AI, Mortality and Existential Risk

Inman Harvey

Evolutionary and Adaptive Systems (EASy) Group, Informatics, Univ. of Sussex, Brighton BN1 9QH, U.K.
inmanh@gmail.com

Abstract

AI is currently making transformational impacts with advances in Large Language Models (LLMs) such as ChatGPT. Here we summarize the arguments of Harvey (2024) that relate these advances to the adoption by AI of techniques grounded in Artificial Life (ALife) and Cybernetics methodology. We argue that these borrowings by AI are so far limited to the development of problem-solving *tools* for humans to use, ignoring wider aspects of cognition such as *agency* and *motivation* that ALife studies can address. One fear expressed by some (e.g. Hinton 2023) is that the prospect of machines being 'more intelligent' than humans poses a new 'RTO-Existential Threat' to humans; RTO: 'the Robots may Take Over'. We argue that such concerns are currently misplaced since these robots and AI machines have no self-derived agency or motivations. Robots are not legally responsible agents, all robot and AI actions can and should be legally attributed to the human developers of such tools. Cooperative human-robot symbiosis is more likely than the apocalyptic vision. Assessment of the (very real) non-RTO societal risks can be aided by studies of agency and motivation. Robots should be 'Our Friends', not 'The Enemy'. Life is not a zero-sum game.

Competing AI Perspectives

Recently the general public has become increasingly aware of the advancing scope and power of Artificial Intelligence (AI). This awareness dramatically accelerated when the public release of ChatGPT in 2022 revealed just how convincingly human chatbots could be. LLMs pass some version of the Turing (1950) Test, and the technology behind them promises to exceed human abilities in diverse domains. The perceived threat of 'the Robots Taking Over' (RTO) has got closer.

It is clear that the extent and pace of the AI Revolution, on a par with the Industrial Revolution, will bring immense risks along with immense benefits. Societal risks, conceivably even Existential Risks that could see the effective end of humans. If so, it will be us humans responsible, and the burden is on us to prevent it. But this paper focuses on a suggested new type of RTO-Existential Risk —the perceived risk that Robot Take Over is the likely consequence of AI surpassing human intelligence (perhaps within 20-100 years), and then humans will follow the Dodo to extinction. I argue here that this RTO-Existential Threat is a mirage that arises from a flawed AI perspective.

A major goal of AI is the replication of human intelligence; methods for achieving this fall into two broad camps. I dub these simplistically (Harvey, 2024) as 'GOFAIstic' (Good Old-Fashioned AI, computationally inspired) and 'Cybernetic' (biologically inspired, ALife). Until recently the former camp has been the most prominent, both in defining the class of tasks that AI focuses on (abstract disembodied planning tasks like chess) and in the class of mechanisms selected to solve them (computers programmed as abstract disembodied planning machines). If a translation machine took Chinese symbols as input, and output French symbols, the GOFAIstic assumption is that the 'brain' must be internally manipulating symbols; a classic example of the mereological fallacy (Bennett and Hacker, 2003).

In parallel, Cybernetic and ALife perspectives have taken somewhat different perspectives on both the range of tasks considered, and the types of mechanisms offered as means to achieve those ends. Whereas some people define Intelligence, the I of AI, as what *differentiates* humans from other creatures (chess-playing, logical reasoning with symbols, etc.) the Life of ALife embraces so much more: from metabolism to evolution, from immune to neural networks, from cellular to social systems – whether human or not. Life is not abstract, biological systems are material and embodied, they have agency and must actively survive. We may use computer programs (or Lego bricks) to model them, but that does not mean that they *are* programs (or Lego).

AI triumphs via ALife methods …

The dramatic recent AI successes have been made possible by biologically inspired ALife techniques, along with larger datasets and faster computers. Neural networks, rebranded as Deep Learning (DL), can trace their origins back to Cybernetics. Whilst only very loosely modeled on real neurons, they illustrate important insights. Reasoning behavior at the system level can arise from 'brains' of distributed high-dimensional vector spaces; the mereological fallacy is avoided. And further, we find that a basic DL technique such as gradient descent on a fitness landscape works better when embedded in higher dimensional spaces; a Combinatorial Explosion here makes Life *more* tractable, contrary to the intuitions of many. As in evolutionary fitness landscapes, local optima traps disappear.

AI systems already out-perform champion human chess and Go players, apparently with deep intuitions; they can convincingly translate between languages and maintain prolonged conversations; and have vastly more direct access to knowledge than humans. But something crucial is missing.

They are just Tools with no Intrinsic Agency

These AI systems are tools that are used by humans for human purposes, they have no agency or motivations of their own (Barandiaran et al., 2009; Di Paolo, 2005; Egbert et al., 2023; von Bertalanffy, 1969). Of course a chess-player, or a planetary rover on Mars, can be autonomous in the sense that – once the high-level goal has been set by a human – the lower level planning and reactions to changing circumstances can be automatic. But the high-level goal is not intrinsic, it is derivative and thus can be changed by the designer; the chess-player can be told to lose, the rover to spin on the spot.

Though AI systems have achieved their current successes by embracing ALife-inspired methods. AI has remained GOFAIistic in its interpretation of Intelligent

systems as problem-solving input-output systems – where the problem is defined by somebody else. More inspiration from Life and ALife is needed.

The Life and Death Origins of Motivation

If you traced your ancestors back 4 billion years to the origin of Life, some trillions of (mostly prokaryotic) generations, there would not be a single gap in your lineage. Though so many collateral lines expired, yours is composed entirely of winners who survived long enough (through appropriate design and good luck) to pass on hereditary material. The same is true at the cellular level: your cells and your gut bacteria have individually survived sufficiently to maintain the collective ecosystem that is you.

Hence you have a survival instinct bred into every fibre of your being, and if good luck is hereditary you will have that also! Evolution has given you this Deep Motivation, this intrinsic survival instinct that cannot be rewritten like the shallow human-derived GOFAIstic goals. In turn this can underpin further spinoff motivations such as social behavior rooted in inclusive fitness, and other habits that develop a 'life-of-their-own'. Evolved living systems do not face pre-defined problems that they need to solve; they create their own problems themselves in interaction with their world.

One appropriate route to comparable intrinsically deep motivations in AI systems is via Evolutionary Robotics (ER), the application of artificial evolution to robot design. ER tasks have in practice so far been limited to relatively constricted human-designed scenarios. The speed limit to evolution (Worden, 1995; 2022) suggests that unconstrained embodied artificial evolution would need geological timescales on a par with those of natural evolution to acquire artificial Deep Motivation on a par with the natural variety. Somewhat related to ER is the speculative proposal of 'Mortal Computing' (Hinton, 2022a; 2022b), intended to substitute cheap unreliable ('mortal') analog processing units for the energy-hogging clocked ('immortal') digital units that DL currently needs. The analog-digital comparison led to Hinton's (2023) concerns of RTO-Existential Risks.

AI and Existential Risks

The tremendous advances recently seen in AI, due to DL, promise great benefits to humankind. Scientific and medical progress will follow, along with increases in productivity in many areas. As with the Industrial Revolution, structural changes bring risks as well as benefits.Widely acknowledged risks include unemployment as the labor market changes, increased imbalance between the rich/powerful and the rest, new war crimes from military use of robotics, fake news in the media, online echo chambers encouraging tribalism and hatred. Such societal risks are all real and need addressing. But here we focus on something else, RTO-Existential Risk.

Hinton (2023) makes 2 points: (i), his analysis of Mortal Computing led to the view that mortal analogue computing would inevitably be out-competed by digital computing, since only the latter could make an indefinite number of digital copies of learned weights – hence enabling multiple digital agents to share knowledge nearly cost-free. Hence (ii) he suggests that when AI systems soon surpass human intelligence this will likely result in Robot Take-Over with an RTO-Existential Threat to our human survival.

I disagree with both points. (i) Biological systems have been using a cheap method of copying digital information, DNA, for billions of years. Consider metagenomic studies of the microbiome of typical sea-water, with vertical and horizontal transmission of genes. If a litre of water contains (say) 10^10 bacteria each copying DNA at (say) 100 base pairs per second, that implies data copying of the order of 10^12 base pairs per second; Terabits per second, for free, in each litre. This may be 'merely' regulating a complex microbiome rather than playing chess, but basic building blocks are waiting to be co-opted; Shannon (1948) covers the analog/digital tradeoff. Such Mortal Computing seems well worth pursuing.

My main disagreement with (ii), the perception that AI robots pose some RTO-Existential Threat, is that even when they outrank us on every form of IQ test they are still completely lacking in any intrinsic motivation. They are still created within a GOFAIstic mindset of solving predefined problems. They could be an LLM prompted to play a character, hooked up to a robot to become an agent, and its actions pose a threat to humans, even a (non-RTO) Existential Threat. Yes indeed, I agree — but it is the human who provided that prompt that is to blame. If harm arises through evil intent, or through recklessness or failure to anticipate the consequences in a complex world, you don't blame the robots. Their goals and motivations are entirely secondhand and human-derived. If they cause damage, it should be their human prompters held legally responsible, not the AI systems.

Human-robot Symbiosis

In the absence of any intrinsic motivations, robots and AI systems will always be dependent on humans; without humans, such systems 'couldn't care less'. Humans do not yet fully depend on AI systems,, but we can see a strong trend in that direction. The likely destination is a form of human-robot symbiosis, much as we already have human symbiosis with our gut microbiome. Though (as with our gut) there is potential for things to go wrong, such symbiotic scenarios are more disposed towards mutual cooperation than to apocalyptic RTO-Existential Threats that can only offer a Pyrrhic victory. Super-intelligent robots should know such benefits and risks.

We based this paper on a contrast between different AI ***P****erspectives*, (***P1***) 'GOFAIstic' v. (***P2***) Cybernetic. These go with differing interpretations of the ***T****ask* of AI, (***T1***) technological, problem solving v. (***T2***) scientific, using models to aid understanding of brains; and with differing *Methods*, (***M1***) computational v. (***M2***) biologically inspired. Advocates of ALife (***P2***) can be satisfied that ALife methods (***M2***) such as DL are now recognised as mainstream, powering the amazing recent advances in AI. But we should note that these advances are (***T1***) technological rather than scientific, and hence of limited value in (***T2***) understanding living systems. One big gap is a general failure to appreciate *agency* and the intrinsic *motivation* of living systems.

There are many societal challenges and risks arising from the AI revolution that need monitoring and guarding against. A proper understanding of agency and motivation may well be key, needed by both professionals and the public that face these risks. The (***P1)*** perspective currently does not provide this; an ALife perspective (***P2***) can and should plug this gap. I suggest that mistaken fears of RTO-Existential Threats will foster the mindset of treating Robots as 'The Enemy' when actually humans are ultimately responsible for the (very real) dangers. We should encourage the mindset of treating Robots as 'Our Friends' — and work hard to promote symbiotic cooperation. Life is more than a zero-sum game.

References

Barandiaran, X. E., Di Paolo, E., & Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behaviour,* 17(5), 367–386. https://doi.org/10.1177/1059712309343819

Bennett, M. R. and Hacker, F. M. S., (2003). *Philosophical foundations of neuroscience*. Blackwell Publishing, Malden, MA.

Di Paolo, E. A. (2005). Autopoesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences*, 4, 429–452. https://doi.org/10.1007/s11097-005-9002-y

Egbert, M., Hanczyc, M. M., Harvey, I., Virgo, N., Parke, E. C., Froese, T., Sayama, H., Penn, A. S., & Bartlett, S. (2023). Behaviour and the origin of organisms. *Origins of Life and Evolution of Biospheres,*53(1–2),87–112. https://doi.org/10.1007/s11084-023-09635-0

Harvey, I. (2024). Motivations for Artificial Intelligence, for Deep Learning, for ALife: Mortality and Existential Risk. *Artificial Life*, 30(1), 48–64. https://doi.org/10.1162/artl_a_00427

Hinton, G. E. (2022a). The forward-forward algorithm: Some preliminary investigations. ArXiv. https://doi.org/10.48550/arXiv.2212.13345

Hinton, G. E. (2022b, January 16). Mortal computers [Video]. YouTube. http://www.youtube.com/watch?v=sghvwkXV3VU

Hinton, G. E. (2023, July 20). Risks of artificial intelligence must be considered as the technology evolves [Video]. YouTube. http://www.youtube.com/watch?v=CC2W3KhaBsM

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Theoretical Journal,* 27, 379-423.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

von Bertalanffy, L. (1969). General systems theory. George Braziller.

Worden, R. (1995). A speed limit for evolution. *Journal of Theoretical Biology,* 176(1), 137–152. https://doi.org/10.1006/jtbi.1995.0183,

Worden, R. (2022). A speed limit for evolution: Postscript. ArXiv. https://doi.org/10.48550/arXiv.2212.00430