law of the Stack. This has implications that apply to all systems, whether it is a human economy or a computer. I then explored how adaptive systems can learn causality, by learning the objects and properties that cause valence.

[598] By which I mean they still complete the task they need to.

Using this I proposed the Psychophysical Principle of Causality, to explain why our subjective experience of an environment is simplified into the objects and properties that it is. A vast orchestra of cells playing a symphony of valence, classifying and judging. I explained the construction of selves and phenomenal consciousness in causal terms. I related this to language and semiotics, formalising Gricean pragmatics and Peircean triadic symbols as tasks. This led to an alternative, stricter definition of access consciousness as the contents of 2ND and higher order selves. This makes a philosophical zombie impossible in all conceivable worlds.

I then explained the emergence of normative meanings in terms of cancer, collective identity, and my The Mirror Symbol Hypothesis[599]. I refuted the strong orthogonality thesis, showing goals and intelligence to be intrinsically linked. Then I sought to explain the origins of life, and prove a version of The Law of Increasing Functional Information.

[599] Michael Timothy Bennett and Yoshihiro Maruyama. Philosophical specification of empathetic ethical artificial intelligence. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2): 292–300, 2022a

FINALLY, I USED ALL THIS to show that consciousness exists because it aids adaptation. I explained qualia as causal identities, and causal identities as tapestries of valence in biological polycomputers with persistent structure. Intelligence[600] is necessary and sufficient for consciousness. I concluded by enumerating the features a conscious machine must have, and I proposed The Temporal Gap in relation to time and substrate independence. If we want to be sure we are building a conscious machine, we should aim for a highly delegated solid brain in which tapestries of valence are realised at a point in time, rather than smeared across it.

[600] As in sample and energy efficient adaptation.

APPENDIX A: TECHNICAL APPENDIX

Here I summarise all the mathematical definitions, proofs, experiments and gives examples. It is quite self contained, but contains very little discussion. It contains examples and pseudocode as well. This appendix, as well as the code I used for the experiments, is available on GitHub: `https://github.com/ViscousLemming/Technical-Appendices`

Bibliography

Gabrielle S. Adams, Benjamin A. Converse, Andrew H. Hales, and Leidy E. Klotz. People systematically overlook subtractive changes. *Nature*, 2021.

Salvatore J. Agosta, Niklas Janz, and Daniel R. Brooks. How specialists can be generalists: resolving the "parasite paradox" and implications for emerging infectious disease. *Zoologia (Curitiba)*, 27(2):151–162, Apr 2010. ISSN 1984-4670. DOI: 10.1590/S1984-46702010000200001. URL `https://doi.org/10.1590/S1984-46702010000200001`.

Brett P. Andersen, Mark Miller, and John Vervaeke. Predictive processing and relevance realization: exploring convergent solutions to the frame problem. *Phenomenology and the Cognitive Sciences*, 2022.

John R. Anderson, Daniel Bothell, Michael D. Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. *Psychological Review*, 2004. Because apparently six authors are needed to figure out how your brain works.

W. R. Ashby. Principles of the self-organizing dynamic system. *Journal of General Psychology*, 1947.

Bernard Baars. *In the Theater of Consciousness: The Workspace of the Mind*. 1997.

Michael F. Barnsley. *Fractals Everywhere*. Academic Press, second edition edition, 2012.

Andrew B. Barron and Colin Klein. What insects can tell us about the origins of consciousness. *Proceedings of the National Academy of Sciences*, 2016.

Jacob D. Bekenstein. Universal upper bound on the entropy-to-energy ratio for bounded systems. *Phys. Rev. D*, 23:287–298, Jan 1981.

Martha Ann Bell and Kirby Deater-Deckard. Biological systems and the development of self-regulation: Integrating behavior, genetics, and psychophysiology. *Journal of developmental and behavioral pediatrics*, 2007.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097.

Michael Timothy Bennett. Symbol emergence and the solutions to any task. In *Artificial General Intelligence*. Springer Nature, 2022a.

Michael Timothy Bennett. Compression, the fermi paradox and artificial super-intelligence. In *Artificial General Intelligence*. Springer Nature, 2022b.

Michael Timothy Bennett. Computable Artificial General Intelligence. *Under Review*, 2022c.

Michael Timothy Bennett. The optimal choice of hypothesis is the weakest, not the shortest. In *Artificial General Intelligence*. Springer Nature, 2023a.

Michael Timothy Bennett. Emergent causality and the foundation of consciousness. In *Artificial General Intelligence*. Springer Nature, 2023b.

Michael Timothy Bennett. On the computation of meaning, language models and incomprehensible horrors. In *Artificial General Intelligence*. Springer Nature, 2023c.

Michael Timothy Bennett. Computational dualism and objective superintelligence. In *Artificial General Intelligence*. Springer Nature, 2024a.

Michael Timothy Bennett. Is complexity an illusion? In *Artificial General Intelligence*. Springer Nature, 2024b.

Michael Timothy Bennett. Technical appendices, 2025a. URL `https://github.com/ViscousLemming/Technical-Appendices`.

Michael Timothy Bennett. Are biological systems more intelligent than artificial intelligence? 2025b. Forthcoming 2026.

Michael Timothy Bennett. What the f*ck is artificial general intelligence? *Artificial General Intelligence*, 2025c.

Michael Timothy Bennett. Lies, damned lies, and the orthogonality thesis. *Preprint*, 2025d.

Michael Timothy Bennett. Optimal policy is weakest policy. *Artificial General Intelligence*, 2025e.

Michael Timothy Bennett. A formal theory of optimal learning with experimental results. *IJCAI*, 2025f.

Michael Timothy Bennett and Yoshihiro Maruyama. Philosophical specification of empathetic ethical artificial intelligence. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):292–300, 2022a.

Michael Timothy Bennett and Yoshihiro Maruyama. The artificial scientist: Logicist, emergentist, and universalist approaches to artificial general intelligence. In *Artificial General Intelligence*. Springer Nature, 2022b.

Michael Timothy Bennett and Ricard Solé. Does suspended animation kill consciousness? *Under review*, 2025.

Michael Timothy Bennett, Sean Welsh, and Anna Ciaunica. *Why Is Anything Conscious?* Preprint, accepted to and presented at ASSC27 and MoC5, 2024.

Jan Bielecki, Sofie Katrine Dam Nielsen, Gosta Nachman, and Anders Garm. Associative learning in the box jellyfish tripedalia cystophora. *Current Biology*, 2023.

Ned Block. On a confusion about a function of consciousness. *Brain and Behavioral Sciences*, 1995.

Lenore Blum and Manuel Blum. A theory of consciousness from a theoretical computer science perspective: Insights from the conscious turing machine. *Proceedings of the National Academy of Sciences*, 119(21):e2115934119, 2022. DOI: 10.1073/pnas.2115934119. URL `https://www.pnas.org/doi/abs/10.1073/pnas.2115934119`.

Manuel Blum and Lenore Blum. A theoretical computer science perspective on consciousness. *J. Artif. Intell. Conscious.*, 8:1–42, 2020.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Occam's razor. *Information Processing Letters*, 1987.

Piotr Boltuc. The engineering thesis in machine consciousness. *Techné: Research in Philosophy and Technology*, 2012.

Joshua Bongard and Michael Levin. There's plenty of room right here: Biological systems as evolved, overloaded, multi-scale machines. *Biomimetics*, 8(1), 2023.

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

Nick Bostrom. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22(2): 71–85, May 2012. ISSN 1572-8641. DOI: 10.1007/s11023-012-9281-3. URL https://doi.org/10.1007/s11023-012-9281-3.

Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford, UK, 2014. ISBN 9780199678112.

Piotr Bołtuć. Consciousness for agi. *Procedia Computer Science*, 2020. BICA 2019.

Richard Brown, Hakwan Lau, and Joseph E. LeDoux. Understanding the higher-order approach to consciousness. *Trends in Cognitive Sciences*, 23(9):754–768, 2019. DOI: 10.1016/j.tics.2019.06.009.

Tom B Brown et al. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, NY, 2020.

Scott Camazine. Patterns in nature. *Natural history*, 2003.

Scott Camazine, Nigel Franks, J Sneyd, Eric Bonabeau, Jean-Louis Deneubourg, and Guy Theraulaz. *Self-Organization in Biological Systems*. Princeton University Press, NJ, 2001.

Murray Campbell, A.Joseph Hoane, and Feng hsiung Hsu. Deep blue. *Artificial Intelligence*, 2002.

Rosa Cao and Daniel Yamins. Explanatory models in neuroscience, part 2: Functional intelligibility and the contravariance principle. *Cognitive Systems Research*, 85:101200, 2024.

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

D. J. Chalmers. *The conscious mind: In search of a fundamental theory*. Oxford University Press, 1996.

David Chalmers. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 1995.

François Chollet. On the measure of intelligence, 2019.

Anna Ciaunica, Evgeniya V. Shmeleva, and Michael Levin. The brain is not mental! coupling neuronal and immune cellular processing in human organisms. *Frontiers in Integrative Neuroscience*, 2023.

Andy Clark. *Being There: Putting Brain, Body, and World Together Again*. MIT Press, 1997.

Joel Crampton, Celine H. Frère, and Dominique A. Potvin. Australian magpies gymnorhina tibicen cooperate to remove tracking devices. *Australian Field Ornithology*, 39:7–11, 2022. DOI: http://dx.doi.org/10.20938/afo39007011. URL https://afo.birdlife.org.au/afo/index.php/afo/issue/view/221.

Fiona R. Cross, Georgina E. Carvell, Robert R. Jackson, and Randolph C. Grace. Arthropod intelligence? the case for portia. *Frontiers in Psychology*, Volume 11 - 2020, 2020.

Charles Darwin. *On the Origin of Species*. 1859.

P C W Davies and C H Lineweaver. Cancer tumors as metazoa 1.0: tapping genes of ancient ancestors. *Physical Biology*, 8(1), feb 2011.

Jordi Delgado and Ricard V. Solé. Collective-induced computation. *Phys. Rev. E*, 55:2338–2344, Mar 1997. DOI: 10.1103/PhysRevE.55.2338. URL https://link.aps.org/doi/10.1103/PhysRevE.55.2338.

Daniel C. Dennett. *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. Simon & Schuster, 1995.

Jacques Derrida. Writing and difference. *U of Chicago P*, 1978.

David Deutsch. *The Fabric of Reality: The Science of Parallel Universes–and Its Implications*. Penguin Books, 1997.

Jacob Devlin et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019.

Hubert L. Dreyfus. *What Computers Can't Do: A Critique of Artificial Reason*. Harper & Row, 1972.

Hubert L. Dreyfus. Why heideggerian ai failed and how fixing it would require making it more heideggerian. *Philosophical Psychology*, 20(2):247–268, 2007. DOI: 10.1080/09515080701239510. URL https://doi.org/10.1080/09515080701239510.

Stefan Edelkamp and Stefan Schrödl. Chapter 9 - distributed search. In Stefan Edelkamp and Stefan Schrödl, editors, *Heuristic Search*, pages 369–427. Morgan Kaufmann, San Francisco, 2012.

Gerald M Edelman and Joseph A Gally. Reentry: a key mechanism for integration of brain function. *Front Integr Neurosci*, 7:63, August 2013.

Nicholas Epley, Adam Waytz, and John T. Cacioppo. On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4):864–886, 2007. DOI: 10.1037/0033-295X.114.4.864. URL https://doi.org/10.1037/0033-295X.114.4.864.

Ben Goertzel et al. Opencog hyperon: A framework for agi at the human level and beyond. Technical report, OpenCog Foundation, 2023.

John Schulman et al. Proximal policy optimization algorithms, 2017.

Ramon Ferrer i Cancho and Ricard Solé. The small world of human language. *Proceedings of the Royal Society B: Biological Sciences*, 268 (1482):2261–2265, 2001. DOI: 10.1098/rspb.2001.1800.

Chris Fields and Michael Levin and. Life, its origin, and its distribution: a perspective from the conway-kochen theorem and the free energy principle. *Communicative & Integrative Biology*, 18(1):2466017, 2025.

Chris Fields and Michael Levin. Scale-free biology: Integrating evolutionary and developmental thinking. *BioEssays*, 42, 06 2020.

Chris Fields, Mahault Albarracin, Karl Friston, Alex Kiefer, Maxwell JD Ramstead, and Adam Safron. How do inner screens enable imaginative experience? applying the free-energy principle directly to the study of conscious experience. *Neuroscience of Consciousness*, 2025.

J. A. Fodor. Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Sciences*, 3(1): 63–73, 1980. DOI: 10.1017/S0140525X00001771.

Jerry A. Fodor. *The Language of Thought*. Harvard University Press, 1975.

Bas C. van Fraassen. *Laws and Symmetry*. Oxford University Press, 1989.

Stan Franklin, Bernard J Baars, Uma Ramamurthy, Gilbert Harman, Antonio Chella, Michael Wheeler, Terrell Ward Bynum, and John Barker. Apa newsletters, 2008.

Sarah A. Fricke and Christina M. Frederick. The looking glass self: The impact of explicit self-awareness on self-esteem. *Inquiries Journal*, 9(10), 2017. URL http://www.inquiriesjournal.com/articles/1711/the-looking-glass-self-the-impact-of-explicit-self-awareness-on-self-esteem. Accessed: 2025-05-03.

M. Friedman and R.D. Friedman. *Capitalism and Freedom*. University of Chicago Press, 1962.

Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

Karl Friston. Life as we know it. *Journal of The Royal Society Interface*, 10(86):20130475, 2013. DOI: 10.1098/rsif.2013.0475. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2013.0475.

Karl Friston, Lancelot Da Costa, Dalton A.R. Sakthivadivel, Conor Heins, Grigorios A. Pavliotis, Maxwell Ramstead, and Thomas Parr. Path integrals, particular kinds, and strange things. *Physics of Life Reviews*, 47:35–62, 2023. ISSN 1571-0645. DOI: https://doi.org/10.1016/j.plrev.2023.08.016. URL https://www.sciencedirect.com/science/article/pii/S1571064523001094.

Thomas Fuchs. *Ecology of the Brain: The phenomenology and biology of the embodied mind*. Oxford University Press, 2017.

Shaun Gallagher and Dan Zahavi. *The Phenomenological Mind*. Routledge, New York, NY, 2021.

Ashitha Ganapathy and Michael Timothy Bennett. Cybernetics and the future of work. In *2021 IEEE 21CW*, 2021. DOI: 10.1109/21CW48944.2021.9532561.

Robin Gandy. Church's thesis and principles for mechanisms. In *The Kleene Symposium*. North-Holland, 1980.

A. Garcez, M. Gori, L. C. Lamb, L. Serafini, M. Spranger, and S. N. Tran. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. 2019.

Marta Garnelo, Kai Arulkumaran, and Murray Shanahan. Towards deep symbolic reinforcement learning, 2016.

James J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979.

Ben Goertzel. *The Hidden Pattern: A Patternist Philosophy of Mind*. BrownWalker Press, USA, 2006.

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

Ben Goertzel. Generative ai vs. agi: The cognitive strengths and weaknesses of modern llms, 2023. URL https://arxiv.org/abs/2309.10371.

Ben Goertzel. Actpc-chem: Discrete active predictive coding for goal-guided algorithmic chemistry as a potential cognitive kernel for hyperon and primus-based agi, 2024.

Paul Grice. Meaning. *The Philosophical Review*, 66(3):377–388, 1957.

Paul Grice. Utterer's meaning and intention. *The Philosophical Review*, 78(2):147–177, 1969.

Hermann Haken. *Advanced Synergetics: Instability Hierarchies of Self-Organizing Systems and Devices*. Springer-Verlag, Berlin, 1983.

Patrick Hammer and Tony Lofthouse. 'opennars for applications': Architecture and control. In Ben Goertzel, Aleksandr I. Panov, Alexey Potapov, and Roman Yampolskiy, editors, *Artificial General Intelligence*, pages 193–204, Cham, 2020. Springer Nature.

Gilbert H. Harman. The inference to the best explanation. *The Philosophical Review*, 74(1):88–95, 1965. ISSN 00318108, 15581470. URL http://www.jstor.org/stable/2183532.

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346, 1990. ISSN 0167-2789. DOI: https://doi.org/10.1016/0167-2789(90)90087-6. URL https://www.sciencedirect.com/science/article/pii/0167278990900876.

Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968. DOI: 10.1109/TSSC.1968.300136.

FA Hayek. The use of knowledge in society. *American Economic Review*, 35(4), 1945.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Fritz Heider and Marianne Simmel. An experimental study of apparent behavior. *The American Journal of Psychology*, 57(2):243–259, apr 1944. URL https://www.jstor.org/stable/1416950.

Geoffrey Hinton. The forward-forward algorithm: Some preliminary investigations, 2022.

C. Horsman, S. Stepney, R. C. Wagner, and V. M. Kendon. When does a physical system compute? *Proceedings of the Royal Society A*, 470 (2169):20140182, 2014.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems, 2021.

David A. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 1952.

David Hume. *A Treatise of Human Nature*. 1739.

Marcus Hutter. *Universal Algorithmic Intelligence: A Mathematical Top→Down Approach*, pages 227–290. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer Nature, Heidelberg, 2010.

Marcus Hutter, David Quarel, and Elliot Catt. *An Introduction to Universal Artificial Intelligence*. Chapman and Hall/CRC, 1st edition, 2024. DOI: 10.1201/9781003460299.

Daniel Hutto and Erik Myin. Radical enactivism: Basic minds without content, 2013.

Takashi Ikegami. Simulating active perception and mental imagery with embodied chaotic itinerancy. *Journal of Consciousness Studies*, 14(7):111–125, 2007.

Takashi Ikegami and Keisuke Suzuki. From a homeostatic to a homeodynamic self. *Biosystems*, 91(2):388–400, 2008.

Tony Ingesson. *The Politics of Combat: The Political and Strategic Impact of Tactical-Level Subcultures, 1939-1995*. Doctoral thesis (monograph), Department of Political Science, Lund University, 2016.

Johannes Jaeger, Anna Riedl, Alex Djedovic, John Vervaeke, and Denis Walsh. Naturalizing relevance realization: Why agency and cognition are fundamentally not computational. *Frontiers in Psychology*, 15, 2024.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie,

Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 2021.

Friston K., FitzGerald T., Rigoli F., Schwartenbeck P., O. Doherty J., and Pezzulo G. Active inference and learning. *Neurosci Biobehav Rev.*, pages 862–879, 2016.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

Stuart A. Kauffman. *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, 1993.

Henry Kautz and Bart Selman. Planning as satisfiability. In *IN ECAI-92*, pages 359–363, New York, 1992. Wiley.

Scott Kelso. *Dynamic Patterns: The Self-Organization of Brain and Behavior*. MIT Press, Boston, 1997.

M. Khajehnejad, F. Habibollahi, A. Paul, A. Razi, and B. J. Kagan. Biological neurons compete with deep reinforcement learning in sample efficiency in a simulated gameworld. arXiv preprint arXiv:2405.16946, 2024.

Jaegwon Kim. *Philosophy of Mind*. Routledge, New York, 3rd ed. edition, 2011.

David Kirk. Nvidia cuda software and gpu parallel computing architecture. In *Proceedings of the 6th International Symposium on Memory Management*, ISMM '07, page 103–104, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595938930. DOI: 10.1145/1296907.1296909. URL `https://doi.org/10.1145/1296907.1296909`.

Zoe Kleinman and Chris Vallance. AI 'godfather' Geoffrey Hinton warns of dangers as he quits Google. *BBC News*, May 2023. URL `https://bbc.com/news/world-us-canada-65452940`. Accessed: 2025-03-13.

A.N. Kolmogorov. On tables of random numbers. *Sankhya: The Indian Journal of Statistics*, A:369–376, 1963.

Stefanie J Krauth, Jean T Coulibaly, Stefanie Knopp, Mahamadou Traoré, Eliézer K N'Goran, and Jürg Utzinger. An in-depth analysis of a piece of shit: distribution of Schistosoma mansoni and hookworm eggs in human stool. *PLoS Neglected Tropical Diseases*, 6(12): e1969, 12 2012. ISSN 1935-2727. DOI: 10.1371/journal.pntd.0001969. URL https://doi.org/10.1371/journal.pntd.0001969.

Sam Kriegman, Douglas Blackiston, Michael Levin, and Josh Bongard. A scalable pipeline for designing reconfigurable organisms. *Proc Natl Acad Sci U S A*, 117(4):1853–1859, January 2020.

Alex Krizhevsky et al. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 2017.

John E. Laird. *The Soar Cognitive Architecture*. MIT Press, MA, 2012.

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

Victor Lamme. Towards a true neural stance on consciousness. *Trends in cognitive sciences*, 2006.

Victor Lamme and Pieter Roelfsema. The distinct modes of vision offered by feedforward and recurrent processing. *Trends in neurosciences*, 2000.

R. Landauer. Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, 5(3):183–191, 1961.

Shane Legg. *Machine Super Intelligence*. PhD thesis, Uni. of Lugano, 2008.

Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and Machines*, pages 391–444, 2007.

Jan Leike and Marcus Hutter. Bad universal priors and notions of optimality. *Proceedings of The 28th Conference on Learning Theory, in Proceedings of Machine Learning Research*, pages 1244–1259, 2015.

L. A. Levin. Universal sequential search problems. *Problems of Information Transmission*, 9(3):265–266, 1973.

Michael Levin. Bioelectrical approaches to cancer as a problem of the scaling of the cellular self. *Progress in Biophysics and Molecular Biology*, 2021. Cancer and Evolution.

Ming Li and Paul M. B. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications (Third Edition)*. Springer Nature, New York, 2008.

Seth Lloyd. Ultimate physical limits to computation. *Nature*, 406 (6799):1047–1054, 2000.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, NY, 2017. Curran.

Ben Lyons and Michael Levin. Cognitive glues are shared models of relative scarcities: The economics of collective intelligence. *Manuscript*, 2024.

Kingson Man and Antonio R. Damasio. Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence*, 1:446 – 452, 2019. URL https://api.semanticscholar.org/CorpusID:208089594.

Gary Marcus. Deep learning: A critical appraisal, 2018.

John Maynard Smith. *Evolution and the Theory of Games*. Cambridge University Press, 1982.

Patrick McMillen and Michael Levin. Collective intelligence: A unifying concept for integrating biology across scales and substrates. *Communications Biology*, 2024.

Daniel W. McShea. A complexity drain on cells in the evolution of multicellularity. *Evolution*, 56(3):441–452, 03 2002. ISSN 0014-3820. DOI: 10.1111/j.0014-3820.2002.tb01357.x. URL https://doi.org/10.1111/j.0014-3820.2002.tb01357.x.

Bjorn Merker. The liabilities of mobility: A selection pressure for the transition to consciousness in animal evolution. *Consciousness and Cognition*, 2005. Neurobiology of Animal Consciousness.

Bjorn Merker. Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences*, 2007.

Aaron Meurer, Christopher Smith, Mateusz Paprocki, Ondřej Čertík, Sergey Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian Granger, Richard Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, and Anthony Scopatz. Sympy: Symbolic computing in python. *PeerJ Computer Science*, 3: e103, 01 2017. DOI: 10.7717/peerj-cs.103.

Qingwei Mi and Tianhan Gao. Adaptive rubber-banding system of dynamic difficulty adjustment in racing games. *ICGA Journal*, 44(1): 18–38, 2022.

Kevin J. Mitchell. *Free Agents: How Evolution Gave Us Free Will*. Princeton University Press, Princeton, NJ, 2023. ISBN 9780691226231.

Volodymyr Mnih et al. Human-level control through deep reinforcement learning. *Nature*, 2015.

Alain Morin. Levels of consciousness and self-awareness: A comparison and integration of various neurocognitive views. *Consciousness and Cognition*, 2006.

John Morrison. Perceptual confidence. *Analytic Philosophy*, 57(1): 15–48, 2016. DOI: 10.1111/phib.12077.

Thomas Nagel. What is it like to be a bat? *Philosophical Review*, 1974.

T. Nakagaki, H. Yamada, and A. Toth. Maze-solving by an amoeboid organism. *Nature*, 407(6803):470, 2000.

A. Newell and H. Simon. The logic theory machine–a complex information processing system. *IRE Transactions on Information Theory*, 2 (3):61–79, 1956.

Eric Nivel et al. Autocatalytic endogenous reflective architecture. Technical report, Reykjavik University, School of Computer Science, 2013.

Georg Northoff. *Unlocking The Brain, Vol. II: Consciousness*, volume 2. Oxford University Press, USA, 2014.

Laurent Orseau. Asymptotic non-learnability of universal agents with neural networks. In Joscha Bach, Ben Goertzel, and Matthew Iklé, editors, *Artificial General Intelligence: 5th International Conference, AGI 2012*, pages 234–243, Berlin, Heidelberg, 2012. Springer Nature.

Laurent Orseau and Mark Ring. Space-time embedded intelligence. In Joscha Bach, Ben Goertzel, and Matthew Iklé, editors, *Artificial General Intelligence*, pages 209–218, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-35506-6.

B. Paroli, G. Martini, M.A.C. Potenza, M. Siano, M. Mirigliano, and P. Milani. Solving classification tasks by a receptron based on nonlinear optical speckle fields. *Neural Networks*, 166:634–644, 2023.

Adam Paszke et al. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., New York, 1st edition, 2018.

Elija Perrier and Michael Timothy Bennett. Position: Stop acting like language model agents are normal agents, 2025. URL https://arxiv.org/abs/2502.10420.

Megan Peters. Towards characterizing the canonical computations generating phenomenal experience, 04 2021.

Gualtiero Piccinini. *Physical Computation: A Mechanistic Account*. Oxford University Press, UK, 2015.

Gualtiero Piccinini and Corey Maley. Computation in Physical Systems. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, Stanford, Sum. 21 edition, 2021.

I. Prigogine and R. Lefever. Theory of dissipative structures. In H. Haken, editor, *Synergetics*, pages 124–135. Vieweg+Teubner Verlag, 1973.

Ilya Prigogine. *From Being to Becoming: Time and Complexity in the Physical Sciences*. W.H. Freeman, 1980.

Hilary Putnam. Psychological predicates. In William H. Capitan and Daniel Davy Merrill, editors, *Art, mind, and religion*, pages 37–48. University of Pittsburgh Press, 1967.

W.V.O. Quine. *Philosophy of Logic: Second Edition*. Harvard University Press, Cambridge MA, 1986. ISBN 9780674665637. http://www.jstor.org/stable/j.ctvk12scx.

Chris R. Reid, David J T Sumpter, and Madeleine Beekman. Optimisation in a natural system: Argentine ants solve the towers of hanoi. *Journal of Experimental Biology*, 214(1):50–58, jan 2011.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322.

Jonathan Richens and Tom Everitt. Robust agents learn causal world models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pOoKI3ouv1.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 1978.

Giovanni Rolla and Nara Figueiredo. Bringing forth a world, literally. *Phenomenology and the Cognitive Sciences*, 2021.

Fernando Rosas, Pedro A.M. Mediano, Martín Ugarte, and Henrik J. Jensen. An information-theoretic approach to self-organisation: Emergence of complex interdependencies in coupled dynamical systems. *Entropy*, 2018.

David M. Rosenthal. *Consciousness and Mind*. Oxford University Press UK, New York, 2005.

S. Russell and P. Norvig. *Artificial intelligence: A modern approach, global edition 4th*. Pearson, London, 2021.

Stuart Russell. *Artificial Intelligence and the Problem of Control*, pages 19–24. Springer Nature, 2022.

Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach, 4th Edition*. Prentice Hall, Hoboken, 2020.

L. J. Savage. *The Foundations of Statistics*. John Wiley & Sons, NY, USA, 1954.

Jürgen Schmidhuber. Discovering neural nets with low kolmogorov complexity and high generalization capability. *Neural Networks*, 10 (5):857–873, 1997.

Christian Schulte and Mats Carlsson. Chapter 14 - finite domain constraint programming systems. In Francesca Rossi, Peter van Beek, and Toby Walsh, editors, *Handbook of Constraint Programming*, Foundations of Artificial Intelligence. Elsevier, 2006.

John Searle. Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3:417–457, 1980.

Thomas D. Seeley. When is self-organization used in biological systems? *The Biological Bulletin*, 2002.

Anil Seth and Tim Bayne. Theories of consciousness. *Nature Reviews Neuroscience*, 2022.

Anil K Seth, Jeffrey L McKinstry, Gerald M Edelman, and Jeffrey L Krichmar. Visual binding through reentrant connectivity and dynamic synchronization in a brain-based device. *Cereb Cortex*, 2004.

Oron Shagrir. Why we view the brain as a computer. *Synthese*.

Murray Shanahan. The brain's connective core and its role in animal cognition. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 367:2704–14, 10 2012. DOI: 10.1098/rstb.2012.0128.

David Silver et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Gabriel Simmons. Comment on is complexity an illusion?, 2024. URL https://arxiv.org/abs/2411.08897.

JJC Smart. Sensations and brain processes. *Philosophical Review*, 68 (April):141–56, 1959. DOI: 10.2307/2182164.

Elliott Sober. *Ockham's Razors: A User's Manual*. Cambridge Uni. Press, 2015. DOI: 10.1017/CBO9781107705937.

Ricard Solé and Luís F Seoane. Evolution of brains and computers: The roads not taken. *Entropy*, 24(5):665, 2022.

Ricard Solé et al. Fundamental constraints to the logic of living systems. *Interface Focus*, 2024.

Mark Solms. *The Hidden Spring*. Profile Books, London, 2021.

R.J. Solomonoff. A formal theory of inductive inference. part i. *Information and Control*, 7(1):1–22, 1964.

Ricard Solé, Melanie Moses, and Stephanie Forrest. Liquid brains, solid brains. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1774):20190040, 2019. DOI: 10.1098/rstb.2019.0040. URL https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2019.0040.

J. Speaks. Theories of Meaning. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, Stanford, Spring 2021 edition, 2021.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

Luc Steels. Evolving grounded communication for robots. *Trends in Cognitive Sciences*, 7(7):308–312, 2003. ISSN 1364-6613. DOI: https://doi.org/10.1016/S1364-6613(03)00129-3. URL https://www.sciencedirect.com/science/article/pii/S1364661303001293.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019. Association for Computational Linguistics.

Richard Sutton. The bitter lesson. *University of Texas at Austin*, 2019.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, MA, 2018.

Keisuke Suzuki and Takashi Ikegami. Spatial-pattern-induced evolution of a self-replicating loop network. *Artificial Life*, 12(4):461–485, 2006.

Jack W. Szostak. Functional information: Molecular messages. *Nature*, 423(6941):689–689, 06 2003. ISSN 1476-4687. DOI: 10.1038/423689a. URL https://doi.org/10.1038/423689a.

Tor Tarantola, Dharshan Kumaran, Peter Dayan, and Benedetto De Martino. Prior preferences beneficially influence social and non-social learning. *Nature Communications*, 8(1):817, 10 2017. ISSN 2041-1723. DOI: 10.1038/s41467-017-00826-8. URL https://doi.org/10.1038/s41467-017-00826-8. The authors declare no competing financial interests.

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

Kristinn R. Thorisson. *A New Constructivist AI: From Manual Methods to Self-Constructive Systems*, pages 145–171. Atlantis Press, Paris, 2012.

Kristinn R Thorisson, Eric Nivel, Bas Steunebrink, Helgi P. Helgason, Giovanni Pezzulo, Ricardo Sanz, Jurgen Schmidhuber, Harris Dindo, Manuel Rodriguez, Antonio Chella, Gudberg K Jonsson, Dimitri Ognibene, and Carlos Corbato-Hernandez. Autonomous acquisition of situated natural communication. *Intl. J. Comp. Sci.& Info. Sys.*, 2014.

Emmanuelle Tognoli and J A Scott Kelso. Enlarging the scope: grasping brain complexity. *Front Syst Neurosci*, 2014.

Giulio Tononi. An information integration theory of consciousness. *BMC Neuroscience*, 5(1):42, 2004.

Giulio Tononi, Melanie Boly, Marcello Massimini, and Christof Koch. Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7):450–461, Jul 2016. ISSN 1471-0048. DOI: 10.1038/nrn.2016.44. URL https://doi.org/10.1038/nrn.2016.44.

Esmeralda G. Urquiza-Haas and Kurt Kotrschal. The mind behind anthropomorphic thinking: attribution of mental states to other species. *Animal Behaviour*, 109:167–176, 2015.

Francisco Varela, Evan Thompson, Eleanor Rosch, and Jon Kabat-Zinn. *The Embodied Mind: Cognitive Science and Human Experience*. 2016.

Ashish Vaswani et al. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, NY, 2017. Curran.

John Vervaeke and Leonardo Ferraro. *Relevance, Meaning and the Cognitive Science of Wisdom*. Springer Netherlands, Dordrecht, 2013a.

John Vervaeke and Leonardo Ferraro. Relevance realization and the neurodynamics and neuroconnectivity of general intelligence. In Inman Harvey, Ann Cavoukian, George Tomko, Don Borrett, Hon Kwan, and Dimitrios Hatzinakos, editors, *SmartData*, NY, 2013b. Springer Nature.

John Vervaeke, Timothy Lillicrap, and Blake Richards. Relevance realization and the emerging framework in cognitive science. *J. Log. Comput.*, 2012.

H. von Foerster. On self-organizing systems and their environments. In *Self-Organizing Systems*. Pergamon Press, 1960.

Erich von Holst and Horst Mittelstaedt. Das reafferenzprinzip. *Naturwissenschaften*, 37(20):464–476, Jan 1950. ISSN 1432-1904. DOI: 10.1007/BF00622503. URL https://doi.org/10.1007/BF00622503.

David Wallace. *The Emergent Multiverse: Quantum Theory according to the Everett Interpretation*. Oxford University Press, 05 2012. ISBN 9780199546961. DOI: 10.1093/acprof:oso/9780199546961.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199546961.001.0001.

P. Wang. *Rigid Flexibility: The Logic of Intelligence*. Applied Logic Series. Springer Nature, 2006.

Pei Wang. On defining artificial intelligence. *Journal of Artificial General Intelligence*, 10(2):1–37, 2019.

Pei Wang. A constructive explanation of consciousness. *Journal of Artificial Intelligence and Consciousness*, 07(02):257–275, 2020.

Pei Wang. *A Constructive Explanation of Consciousness and its Implementation*. World Scientific, 2023.

Michael Wheeler. Martin Heidegger. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, Fall 2020 edition, 2020.

Alfred North Whitehead. *Process and Reality*. 1929.

M. Wilson. Six views of embodied cognition. *Psychonomic Bulletin & Review*, 9(4):625–636, 2002.

S. Wolfram. *A new kind of science*. Wolfram Media, 2002.

D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997. DOI: 10.1109/4235.585893.

Michael L. Wong, Carol E. Cleland, Daniel Arend, Stuart Bartlett, H. James Cleaves, Heather Demarest, Anirudh Prabhu, Jonathan I. Lunine, and Robert M. Hazen. On the roles of function and selection in evolving systems. *Proceedings of the National Academy of Sciences*, 120(43):e2310223120, 2023. DOI: 10.1073/pnas.2310223120. URL `https://www.pnas.org/doi/abs/10.1073/pnas.2310223120`.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3320–3328, Cambridge, MA, USA, 2014. MIT Press.

Eliezer Yudkowsky et al. Orthogonality thesis. `https://www.lesswrong.com/w/orthogonality-thesis`, 2025. Wiki page from LessWrong with multiple contributors. Accessed: 2025-03-18.

Yichao Zhou and Jianyang Zeng. Massively parallel a* search on a gpu. *Proceedings of the AAAI Conference on Artificial Intelligence*, (1), 2015.

J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, 1977. DOI: 10.1109/TIT.1977.1055714.