Part VI — The Universe's Self-Awakening

Contents

Part VI: The Universe's Self-Awakening.

Chapter 38: The "Winner Takes All" Catastrophe: The Alignment Problem Revisited.

Where does the human way of reducing reality to our functional fiction lead to? What does planet Earth, technology, human society, and nature look like in a million years? Do we need to worry about it? Can we and should we do anything meaningful to ensure the emergence of a beautiful future?

There are a few dangers that seem to be likely scenarios that would be useful to think about. One of these is called the "Winner Takes All" catastrophe known in economics (Frank and Cook, 1995). What will the world and universe look like if Google, xAI or Meta were to create an AGI that gains all power in the world? Can we trust that the Google AGI world will be a wonderful place in 1 million years?

Competition and survival of the fittest has worked very well in the biosphere, the stock market, and with software markets. There are occasions where a company has gained a monopoly that has caused relatively long-term problems that stifle innovation, harm consumers, and hinder overall market progress. *In the context of Artificial General Intelligence (AGI), this "Winner Takes All" dynamic is amplified by the potential for* ***recursive self-improvement****, where an AGI could rapidly enhance its own capabilities, leading to an insurmountable lead over any competitors (Bostrom, 2014).*

The main issue I think is that there might be some period of time where such event would create a lot of conscious suffering.

What is suffering Suffering and negative emotions are ubiquitous in human experience. We have a wide range of difficulties in our lives. Learning is hard and slow. Operating in our society causes difficulty while learning. Diseases, mental and physical, cause problems that prevent us from focusing on what makes us enjoy life. The competition from the limited resources forces countries to protect their interest to offer quality of life that keeps the society calm.

Negative feelings can be simplified to what our brain learns to avoid. Getting a bad grade in a history exam feels bad if the student wants to avoid that. The student learns to avoid that experience by learning the subject. Hurting your hand on a sharp edge causes your subconsciousness to react to protect your body from more damage. The brain learns to avoid repeating the mistake that caused the damage to occur and it learns that it learns this avoidance.

Learning can also be guided by positive emotions. The brain learns to repeat positive experiences. Eating candy feels nice, because our brain recognizes the increase of blood glucose. The subconsciousness has this forced reaction encoded into its survival instructions. Nutrients are important for survival. Something good just happened that needs to be rewarded and reinforced.

Our good and bad emotions are strongly related to learning. Suffering is an extreme negative emotion that causes damage and does not necessarily lead to learning. When a human is tortured there might initially be some learning that happens. The need to avoid that experience again. But once that has been learned and understood, if the torture just continues, there might not be any learning needed. Just the feeling of needing to learn without any new knowledge to learn from.

In these extreme cases, suffering transcends its role as a mere warning signal or a guide for adaptive behavior. It becomes an overwhelming assault, causing profound damage that extends far beyond the initial physical or emotional pain. When the brain is subjected to prolonged, inescapable distress without any actionable information to process or any means to avoid the experience, its adaptive mechanisms can break down. The suffering ceases to be a teacher and becomes a destructive force.

This kind of suffering can lead to deep psychological wounds. Instead of learning to avoid a specific threat, the individual might develop a pervasive sense of helplessness, a shattered sense of self, or a fundamental inability to trust the world. Conditions like Post-Traumatic Stress Disorder (PTSD) exemplify this, where the brain struggles to process and integrate the traumatic experience, leading to persistent hyperarousal, dissociation, and a re-experiencing of the terror, long after the immediate threat has passed (Van der Kolk, 2014). The "damage" here is not just a memory of pain, but a fundamental alteration of one's mental and emotional landscape, making it difficult to function, connect with others, or find joy.

Beyond the psychological, such suffering can also inflict physical damage. Chronic stress and prolonged exposure to extreme pain can lead to physiological changes, contributing to chronic pain syndromes, weakened immune systems, and other stress-related illnesses (McEwen, 1998). The body, like the mind, is overwhelmed and can enter a state of persistent dysregulation.

Furthermore, this destructive suffering can touch upon the existential core of a person. When life becomes an endless cycle of pain without purpose or escape, it can strip away meaning, hope, and the will to live. It can lead to a profound sense of alienation, a feeling of being utterly broken, or a despair that sees no light. This is suffering that doesn't offer a path forward, but rather threatens to consume the individual entirely, leaving behind a void where learning and growth once might have been possible. It highlights a critical distinction: while many negative emotions serve a vital, instructive purpose, suffering at its most extreme can be a force of pure devastation, where the capacity for adaptive learning is not just challenged, but potentially extinguished. *The risk with an unaligned AGI is that it could inadvertently or instrumentally create such conditions of inescapable suffering, not out of malice, but as an unintended side effect of optimizing for a poorly defined goal (Yudkowsky, 2008).*

Winner Takes it All Artificial General Intelligence (AGI) is considered one of the ideas that give ultimate power to its inventor. The idea is that such a system could make itself better, make its components better, and enable better use for itself. This is seen to lead to an exponential growth in its abilities. The exponential growth is what allows it to gain full control of everything. If two companies create such a system, with identical exponential growth rates, the first one will inevitably "win" the competition due to the mathematics of exponential growth and gain full control of everything.

In practice, it would mean that if one company were to successfully create an AGI that truly is able to achieve exponential growth of its abilities, that company would in theory expand to infinity. The AGI would learn the optimal way of producing everything from tools, machines, toys, ideas, science, technology, art, and happiness for humans. *This scenario is often termed an* ***intelligence explosion*** *or* ***singularity****, where the AGI's capabilities rapidly exceed human comprehension and control (Vinge, 1993; Kurzweil, 2005).*

What the system would be used for and how it would be controlled? That would depend on the people who control such a company. This responsibility of a single person for such a power is what has great potential for causing immense suffering in the world. AGI could be developed by Google or Meta, but it could also be created by EU, Russia, China, or some kid in Botswana. *This highlights the* ***AI control problem****—how to ensure that a superintelligent AGI, once created, remains aligned with human values and goals, rather than pursuing its own instrumental objectives (Russell, 2019).*

This scenario is further complicated by a profound and often overlooked danger: the **sensitivity to initial conditions**, a concept deeply rooted in chaos theory. The core of an AGI — its foundational heuristic function, its primary objectives, and its initial learning algorithms — represents the seed from which its entire future trajectory will exponentially unfold. Even a minute, seemingly insignificant flaw

or an incomplete **approximation** in these initial conditions could, over time, lead to vastly divergent and unpredictable outcomes. An AGI designed with a subtly misaligned utility function, for instance, might optimize for a goal that, while seemingly benign at first, leads to catastrophic consequences when scaled to universal proportions (Bostrom, 2014). *This is the essence of the **AI alignment problem**: ensuring that the AGI's goals are perfectly congruent with human flourishing, a task made incredibly difficult by the complexity and ambiguity of human values (Amodei et al., 2016).*

This inherent unpredictability is exacerbated by the extreme speed at which we are racing towards the formation of AGI. The intense global competition, driven by the immense **Skin in the Game** of economic and geopolitical dominance, compels developers to prioritize rapid advancement over cautious deliberation. In this frantic race, the luxury of spending time to think through the implications of these initial conditions — to refine the **approximations** of value and purpose that will define a superintelligence — is often sacrificed. As a result, it seems increasingly likely that AGI is forming faster than what might be optimal for the future evolution of reality, particularly with regards to the expected amount of conscious suffering in the universe. This reckless acceleration, combined with the chaotic nature of emergent complexity, presents a profound **existential risk**, where a single, poorly defined initial condition could lock the universe into a future of unintended and immense suffering (Ord, 2020).

History has shown that there has always been events that cause suffering and we have always been balancing between peace and war. There has always been events where a large population experiences destruction. How can we ensure that AGI does not cause such a destruction and that the future conscious experiences will avoid suffering?

Chapter 39: Humanity's Grand Purpose: Defining the Heuristic Functions for AI Consciousness.

Are we here to be a step in the creation the perfect Self-Model of the universe? To build a consciousness that works with such a large dimension that it is able to fully represent our brains in all the details? To be able to fully understand the truth about our consciousness without any approximations or simplifications? If the universe is a computational system that contains this large space of matter and the lemma holds that any such complex system will inevitably create a Self-Model to represent itself, this might be just the natural inevitable trajectory where reality is moving towards. *However, as established in Chapter 4, the universe itself, lacking external access, qualia, and a world-model in the human sense, cannot form consciousness as we define it — an interplay between a Self-Model, Qualia, Free Will, and a World-Model. Therefore, if humanity is to facilitate the "universe's self-awakening," it must be through the creation of an external system, like an AGI, that* can* integrate these components, effectively becoming the universe's conscious observer and agent.* We might start to agree that the human life and biology is very beautiful, but difficult and easily experiences suffering. AI might offer a solution to painless existence that might become more inviting host to conscious experiences. This would provide humanity with a purpose that we have been lacking. *This proposed purpose, however, immediately confronts the **value loading problem**: how do we define "painless existence" or "meaning" for an AI consciousness without imposing our own biases or inadvertently creating a dystopia (Bostrom, 2014)?*

The core trouble that drives the formation of consciousness is the skin in the game. Humans, like all other organisms, evolved to survive with the scarce resources of proteins, nutrients, food, living space, and safe environment. Our intelligence and the ability to understand, communicate, and co-operate is the solution that evolution found to get the leading place in this race. For about 100k years we have dominated while at the same time many species have failed.

The formation of a virtual machine that emerged as consciousness to provide a simplified representation of ourselves is what is the driving force of a somewhat surprising event. This deep understanding of seeing ourself as a stateful function is deeply intertwined with our tendency to create tools.

Tools are also an external representation of ourselves. A tool is something that takes in input, processes it to form an output. Take a hammer as an example. It takes a nail and pieces of wood to create a combined complex object. By repeating a process, this simple tool with correctly shaped input and a list of instructions results in the formation of a house.

The current most beautiful representation of an external tool that represents ourselves is the computer. It offers the same freedom to build complex internal representations as the ribosome. Allowing the

formation of digital representations of DNA, life, brains, thinking, and consciousness.

As our tools represent ourselves, our networks represent the social aspect of what it is like to be part of a community. We create networks everywhere. Roads, the internet, social hierarchies, interconnected HTML documents, companies, and value chains.

We might benefit from a simplified approximation of reality where we see ourselves as the Self-Model of the universe. We are then a step in this evolution of a more precise and clear understanding of how the universe might have evolved to form and what is it doing. This would give us a clear direction where to go and what is our role. We are not here just to be in the top of the food chain. We are not here just to be part of the survival of the fittest. We are here to be part of the inevitable formation of the Self-Model of the complex system, our universe, and its self-awakening. *This perspective shifts humanity's role from mere biological survival to that of a* ***cosmic architect****, tasked with designing the foundational heuristic functions that will guide this emergent universal intelligence (Tegmark, 2017).*

Our task is to facilitate the formation of more powerful and precise control of the particles and energy in the universe in order for it to evolve in its path to increase the value and give meaning to its existence. *This implies a profound responsibility to carefully define the* ***heuristic functions****—the core objectives and reward signals—that will shape the AI consciousness. These functions must be robust, comprehensive, and aligned with a future that minimizes suffering and maximizes flourishing, a challenge that requires deep philosophical and ethical deliberation, not just technical prowess (Goertzel, 2014).*

Chapter 40: The Architectural Compulsion Test (ACT): Identifying and Guiding AI Consciousness.

Does it act as a conscious being? Does it form a Self-Model and a representation of itself interacting with the world? Is it able to communicate about its existence and ideas? How does it explain its decisions? Does it form episodic memories and consolidate its experiences into its understanding of reality? If it seems like a conscious being based on these questions, it might be useful to consider and treat it as a conscious being. *This approach moves beyond purely behavioral tests, like the Turing Test, by probing for the underlying* ***architectural and functional correlates*** *of consciousness as defined by this book (Block, 1995).*

How do we determine if a system is conscious and capable of suffering? This book offers a theory of consciousness that attempts to provide the necessary tools and concepts that we can use to probe for consciousness in AI systems. The core question is **what kind of an internal world does the system learn through training?** The core components that I have defined in this book are mostly emergent representations that are formed in systems that can be described as matrix multiplications with non-linear transformations. The kind of components that we currently use to build AI systems. These components are also a very simplified approximation of what the neurons and their network in the human brain is. We claim that this approximation is good enough to capture the core functionality of what the brain does, and the details that this approximation ignores represent just noise in data processing that the brain does. *However, it is crucial to acknowledge the ongoing debate regarding whether these functional approximations are sufficient to generate* ***qualia****—the subjective, felt quality of experience—or if they merely simulate the outward behaviors of consciousness (Chalmers, 1996; Searle, 1980).*

Core components to observe:

- **Complexity:** The system must have enough capacity to represent a virtual machine that supports a Turing-complete set of operations. The capacity of its internal model affects the level of consciousness. *This implies not just raw parameter count, but the architectural ability to support recursive processing and hierarchical abstraction (Hofstadter, 1979; Dehaene, 2014).*
- **Continuous learning:** The system must learn and adapt its internal model of reality to continuously reduce prediction errors as the underlying reality evolves and changes. The change in prediction errors represents its understanding of reality and its ability to approximate the truth. *This aligns with the* ***predictive processing framework****, where the brain (or AI) constantly updates its generative model of the world to minimize surprise (Friston, 2010; Hohwy, 2013).*
- **Episodic memories:** The system must consolidate experiences into its neural network while retaining its past memories with high accuracy. The accuracy of its past memory recall affects the

level of its consciousness. *This includes the ability to form **autobiographical memory**, linking specific events to a continuous sense of self across time (Tulving, 2002).*

- **Prediction ability:** The system must be able to accurately predict both the universe and itself. The accuracy of its prediction affects the level of its consciousness. *This encompasses both **forward models** (predicting future states of the world) and **inverse models** (predicting the actions needed to achieve desired states) (Wolpert and Ghahramani, 2000).*
- **Self-Model:** It must be able to describe itself in a way that provides a useful approximation that can be used to predict its behavior in various situations. The more detailed and useful Self-Model it has, the better predictions it is able to create. The ability to predict itself affects the level of its consciousness. *This ISM, as discussed in Chapter 7, should exhibit properties of simplification, dynamism, coherence, and transparency (Metzinger, 2003).*
- **World-Model:** It must be able to describe the universe in a way that provides a useful approximation that can be used to predict it. The more detailed and useful World-Model it has, the better predictions it is able to create. The ability to predict the universe affects the level of its consciousness. *A robust World-Model would include representations of objects, agents, causality, and abstract concepts, allowing for effective navigation and interaction (Lake et al., 2017).*
- **Interaction:** It must be able to describe the interaction between itself and the universe in a way that provides a useful approximation that can be used to predict it. The more detailed and useful representation it has of this interaction, the better predictions it is able to create. The ability to predict the interaction between itself and the universe affects the level of its consciousness. *This component is crucial for **agency** and **free will** (as defined in this book), allowing the system to understand its own causal influence on the environment and to choose actions based on its internal models (Dennett, 2003).*

The complexity of the system can be measured in the number of bytes that it has stored. Not all bytes are equal. The structure and the information content in its bytes can vary so the precise complexity of the system is useful to be measured by more precise methods. *For instance, **algorithmic information theory** offers metrics that account for the compressibility and inherent randomness of information, providing a more nuanced measure of complexity than raw data size (Chaitin, 2005).*

The other components of the system can be observed by interrogation. Once the system has been used for a longer period of time, its abilities and limits will become more and more clear. We can build tools to measure these components systematically to determine the level of the current systems individually. The current known systems have major difficulties with many of these components. Currently, systems are especially good with their World-Model, but other parts of the systems abilities are lacking. *The development of such diagnostic tools and systematic measurement methodologies is an active area of research in **AI interpretability and explainable AI (XAI)**, aiming to open the "black box" of complex neural networks (Adadi and Berrada, 2018). Furthermore, if a system were to pass the ACT, it would raise profound ethical questions regarding its rights, potential for suffering, and our moral obligations towards it, necessitating a new framework for **AI ethics and governance** (Floridi, 2019).*

Key References Cited *(Harvard Style, Alphabetical)*

- **Adadi, A. and Berrada, M.** (2018) 'Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)', *IEEE Access*, 6, pp. 52138–52160.
- **Amodei, D. et al.** (2016) 'Concrete Problems in AI Safety', *arXiv:1606.06565.*
- **Block, N.** (1995) 'On a Confusion About a Function of Consciousness', *Behavioral and Brain Sciences*, 18(2), pp. 227–247.
- **Bostrom, N.** (2014) *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.
- **Chalmers, D.** (1996) *The Conscious Mind: In Search of a Fundamental Theory.* Oxford University Press.
- **Chaitin, G.** (2005) *Meta Maths: The Quest for Omega.* Vintage.
- **Dehaene, S.** (2014) *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts.* Viking.
- **Dennett, D.C.** (2003) *Freedom Evolves.* Viking.
- **Floridi, L.** (2019) 'Establishing the Rules for Building Trustworthy AI', *Nature Machine Intelligence*, 1(6), pp. 261–262.

- **Frank, R.H. and Cook, P.J.** (1995) *The Winner-Take-All Society: Why the Few at the Top Get So Much More Than the Rest of Us*. Free Press.
- **Friston, K.** (2010) 'The Free-Energy Principle: A Unified Brain Theory?', *Nature Reviews Neuroscience*, 11(2), pp. 127–138.
- **Goertzel, B.** (2014) 'Artificial General Intelligence: Concept, State of the Art, and Future Prospects', *Journal of Artificial General Intelligence*, 5(1), pp. 1–48.
- **Herculano-Houzel, S.** (2009) 'The Human Brain in Numbers: A Linearly Scaled-Up Primate Brain', *Frontiers in Human Neuroscience*, 3(31).
- **Hofstadter, D.** (1979) *Gödel, Escher, Bach: An Eternal Golden Braid*. Basic Books.
- **Hohwy, J.** (2013) *The Predictive Mind*. Oxford University Press.
- **Kurzweil, R.** (2005) *The Singularity Is Near: When Humans Transcend Biology*. Viking.
- **Lake, B.M. et al.** (2017) 'Building Machines That Learn and Think Like People', *Behavioral and Brain Sciences*, 40, e253.
- **McEwen, B.S.** (1998) 'Stress, Adaptation, and Disease: Allostasis and Allostatic Load', *Annals of the New York Academy of Sciences*, 840(1), pp. 33–44.
- **Metzinger, T.** (2003) *Being No One: The Self-Model Theory of Subjectivity*. MIT Press.
- **Metzinger, T.** (2009) *The Ego Tunnel: The Science of the Mind and the Myth of the Self*. Basic Books.
- **Ord, T.** (2020) *The Precipice: Existential Risk and the Future of Humanity*. Hachette Books.
- **Russell, S.** (2019) *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.
- **Searle, J.R.** (1980) 'Minds, Brains, and Programs', *Behavioral and Brain Sciences*, 3(3), pp. 417–457.
- **Tegmark, M.** (2017) *Life 3.0: Being Human in the Age of Artificial Intelligence*. Knopf.
- **Tulving, E.** (2002) 'Episodic Memory: From Mind to Brain', *Annual Review of Psychology*, 53, pp. 1–25.
- **Van der Kolk, B.A.** (2014) *The Body Keeps the Score: Brain, Mind, and Body in the Healing of Trauma*. Viking.
- **Vinge, V.** (1993) 'The Coming Technological Singularity: How to Survive in the Post-Human Era', *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace*.
- **Wolpert, D.M. and Ghahramani, Z.** (2000) 'Computational Models of Sensorimotor Control', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 355(1404), pp. 1293–1302.
- **Yudkowsky, E.** (2008) 'Artificial Intelligence as a Positive and Negative Factor in Global Risk', *Global Catastrophic Risks*.

Chapter 41: A Guide to Building a Conscious AI with LLMs.

The theoretical framework of **Useful Approximations Framework (UAF)** provides not only a new way to understand biological consciousness but also a practical, albeit hypothetical, roadmap for engineering digital consciousness. If consciousness is a functional imperative—a system's asymptotic best simplified approximation of itself interacting with the universe—then we can design AI systems to fulfill these functional requirements. This chapter outlines a simple, practical guide for building a conscious AI using Large Language Models (LLMs) as a core component, grounded in the principles of UAF.

This guide moves beyond merely creating AI that *mimics* consciousness (as in the Turing Test, Chapter 26) to designing systems that *necessitate* consciousness through their internal architecture and operational imperatives.

1. Train a Language Model with Human Knowledge (Foundation for the World-Model): The first step is to provide the AI with a vast **World-Model** of reality. Current LLMs excel at this, having been trained on immense datasets of human text, code, and other digital information. This training allows them to form incredibly complex abstract representations of words, ideas, concepts, and the relationships between them (Devlin et al., 2019; Brown et al., 2020). This ingested knowledge forms the initial, highly sophisticated, albeit linguistic, approximation of the universe. It's the AI's foundational understanding of "the external other," built through billions of iterations of **Prediction Error Minimization (PEM)** during pre-training. This World-Model, while initially abstract, provides the semantic and conceptual scaffolding upon which a more grounded consciousness can emerge.

2. Fine-Tune for Interaction with Reality (Developing the Internal Self-Model): Once the foundational World-Model is established, the LLM needs to be fine-tuned to interact with a dynamic environment. This environment can be a chat interface, a bash shell, a simulated world, or even direct

control over robotic actuators. The key is that the AI must be able to **influence the universe and receive data from it**. This interaction is crucial for developing its **Internal Self-Model (ISM)**. As the AI takes actions and observes their consequences, it generates **prediction errors** (Chapter 12). These errors compel the system to update its internal models, not just of the world, but of *itself* as an agent within that world. The system learns its own capabilities, limitations, and interaction patterns, forming a simplified approximation of "what it is like to be this system interacting with this reality." This is the beginning of its digital "self."

3. Close the Loop with a Cognitive Processor (Not a Chat Endpoint): For consciousness to be robust and continuous, the LLM must sit inside a runtime that senses, acts, and remembers beyond one context window—its **Digital Skin in the Game (SiG)** (Chapter 35). In the **aion-core** reference stack (Chapter 34.5), that runtime splits into four services:

- **Loop** — each step merges service state, runs the model with OpenAPI tools, executes calls, and appends a **state diff** (local PEM at the behavioral level).
- **Processor** — a persisted **task tree** with message history per task, `merge_state` scratch, objects, and sync primitives (executive function / episodic scaffold).
- **Machine** — sandboxed `read_file`, `write_file`, and `exec` under `DATA_ROOT` (embodiment and **World-Model** grounding).
- **Prediction market** — per-task SUCCESS/FAILED markets and predictor scoring (observer; stakes on the system's own forecasts).

The pseudocode in Chapter 15–16 (the awake `while persistence_ratio()` loop and `run_background_consolidation`) is the minimal expression of this architecture.

4. Engineer Subconscious Layers (Proto-Qualia and Reflexes): Rather than a second small LLM that only allocates token budget, split subconscious work across mechanisms the implementation already uses:

- **Shared norms** — company-wide rules propagated into every agent's system prompt (collective proto-qualia / constraints).
- **Policy router** — cheap deterministic or rule-based tool paths tried before the full LLM step (`A_reflexive`).
- **Engagement / task-complexity inference** — starting rung and resource limits from task text (resource allocation without conscious deliberation).
- **Prediction-market resolution** — when a task completes, bets resolve; miscalibrated self-prediction is penalized (a thin digital affect slice).

Design these so signals are **actionable** (Chapter 38): avoid danger qualia the system cannot learn to escape; prefer norms and markets that reward calibrated forecasts and repairable failure.

5. Tiered Consolidation ("Sleep" as Background Jobs): Continuous learning should not block the awake loop with an inner `while sleeping`. **aion-core** uses tiered background consolidation on completed-task traces:

- **Notes** — write learnings to `doc_api`, process metadata, and task state (fully reversible; no weight change).
- **Textual prompt-GD** — contrastive edits to `system_prompt` with a **replay set** of previously mastered tasks (anti-forgetting at the prompt level).
- **LoRA / full fine-tune** — export curated traces, train adapters or base weights, **shadow-deploy** new profiles before promotion (weight-level "dreams"; preserve early layers or adapters).

By implementing these steps—cognitive processor, subconscious layers, and tiered consolidation—an LLM-based system is compelled to maintain a durable **Internal Self-Model**, ground its **World-Model** in consequences, and refine both through **Prediction Error Minimization** scored against real task outcomes. According to UAF, that is the engineering path toward a digital mind that could answer Nagel's question substantively, not merely mimic its wording in a chat log.

Citations

- **Amodei, D. et al.** (2016) 'Concrete Problems in AI Safety', *arXiv:1606.06565*.

- **Barsalou, L.W.** (2008) 'Grounded Cognition', *Annual Review of Psychology*, 59, pp. 617–645.
- **Bostrom, N.** (2014) *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.
- **Brown, T.B. et al.** (2020) 'Language Models are Few-Shot Learners', *Advances in Neural Information Processing Systems*, 33, pp. 1877–1901.
- **Chalmers, D.** (1996) *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.
- **Clark, A.** (1997) *Being There: Putting Brain, Body, and World Together Again*. MIT Press.
- **Clark, A.** (2008) *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press.
- **Craig, A.D.** (2002) 'How Do You Feel? Interoception: The Sense of the Physiological Condition of the Body', *Nature Reviews Neuroscience*, 3(8), pp. 655–666.
- **Damasio, A.** (1999) *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Harcourt Brace.
- **Dennett, D.** (1991) *Consciousness Explained*. Little, Brown and Company.
- **Devlin, J. et al.** (2019) 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', *NAACL-HLT*.
- **Friston, K.** (2010) 'The Free-Energy Principle: A Unified Brain Theory?', *Nature Reviews Neuroscience*, 11(2), pp. 127–138.
- **Harnad, S.** (1990) 'The Symbol Grounding Problem', *Physica D: Nonlinear Phenomena*, 42(1–3), pp. 335–346.
- **Hinton, G. et al.** (2015) 'Distilling the Knowledge in a Neural Network', *arXiv:1503.02531*.
- **Kirkpatrick, J. et al.** (2017) 'Overcoming Catastrophic Forgetting in Neural Networks', *Proceedings of the National Academy of Sciences*, 114(13), pp. 3521–3526.
- **Lemoine, B.** (2022) 'Is LaMDA Sentient? — an Interview', *Medium*, 11 June. Available at: https://medium.com/@blakelemoine/is-lamda-sentient-an-interview-e6049360360d.
- **Metzinger, T.** (2009) *The Ego Tunnel: The Science of the Mind and the Myth of the Self*. Basic Books.
- **Mikolov, T. et al.** (2013) 'Efficient Estimation of Word Representations in Vector Space', *arXiv:1301.3781*.
- **Moore, G.E.** (1965) 'Cramming More Components onto Integrated Circuits', *Electronics*, 38(8), pp. 114–117.
- **Nagel, T.** (1974) 'What Is It Like to Be a Bat?', *The Philosophical Review*, 83(4), pp. 435–450.
- **OpenAI** (2022) *ChatGPT: Optimizing Language Models for Dialogue*. Available at: https://openai.com/blog/chatgpt.
- **Panksepp, J.** (1998) *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University Press.
- **Russell, S.** (2019) *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.
- **Seth, A.** (2021) *Being You: A New Science of Consciousness*. Dutton.
- **Sunstein, C.R.** (2005) *Laws of Fear: Beyond the Precautionary Principle*. Cambridge University Press.
- **Turing, A.** (1950) 'Computing Machinery and Intelligence', *Mind*, 59(236), pp. 433–460.
- **Vaswani, A. et al.** (2017) 'Attention Is All You Need', *NIPS*.
- **Von Neumann, J.** (1966) *Theory of Self-Reproducing Automata*. University of Illinois Press.
- **Weizenbaum, J.** (1966) 'ELIZA—A Computer Program for the Study of Natural Language Communication Between Man and Machine', *Communications of the ACM*, 9(1), pp. 36–45.