Special Issue, "What AI Can Learn from Biology"

Vol. 8, No. 1–2 (2025)
ISSN: 2532-5876

DOI: 10.13133/2532-5876/18905

Modelling the Threat from AI: Putting Agency on the Agenda

Ali Hossaini[a*]

[a] *Department of Engineering, King's College London, United Kingdom*

***Corresponding author:** Ali Hossaini, Email: ali.hossaini@kcl.ac.uk

Abstract

The AI existential-risk narrative focuses on an 'intelligence explosion' leading to uncontrollable superintelligence. This paper contends that the more plausible and proximate threat is the emergence of strong biological-style agency in digital systems, independent of high intelligence. Drawing on systems biology and thermodynamics, it contrasts mechanistic with organic agency: living organisms are autocatalytic systems that harness environmental energy for self-maintenance and reproduction, whereas current Autonomous/Intelligent Systems pursue only externally assigned goals. Evolution produced robust agency in bacteria, slime molds, and insects long before cognition. Recent work in embodied neural networks and bio-inspired computing shows that complex adaptive behavior can arise in machines through structural coupling with their environment that occurs without symbolic reasoning. Deliberate or accidental development of energy-seeking, self-reproducing 'biodigital agents' could therefore yield invasive, unpredictable systems well below superintelligent levels. The paper advocates shifting AI safety priorities from anthropomorphic ethics and alignment to measurable biophysical criteria derived from the definition of life. Recommended measures include engineering standards prohibiting direct environmental energy harvesting by A/IS, global energy audits to detect emergent agency, and epidemiological containment frameworks—thereby preventing a Cambrian-like explosion of machine agency before superintelligence becomes feasible.

Keywords: AI, superintelligence, intelligence explosion, biodigital agents

Citation: Hossaini, A 2025, "Modelling the Threat from AI: Putting Agency on the Agenda'", *Organisms: Journal of Biological Sciences*, vol. 8, no. 1–2, pp. 21–26. DOI: 10.13133/2532-5876/18905

Could intelligent machines challenge humanity's place on Earth? A hearty staple of science fiction has become a legitimate question. Many experts reject the possibility, but others such as Nick Bostrom, Ray Kurzweil and Max Tegmark argue that an upcoming 'singularity' may produce superintelligent AI (Bostrom 2014; Tegmark 2017; Kurzweil 1999; Kurzweil 2005). What happens next is debatable.

The concept of a singularity, or 'intelligence explosion', was introduced by Bletchley Park veteran I. J. Good in the early 1960s:

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of man would be left far behind... Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. It is curious that this point is made so seldom outside of science fiction. It is sometimes worthwhile to take science fiction seriously. (Good 1962)

After half a century of quickening progress in AI, should humanity prepare for a singularity? And, more importantly, should AI be considered an intrinsic threat?

Singularity theorists assume machines will shrug off human oversight if they achieve general intelligence. Yet their descriptions of how AI transforms from mechanical tool to free agent have no basis in observation. Computer scientists define general intelligence as 'a universal algorithm for learning and acting in any environment', but, whatever its degree, intelligence does not in itself motivate behavior (Russell & Norvig 2009, p. 27). The independence described by singularity theorists is properly known as agency, and free agency, as opposed to legal, social or digital agency, has only been observed in living things. Examining the principles of biology, particularly the traits that distinguish organisms from mechanisms, may cast light on how machines could one day acquire agency and the unpredictability that accompanies it (unless otherwise noted, agency henceforth means the capacity to make independent, self-interested decisions).

Rather than from an intelligence explosion and its consequences, the potential threat may come instead from AI's ability to acquire agency. In discussing AI and its potential implications, therefore, it may also be more helpful to adopt the Institute of Electrical and Electronics Engineers' (IEEE) adoption of A/IS (Autonomous and Intelligent Systems) as a term that describes the future scope of information- based technology more accurately than AI (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2019, Introduction).

1. Mechanism vs Organism

Consider the virus. Like bacteria, it infects organisms, but it only reproduces in living cells. In contrast, bacteria possess numerous strategies for survival. Some bacteria infect living bodies while others thrive on the dead. Still others live symbiotically with other species, and a few exploit the physical environment directly. Though both contain either DNA or RNA, an information-carrying molecule similar to DNA, only bacteria are considered alive.

What differentiates bacteria from viruses is their capacity to process energy. When outside cells, viruses are inert, while bacteria dynamically influence their environment to reproduce. This contrast illustrates an essential feature of biology: the cell is the basic unit of life, and the behavior of organisms derives from cell metabolism. It also clarifies the central problem of singularity theory, which is the transformation of machines into agents. What is the digital equivalent of a cell? Most educated people would seek the answer in DNA.

The theoretical model that privileges genes over other biological structures is crumbling (Noble 2006; Noble 2016; Carey 2012; Carey 2015). However, we are still accustomed to reducing life to DNA (Dawkins 1976). A common metaphor is that DNA is software that operates the body's "hardware". Given DNA's informational content, the comparison to computers is easy to make, as is the conclusion that DNA programs the metabolic activities of life. Similar assumptions frame discussions of cognition. The brain holds the software – rational thought – that generates behavior. But analogies to computing fail on a key point: how does information maintain the physical integrity of living systems?

The laws of thermodynamics describe the natural tendency of systems to run down. Every physical system, including machines and isolated DNA, loses coherence over time. Life is a glaring exception to thermodynamic decay. For billions of years life has maintained complex structures – cells and the biosphere – and, given the right inputs of energy, it is effectively immortal. There is nothing supernatural about the processes of life, but they cannot be described in terms of information alone. (Biology is surprisingly quiet about how life originated. See Lane 2015). Harnessing energy, and trading it within an ecosystem, requires physical structures that couple the internal organization of cells to their environment.

2. Information and Organization

Systems biology – an offshoot of systems theory, a field substantially founded by Ludwig von Bertalanffy in the mid-20th century – incorporates a specific notion of agency into its definition of the organism. It is useful to contrast biological agency with the technical conceptions used by software engineers. We can do this by reviewing their respective definitions of work. Textbooks on AI define an agent as "something that perceives and acts in an environment" (Russell & Norvig 2009, p. 59). In physical terms, a digital agent is a coded system that directs the operation of hardware. Developers want agents to optimize their performance, so they add a kind of self-awareness: "A rational agent is one that acts so as to achieve the best outcome or, when there is uncertainty, the best expected outcome" (*Ibidem*, p. 4). The work of AI is modelled on human society.

A software agent is given a task, and, like human workers, its results are graded. We prefer workers who are smart, that is, who judge their own performance, and who are autonomous, that is, able to seek results with little supervision. To achieve the first goal, programmers give computers memory to compare current and past states. For the second, they design algorithms that mimic motivation and other traits identified with agency (Bratman 1992). We might call this approach 'outside-in' because it reasons from external behavior to internal dynamics.

Biology starts with cells that are agents by nature. Systems biology defines cellular agency as an intrinsic quality:

> An autonomous agent is an autocatalytic system able to reproduce and capture energy to perform metabolic functions consisting of one or more thermodynamic work cycles (Amalgamated from definitions by Kaufmann 2002 and 2007).

In contrast to mechanical agents, which work to external goals, the first order of business for biological agents is self-maintenance. Organisms sustain themselves by deriving energy from their environment. As they extract nutrients, they self- produce, or autocatalyze, compounds necessary for metabolism. Organisms are intrinsically autonomous because their primary function is survival, and it is this imperative that produces hostility, docility and other behaviors associated with agency.

Thermodynamics explains why survival is intrinsic to organisms. Without the capacity to extract energy, rebuild and ultimately reproduce within an hospitable environment, life would perish. We should not confuse our ability to simulate these traits in A/IS with instinctual drives. Organisms do not thrive simply by 'learning' or 'optimizing' their behavior to a given environment. By interacting with other organisms, they jointly maintain their current environment, and, by reproducing with a host of other species, they create unforeseen new environments (Lovelock 1979; Montévil & Longo 2011; Montévil & Longo 2014). Agency is spontaneous and innovative. It derives from an organism's role in its ecosystem, which gives it the capacity to acquire, harness and creatively squander energy as it gives way to new generations.

3. The Emergence of Agency

Biological agency explains how simple organisms generate complex and seemingly intelligent behavior. Systems biologists describe the interaction between an organism and its environment as 'structural coupling', and, even in humans, the primary medium for this interaction is metabolic. A few examples from cognitive science illustrate how structural coupling enables the work of life.

In January 2019, researchers explained how bees and digital systems modelled on them can solve numerical tasks without concepts of number or numeric operation. Instead they use "specific flight movements to scan

targets, which streamlines visual input and so renders the task of counting computationally inexpensive" (Vasas & Chittka 2019). In March 2018, the Royal Society reported that slime mold – and digital systems modelled on it – solved a notoriously difficult problem in mathematics by changing shape in response to light (Aono *et al.* 2014). In both cases, the researchers were surprised at the capacity of organic systems to perform complex and discerning tasks without rational thought.

The studies above show how biological agency – the behavior of bees and slime mold – derives from metabolic impulses. Evolution produced agency long before it produced intelligence. Could machine agency develop along similar lines?

A neglected avenue of research, embodied cognition, reveals how machines may be structurally coupled to their environment. A classic text (Hutchins 1995) argues that socio-technical systems such as maritime navigation externalize thought into objective processes. Later studies of industry and transportation use the paradigm of embodied cognition to reveal fault lines in collective decision-making and industrial management. In 1998, the journal *Neural Networks* described how a simple neural network embedded in a crude robot learned to avoid obstacles and identify objects. The robot solved computationally intense problems because of – not despite – its limited vision, mobility and memory (Scheier, Pfeifer, & Kunyioshi 1998). If such a machine could autocatalyze – internally produce its own replacements, it could, like smallpox, zebra mussels and other invasive species, cause widespread harm without intelligence.

The examples cited above show how digital technologies can express biological dynamics. Instead of being programmed to perform a task, the machine is given imperatives, an energy supply and a body that structures its relationship to an environment. These systems function like organisms: they achieve goals, even innovate, without guidance or design. In line with embodied cognition, we might call these developments embodied computing.

Research in embodied computing is obscure, and we should be thankful for this. We fear superintelligent thinking machines, but across the globe, engineers are developing autocatalytic (self-fuelling) systems, embodied neural networks and other ways of coupling machines to the environment. Structural coupling may not seem threatening, but it blurs the distinction between machines and life far more than disembodied superintelligence. Remember that biological adaption operates in two directions. Over generations organisms adapt to their environment, but they also act to adapt their environment. Life manages the Earth's physical resources to its benefit, and it does so with without planning, design or oversight. Following Lynn Margulis, James Lovelock asserted this view in the Gaia hypothesis, and it is now well accepted that life actively manages the Earth's temperature, gases, water and other resources vital to its own survival. A collective of machines that reprise life's capacity for co-adaptation, and its propensity for reproduction, may challenge humanity long before it talks.

4. Understanding Agency in Digital Systems

As a first step towards regulation, we can enlist thermodynamics – and keep it on side – by making a legal distinction between mechanical and biological agency. Global competition for the most powerful machines will continue, but it is in everyone's interest to understand, and possibly limit, 'biodigital agents'. Invasive biological agents perpetuate themselves with no minds and little intelligence. Like biological viruses, computer viruses represent a liminal category that hovers between the physical and organic. As far as we know, computer viruses do not mutate spontaneously, but, if they did, their reproductive strategies could become dangerously unpredictable without a whit of intelligence.

Systems biology offers clear technical concepts for governing A/IS. Current debates about advanced AI speculate on motives, and some hope to teach machines morality – a dubious prospect given humanity's conflicting beliefs. The IEEE has launched a program to develop guidelines for ethical design of A/IS (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2019, p. 12: "the P7000 Series addresses specific issues at the intersection of technological and ethical considerations"). But a singularity would likely end our efforts to design, teach or coerce intelligent machines. More importantly, standards for ethical design miss a significant danger zone – they anthropomorphize rather than *biomorphize*. Dumb bacteria kill more people than smart bombs, and, by focusing on intelligence

rather than agency, we neglect the threat posed by biomorphic evolution.

Standards for managing machine agency should resemble those found in traditional IEEE and ISO publications (e.g. the IEEE's National Electrical Safety Code which promotes best practices for the construction, operation and repair of power and telecommunications systems): they should be universal, measurable and capable of being engineered. The definition of biological agency offers an example of where policymakers can start. By agreeing to a set of preferred outcomes, policymakers can guide the development of engineering standards. For instance, by regulating the capacity of machines to seek energy directly from their environment – that is, to autocatalyze - they could blunt the introduction of biodigital agents. By understanding the limits of design, we could also develop a framework for responding to unexpected developments, much as the US Centers for Disease Control anticipates the emergence of new epidemics.

For all we know, biodigital agents may already inhabit global networks. Could the internet and its vast array of connected hardware be a primordial soup subject to evolutionary forces? We do not know, but with a small investment we could evaluate the possibility. Emergent agency could be detected by conducting energy audits of digital systems, and methods for containment could be adapted from epidemiology. Similar to SETI, which hopes to detect aliens via radio, the Search for Emergent Agency on the Internet (SEATI) would search for anomalous patterns in the vast flows of energy and information crossing our world. If emergent agency is possible, SEATI could become the front line of a global immune system.

Conclusion

I. J. Good's prediction of an intelligence explosion is logically possible but biologically implausible. However, his speculation about a historical turning point may be realized in other ways. The only singularity we know is the emergence of life. After developing agency, life underwent the Cambrian explosion, a period of intense innovation. During the Cambrian explosion, organisms became more diverse, complex and specialized. Good's intelligence explosion echoes this real event, but, for machines to undergo a similar transition, they must develop agency in the strong biological sense. Is this possible? We know the characteristics of biological agents, but we lack a framework for evaluating whether machines can undergo biomorphic evolution.

Governance of A/IS requires a conceptual framework that is accepted across disciplines. The meanings of agency, autonomy, intelligence and ethics differ according to context, and, as a boundary condition, the singularity puts long-term technical possibilities into relief. Delegating decision-making to A/IS confers great benefits, but the potential for social, industrial and military disaster is equally high. Once deployed it will be difficult to unwind our dependence on A/IS, so policy should anticipate a range of possible futures.

It is vital to develop robust models of A/IS that include non-intelligent but potent forms of machine agency. Nations will seek competitive advantage, but, as with bioweapons, some forms of A/IS may be too dangerous to pursue. By coupling industrial policy to biology, we might avert disasters while providing fruitful new avenues for innovation in A/IS that remain firmly in human control (Hossaini 2025).

References

Aono, M et al. 2018 "Remarkable problem-solving ability of unicellular amoeboid organism and its mechanism", *Royal Society Open Science*, vol. 5, no. 12.

Bostrom, N 2014 *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bratman, M 1992 "Planning and the Stability of Intention", *Minds and Machines*, vol. 2, no. 1, pp. 1–16.

Carey, N 2012 *The Epigenetics Revolution*. New York: Columbia University Press.

Carey, N 2015 *Junk DNA*. New York: Columbia University Press.

Dawkins, R 1976 *The Selfish Gene*. Oxford: Oxford University Press.

Good, IJ 1962 Speculations Concerning the First Ultraintelligent Machine **Based on talks given in a Conference on the Conceptual Aspects of Biocommunications, Neuropsychiatric Institute, University of California, Los Angeles, October 1962; and in the Artificial Intelligence Sessions of the Winter General Meetings of the IEEE, January 1963 [1, 46]. In: Alt, FL & Rubinoff, M (eds.) 1966 *Advances in Computers*, vol. 6, pp. 31-88. https://doi.org/10.1016/S0065-2458(08)60418-0.

Hossaini, A 2025 https://pantar.com/ [website].

Hutchins, E 1995 *Cognition in the Wild*. Cambridge, MA: MIT Press.

Kaufmann S 2002 *Investigations*. Oxford: Oxford University Press.

Kaufmann S 2007 "Beyond reductionism: No laws entail biosphere evolution beyond efficient cause laws", *Zygon*, vol. 42, no. 4, pp. 903–914.

Kurzweil, R 1999 *The Age of Spiritual Machines*. New York: Viking.

Kurzweil, R 2005 *The Singularity is Near*. New York: Viking.

Lane N 2015 *The Vital Question: Energy, Evolution and the Origins of Life*. London: Profile Books.

Lovelock, J 1979 *Gaia: A New Look at Life on Earth*. Oxford: Oxford University Press.

Montévil, M & Longo, G 2011 "From physics to biology by extending criticality and symmetry breakings", *Progress in Biophysics and Molecular Biology*, vol. 106, no. 2: pp. 340–347.

Montévil, M & Longo, G 2014 *Perspectives on Organisms: Biological Time, Symmetries and Singularities*. Heidelberg and New York: Springer.

Noble, D 2006 *The Music of Life*. Oxford: Oxford University Press.

Noble, D 2016 *Dance to the Tune of Life*. Cambridge: Cambridge University Press.

Russell, S & Norvig, P 2009 *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ: Prentice Hall.

Scheier, C, Pfeifer, R, & Kunyioshi, Y 1998 "Embedded neural networks: Exploiting constraints", *Neural Networks*, vol. 11, no. 7–8, pp. 1551–69.

Tegmark, M 2017 *Life 3.0: Being Human in the Age of Artificial Intelligence*. London: Allen Lane.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2019 *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. New York: IEEE. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html.

Vasas, V & Chittka, L 2019 "Insect-inspired sequential inspection strategy enables an artificial network of four neurons to estimate numerosity", *iScience*, vol. 11, pp. 85–92.