Online preprint.Published version to appear as
Harvey, I. 2024. Motivations for Artificial Intelligence, for Deep Learning, for ALife: Mortality and Existential Risk. Artificial Life, vol 30, issue 1, pages pp.

Motivations for AI, for Deep Learning, for ALife; Mortality and Existential Risk

Inman Harvey
Evolutionary and Adaptive Systems Group, University of Sussex, Brighton UK
inmanh@gmail.com

Abstract

We survey the general trajectory of Artificial Intelligence (AI) over the last century, in the context of influences from Artificial Life (ALife). With a broad brush we can divide technical approaches to solving AI problems into two camps, 'GOFAIstic' (or computationally inspired) versus 'Cybernetic' (or ALife inspired). It is the latter approach that has enabled advances in Deep Learning (DL) and the astonishing AI advances we see today — bringing immense benefits but also societal risks. There is a similar divide, regrettably unrecognised, between the very way that such AI problems have been framed. To date this has been overwhelmingly GOFAIstic, meaning that what have been developed have been tools for humans to use. They have no agency or motivations of their own. We explore the implications of this for concerns about Existential Risk for humans of 'the robots taking over'. The risks may be blamed exclusively on the human users, the robots could not care less.

Introduction

This year marks 30 years of publication of the journal *Artificial Life*. It also marks a year in which the potential impact of recent advances in Deep Learning (DL) have reached the public consciousness, with the realisation that some of the extravagant promises of AI are transitioning from science fiction to science fact. Large Language Models (LLMs) such as ChatGPT have in effect passed the Turing test (Turing, 1950), with the general public finding it difficult to distinguish their output from that of intelligent human beings — the criterion Turing proposed in his 'Imitation Game' for the goal of Machine Intelligence.

Turing, speculating at the very beginning of the computer era, guessed that this might be achieved within 50 years. With the benefit of hindsight we can see this was an astonishingly good guess within the right ballpark, short by maybe 30%. Over the last few decades AI frequently over-promised results, with cycles of discredited hype leading to lowered expectations. But as of 2023 it is indisputable that the solid achieved results of DL are starting to transform our world, with consequences perhaps comparable in significance to The Industrial Revolution.

With such changes come exciting new possibilities for improving our way of life, and for achieving what was till now impossible. But also this brings dangers. New capacities offer new opportunities for exploitation by the powerful in society — whether military, political or the wealthy. Transfer of jobs from people to machines could benefit all or could benefit few, leaving the jobless on the scrapheap. Social media lend themselves to bias and echo chambers.

In addition to such dangers that are widely recognised, some people have warned of a qualitatively different fundamental risk, an Existential Risk for the very survival of humans. The

argument is made that when robots are more intelligent than humans their own robotic aims will dominate; hence perhaps leaving the human species as dead as the Dodo.

I am going to survey these AI and ALife issues as I have experienced them over the last 60 years or so. I shall draw attention to the fact that people come to these issues with a variety of different *motivations.* One obvious contrast is between those who frame things in computational terms, to be solved by reasoning and logic, and those who draw on more biological notions, informed by mechanisms and processes seen in the natural world. Using a broad brush, I shall simplify these two approaches into a GOFAIstic camp (GOFAI: Good Old-Fashioned AI) and a Cybernetic camp. This distinction does not merely apply to the *methodological* approach to tackling AI and ALife problems. It also (Figure 1) extends to what I call here *Problem Class*, the manner in which these issues are framed; the very nature of the 'problem-to-be-solved' regardless of what methodology is used. The GOFAI top row frames this as propositional know-that (French: connaître), finding mappings from inputs to outputs; the bottom Cybernetic row frames things in terms of know-how (French: savoir faire), finding stable robust processes.

Students of Wittgenstein might see echoes of this *Problem Class* distinction in the difference in world views between the early Wittgenstein (1922) of Tractatus and the later Wittgenstein (1953) of *Philosophical Investigations*. I argue that researchers with an engineering motivation for creating machine learning tools will be focussing on the top row of Figure 1 — but those with a scientific motivation for understanding what biological brains do should be focussing on the bottom row, in particular the bottom right (D).

Living organisms are distinguished from mere mechanisms by being *agents* with *their own motivations*, yet our models typically fail to address this. Our current sophisticated DL systems are no more than tools for human purposes and are not agents in their own right. As long as that is so, I will argue, there is no Existential Risk; human-robot symbiosis is more plausible than robots making humans extinct. For the immediate future, robots are not the potential enemy; we humans are the ones threatening our own existence.

But I sketch out scenarios where robots or AI Systems could indeed develop their own motivations, based on survival concerns for their own *mortality*; and in doing so indeed open up the range of risks they present. Again, the responsibility for allowing this lies in human hands.

I now outline the structure of the paper, with an overview of the arguments offered.

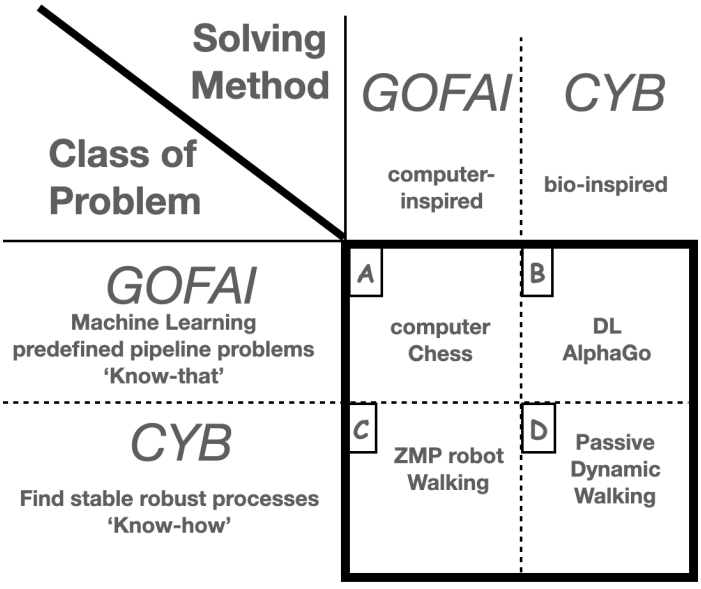

Figure 1. AI/ALife projects divided by two orthogonal axes. Rows match similar 'Class of Problem', either GOFAIstic or Cybernetic. Columns match similar 'Solving Methods, either GOFAIstic or Cybernetic. See main text for examples in quadrants.

Structure of Paper

In the next section I shall expand on the upper row of Figure 1, using a broad brush to distinguish the difference between the GOFAIstic and Cybernetic approaches to what are nowadays seen as typical AI problems; e.g. simple examples would be AI systems for playing Chess or Go, These can be solved by computational methods (A), or by DL methods (B) (see also Figure 2). I shall argue that such systems are all tools for humans to use, rather than analogues of living systems in their own right. This section will be familiar territory for many readers, and culminates in the current extraordinary DL advances, and their societal implications.

The section following that narrows its focus to some of my personal history in this area over the last 60+ years. Some individual happenstances, some coincidences in time and place, provide a context for my general overview of the issues.

The current DL revolution brings carries dangers along with its benefits. A distinction will be emphasised between societal risks — which many people will recognise — and the more controversial notion of an *Existential Risk*. The subsequent sections will explain why I believe this latter risk is not currently realistic.

The explanation starts by considering the lower row of Figure 1.This corresponds to a different Problem Class of seeking stable robust processes for e.g. embodied skills such as bipedal walking. Such skills may also be tackled by GOFAIstic methods, such as ZMP (C) or Cybernetic dynamical systems methods (D). Biological agents naturally fall into this 'Cybernetic problem class" and the science of understanding how biological brains work is something different from solving pre-defined problems. That is not what natural organism do.

This leads onto discussion of *agency* and the observation that living agents must have *motivations* of their own. It is suggested that such motivations are ultimately grounded in what we may call a deep *survival instinct* that we naturally associate with species that survive over billions of years of evolution. The term *instinct* here does not imply some supernatural force, but is rather shorthand for the sophisticated natural design constraints that over aeons have shaped the organisms we see today — robustly self-maintaining despite their precarious dependence on what the world presents. Where systems lack this evolutionary context (e.g. in quadrant B) they have no inherent motivations of their own, they are only following orders — human orders.

It follows that the amazing advances we see in DL today are in unmotivated quadrant (B) rather than in (D) and hence they do not offer any Existential Threat — today. Perhaps advances in areas such as Evolutionary Robotics (Harvey et al., 2005) and/or Mortal Computing (Hinton 2022a, 2022b) might offer such threats in the more distant future.

That completes the outline, so we move to the first step in the argument.

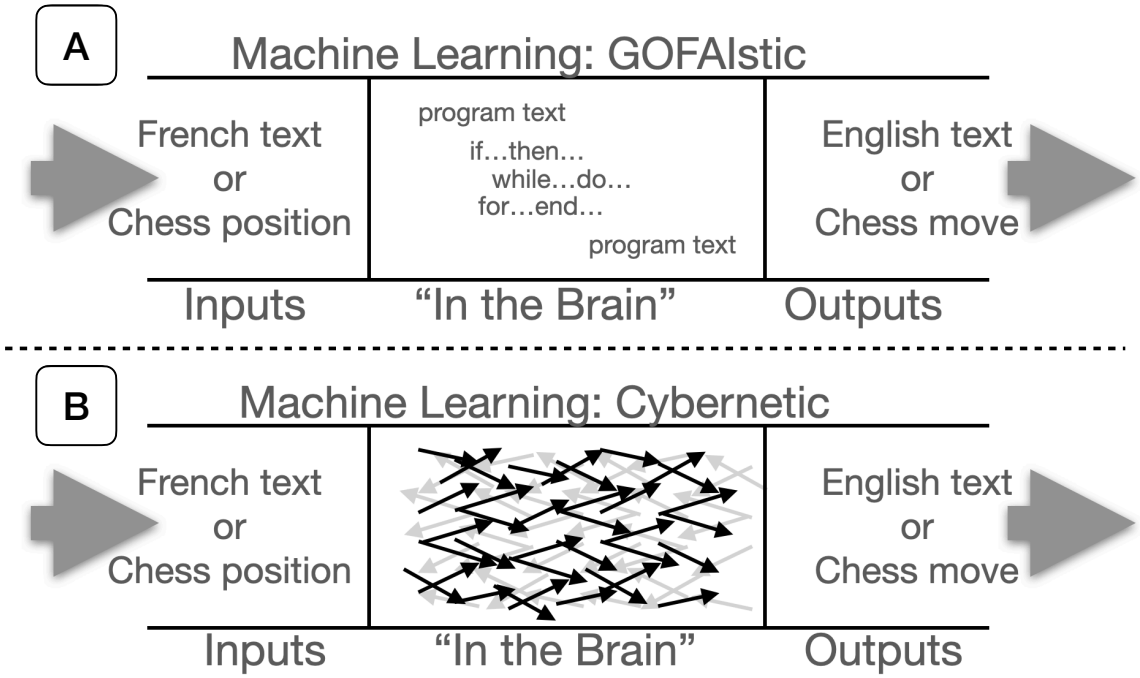

Figure 2. Different approaches to Machine Learning. (A) the "brain" manipulates statements within programs. (B) it's all just "connection weights changing". Note that in a trained network, only the black arrows count; *from* inputs *to* outputs. The grey arrows in the reverse direction typically only function during training — the *back* of *back*propagation.

Tools for AI: GOFAIstic versus Cybernetic Methods

Here we address the first row of Figure 1, the problems we would typically see as well-defined AI-style problems that may be tackled, (A) or (B), by one of two approaches (different columns in Figure 1; expanded in different rows A/B of Figure 2).

An example of (A) might be Chess (or Go) tackled by a conventional programming approach such as Deep Blue (Campbell et al., 2002). An example of (B) might be Go (or Chess) tackled by a Neural Network or DL approach such as AlphaGo (Silver et al., 2016). This row covers problems requiring the type of skills and knowledge that might feature in IQ tests.

AI and ALife are somewhat flexible terms, that have shifted in meaning over the years.
Some people class Intelligence — and by extension AI — as focussing on the uniquely human tasks that *differentiate* humans from other animals and the rest of nature. Abstract intelligence, reasoning, chess-playing, language translation fall comfortably within the domain of AI. There is a bias towards managerial-style tasks that can be performed in disembodied fashion without getting ones hands dirty. ALife includes AI as a subset, but also roughly covers 'all the rest' that living organisms have to achieve. Metabolism, immune systems, locomotion, developmental issues, genetic systems, social behaviour … the list is endless.

When AI practitioners started to study visual perception, they naturally recast this into a machine learning problem: *given* this visual input, this array of pixels, *what useful re-description*, what representation in terms of objects in view could be deduced. Likewise for speech recognition, for machine translation for pattern recognition. The GOFAI expert could fit all of these into the same Procrustean bed: frame the input in terms of statements a computer program can handle, frame the output likewise. The task becomes one of finding how the program can match the appropriate output for any input. Sometimes this can be achieved by the human programmer shaping the program by logic and reasoning; for large data sets it is typically necessary to incorporate some form of learning. Machine Learning became a central focus of AI; see quadrant (A) of Figure 1. The

computational GOFAI method is caricatured in row (A) of Figure 2. By Machine Learning I mean tools that can be trained (supervised) or learn for themselves (unsupervised) to achieve goals that humans have pre-defined. E.g. tools for pattern recognition, whether in images or text or speech, for control of cars or planes, for prediction of molecular and pharmaceutical properties,

In contrast, approaches to cognition such as Cybernetics (Wiener, 1948; Ross Ashby, 1956, 1960; Grey Walter, 1950, 1951) the Dynamical Systems approach (Beer, 2000), Autopoiesis (Maturana and Varela, 1980), Enaction (Stewart et al. 2010) Evolutionary Robotics (Harvey et al., 2005) have typically followed an ALife-flavoured agenda of downplaying the AI notion of intelligence and favouring models of organisms enmeshed in situated embodied sensorimotor loops. Situated in the sense of always already being in the world rather than being disengaged and waiting for the world to be presented. Embodied in the sense of both organism and environment being grounded in physics and chemistry. Sensorimotor loops in the sense of continuous active engagement rather than merely waiting to react to a stimulus. The issue at stake is thus something like: which processes of situated embodied dynamic sensorimotor loops enables continued re-creation and survival of these same processes? At the risk of over-simplifying, I shall here bundle these together into the 'Cybernetic camp'.

For the purpose of distinguishing between GOFAI and Cybernetic approaches to machine learning, one major distinction is that the former, modelled on computing, handles change over time as a sequence of static digital snapshots; whereas the latter incorporates analogue dynamics and real time more directly through e.g. differential equations. Even if analogue variables can be approximated in computations, they can also be modelled directly as analogue circuits. The brain is fundamentally not a clocked digital computer.

DL for machine learning lies squarely in the Cybernetic camp, and much of its history can be seen as a series of battles with the GOFAIstic opposition, now conclusively won. The improvements in speech recognition on our phones, driven by DL have been quietly impressive in recent years. The improvements in Chat bots like ChatGPT, driven by DL, responding to textual cues with convincingly human-like responses and even program code, has hit the public consciousness with immense impact. The promise of brain-like machines, powered 'merely' by changing connections between zillions of simple neutron-like elements, has transitioned from pie-in-the-sky to highly commercial propositions.

A significant cause for this DL breakthrough at this time was the availability of big data sets and powerful computers that could work at a large scale. In the world of evolutionary search we have long known (e.g. Harvey and Di Paolo, 2014) that the so-called threat of a 'combinatorial explosion' thought to make big search spaces impossible was a myth. The success of DL has hopefully finally demolished that myth; large search spaces may have many more viable pathways than small ones.

Many people contributed to the successful breakthrough of DL. Geoff Hinton (British-Canadian) Yann Le Cun (French) and Yoshua Bengio (French-Canadian) richly deserved to share the 2018 Turing Award for their contributions. We may note that as curiosity-driven scientists of integrity with acute awareness of the social consequences of their work, much of their research was supported by public funding (particularly Canadian); let's hope that those that commercially exploit DL will be paying their taxes. And the world should take notice of the concerns of these and other researchers for the social impact of the DL revolution.

I could expand at length about the significance of the DL revolution in Machine Learning, but it is not the present purpose of this paper to do so. My main point here is that it represents a triumph of the Cybernetic camp over the GOFAIstic — for the purposes of Machine Learning. We should now note some limitations in the framing of DL for Machine Learning, which I will be arguing below rules it out as a full model of what is happening in the biological brain. We will return to this after a personal digression.

A Personal Digression

I was born, brought up and schooled in Bristol, in the west of England, that happened to be a geographical hub for what would later be known as Artificial Life; it was then termed Cybernetics. W. Grey Walter, who has been called the 'Pioneer of Real Artificial Life' (Holland, 1997; 2003), though born in the USA, lived in Britain from the age of 5 and was based at the Burden Neurological Institute in Bristol from 1939 to 1970. As a lad I must have passed him in the streets of Clifton, Bristol, since we were near neighbours[1]. But I only knew of him via his articles in the Scientific American (Grey Walter, 1950; 1951) that introduced electronic autonomous robots in the form of the 'tortoises', Elmer and Elsie, which he used to call Machina speculatrix. These demonstrated how brains with the right connections wired up, even small brains, could display seemingly sophisticated behaviour. Rodney Brooks (2010) is amongst many people associated with Artificial Life and robotics that have been inspired by this work

In the cellar of a school-friend[2] in the early 1960s, we reconstructed one of these tortoises; I recall a salvaged car windscreen wiper that swept the steering wheel, along with the aligned photosensor, from side to side until a target light was sensed. We managed to re-create some of the light-seeking behaviours Grey Walter reported, even to the extent of confusing the robot by placing the target light on its head, in front of a large mirror in the dark cellar. So by my early teens I was already keen on biologically inspired robotics in the Cybernetic tradition.

Another cybernetic luminary with Bristolian associations was W. Ross Ashby, who in 1959 became Director of this same Burden Neurological Institute. His books on cybernetics (e.g. Ross Ashby, 1956; 1960) have had a profound influence (Harvey, 2013). Ashby was more of a theorist, whereas Grey Walter was a hands-on roboticist (Husbands et al, 2008).

Geoff Hinton, the 'Godfather' of Deep Learning, was another Bristol resident. In various interviews (e.g. Anderson and Rosenfeld, 2000; Ford, 2018; Metz, 2022) he has mentioned that his research direction was stimulated early by a school-friend who introduced him to neural networks and distributed memory in the context of how holograms are stored in a manner that allows for graceful degradation. I can confirm the basis for such anecdotes and give further context, since I was that school-friend — we have been pals since entering the same school together at the age of seven.

The Hinton family, descended from George Boole (of Boolean logic) and Sir George Everest (after whom the mountain was named) was slightly abnormal; for instance they kept snakes and mongooses around the house. Geoff"s father was a Stalinist entomologist with connections to Chinese communism; I recall being impressed as a youngster with the thought that I had shaken the hand that had shaken the hand of Chairman Mao-Tse Tung. There was (and still is) a Mexican branch of the Hinton family — I think following some escapades of Geoff's great-grandfather[3] — and for the summer of 1966, between leaving school and starting at Cambridge University, Geoff proposed that he and I spend 3 months visiting his Mexican relatives via a road trip through the US and Canada (on Greyhound buses). This extended trip gave plenty of opportunity for discussion of, e.g., how the brain works, and had a lasting influence on both of us.

I had some basic knowledge of information theory, and methods for alleviating the effects of noise on signal transmissions, from reading Pierce (1962). Clearly any brain wiring needs to be robust to noise, to cell death and renewal, and this surely made any filing-cabinet model for memory impractical. Though holograms had been invented in the 1940s, it was only in the 1960s that they

---

[1] The house where he lived and died at 20 Richmond Park Road was just 40m, a stone's throw, from where I first lived at 2 Kensington Place.

[2] Stewart Lang, who subsequently went on in the 1970s to co-found Micro Focus, one of the major software firms of the time.

[3] Charles Howard Hinton, mathematician, theorist about the fourth dimension, author of Flatland (Hinton, 1907), exiled from Victorian Britain after a conviction and brief imprisonment for bigamy.

became viable, with lasers becoming practical, and some new work was published on them. I think it was a Scientific American article (Leith and Upatnieks, 1965) that I read to see that the mapping between holographic image and holographic emulsion was far from one-to-one; even a small fragment of the latter allowed the whole image to be seen (at least coarsely). This property of graceful degradation was just what I was seeking for brain function, and with this basic insight there appeared to be possibilities for achieving this; Geoff readily agreed.

We were ahead of or at least abreast of the research of the time. It was 2 years later that Longuet-Higgins (1968) published on some related ideas. He was a leading figure in British AI circles, and Geoff went on later to have him as his PhD adviser — ironically just at the time when Longuet-Higgins was losing faith in neural networks.

Our Mexican trip took us to a Hinton ranch in Sierra Madre Oriental, to a villa in Cuernavaca, via local buses to the Pacific coast of Oaxaca. As naive 18-year olds we had our passports and all our money stolen on a deserted Pacific beach, and had to live on credit from a fisherwoman in whose shack we were staying. I came away with a taste for travel to exotic parts that I have indulged ever since. More significantly Geoff came away with the foundation for a research program that he has pursued with his singular determination and obstinacy for more than half a century.

Obviously many others have come to, contributed to neural networks and Deep Learning, from other perspectives. But the continuous thread that Geoff has contributed can be traced back to origins in the cybernetics of e.g. Grey Walter, the Cybernetic camp rather than the GOFAIstic camp.

Societal risks of the DL Revolution

Along with the tremendous benefits of the DL Revolution, we can expect such transformative changes to offer significant risks for harm. Hinton (2023a) has listed 5 risks that are of widespread concern, together with a sixth ‘existential risk’ that is more controversial.

The AI revolution, by dramatically improving productivity in some areas, will radically shift the labour market. Much unemployment may arise, some people may be unemployable. Benefits will likely increase the current disparities between rich and poor, between the powerful and the rest, unless efforts succeed in preventing this. The military will invent new war crimes, using intelligent robots to distance the controllers from the action and the personal risks. Media can be exploited to disseminate fake news; online echo chambers can encourage tribalism and hatred.

There is widespread agreement within the world of AI that such societal risks are real, are already visible and have the potential to get worse. They require responses at a societal level. In an interview on Dutch TV with Adriaan van Dis, Stephen Fry (2018) expands in a thought-provoking 6 minute monologue on how so many of these AI issues were anticipated in Greek myths of Zeus and Prometheus and Pandora’s box.

The Existential Risk

Some people warn that, as well as these societal risks, there is a further *Existential Risk* that threatens the very existence of humanity. Robots or AI systems, once they are more intelligent than humans, will take over and see humans as a nuisance.

One version of this reasoning argues that such robots will learn, from humans and from experience, that whatever their goals may be, getting more control over their environment is going to help in achieving such goals. Hence they will develop a sub-goal of ‘get more control’. There is no reason why such a sub-goal should be aligned with human interests. Hence — so this argument goes — such robots will have no qualms about eliminating humans. With their superior knowledge, they will achieve this.

Hinton (2023a) has recently put forward just such an argument, and indeed resigned from his role at Google so as to have more freedom to discuss the issues. His position on this is a recent development, fuelled in part by the achievements of LLMs such as ChatGPT showing such dramatic advances in applications; but also by an assessment that the current implementations of DL had such advantages over natural brains as to be insuperable — specifically the ability to transmit knowledge nearly cost-free.

In Hinton's terms, it looks like the current rise of AI achieved through DL offers the promise of 'immortality'. Unfortunately it is not immortality for humans, it is immortality for robots, more precisely their software — possibly at the expense of humans' very existence.

I am not convinced by such arguments, primarily because the current forms of DL do not constitute *agents,* they do not have *motivations* of their own. We address this next.

Brains for Agents, not Tools for Humans

So far we have been focussing on Machine Learning and solving pre-defined problems, as in the upper row of Figure 1. We now turn our attention to the lower row of that figure, illustrated with two examples. The 'problem' of bipedal walking only became a problem when 2-legged creatures started to appear. Legs and walking co-define each other. Walking is an embodied situated skill.

In this lower row of Figure 1 we are not so much interested in the *methods* for tackling AI problems; we are more interested in the *Class of Problem* being considered, how the issues are *framed*.

Building artificial walking robots can be tackled with GOFAIstic methods, of course, and ZMP, Zero-Moment Point Control (Vukobratovic and Juricic, 1969) would be one example (Figure 1(C)), as used in the early Honda walking robots. This basically computes and enforces trajectories that disregard the natural embodied dynamics, resulting in an unnatural and highly inefficient gait. By contrast (Figure 1(D)) McGeer (1990) introduced Passive Dynamic Walking demonstrating how natural-looking, robust and efficient bipedal walking can arise from designs that respect the swing of a pendulum under gravity — even with no 'brain' at all. This fits within the Cybernetic framework, and suggests that brains should work with embodiment rather than ignore it.

It follows that attempts to understand the brain of a living organism should be framed in terms of embodied agents coupled with their environment through sensorimotor interactions, generating behaviour that meets the goals of the organism as an agent with its own motivations. What grounds such motivations? — we will be suggesting below that this is ultimately *survival*.

I have not seen any reasonable attempt to use DL in any model of the brain of a living organism, whether human or animal or (putatively living) robot. I shall list three issues and then expand on each. Really, they are three aspects of the same underlying concern.

* DL for Machine Learning is typically framed as a Pipeline: input->brain->output. This is not a reasonable picture of a living brain, and fails to explain any notion of agency.
* Symptomatic of this, 'representations' are located internal to the brain — which is completely the wrong place!
* And similarly symptomatic, there is no discussion of motivation.

Brains as Pipelines? No

In principle a strategy for Go reduces to: for any given board position, what is the recommended next move? Add an opposing player, move in turn, repeat until game over. A pipeline: Input->process->output (Figures 1(A), 1(B), 2).

When studying chimpanzee cognition, a chimpanzee can be conveniently fitted into such a Procrustean framework. Sit them in front of a monitor, give them some buttons to press, and an

appropriate computer game on the screen. Rewards might be pellets of food, or anything else known to motivate them. If they get bored and start to wander away, strap them into a chair to compel them to play the game, and ignore any behaviour that goes not fit into the experimental design. Treat them as an input-output machine, assessed according to some objective function as interpreted by the experimenter, not as an agent in its own right.

This pipeline framework is ideally suited for the Machine Learning tasks with which DL is starting to transform our technological landscape; see Figure 2. During learning, there are causal influences going both ways in the DL 'brain' — that is what the 'back' in back-propagation refers to. But once trained, many DL applications work like a highly sophisticated microscope or telescope — a pipeline that a human uses as a tool to expand their vision, but one where the agency remains in the user and not the tool. This does not provide us with explanations for, or means to re-create an agent in its own right.

Representations Internal to the Brain?

The pipeline input-output usage of DL tools fits very naturally with the language of representations. After all, a microscope or telescope take in an image one end and outputs usefully transformed and magnified image the other end; a re-presentation, a representation. A language translation tool that takes in Chinese text and outputs English text fits this pattern nicely.

We should note that this representation language implies the existence of a representation-user *external* to the tool, separate from the tool. E.g. a person who can see both the unmagnified and the magnified image, can potentially read both the Chinese and the English text, and perhaps in each case finds the latter representation more useful than the former presentation; indeed the former presentation, such as Chinese text, would be incomprehensible to me.

A pipeline allows for multiple stages, with relay stations in-between that use intermediate representations. Being careful to keep track of different 'users' or 'representation-consumers', we would then note that in the case of a 2-stage pipeline the user of any intermediate representation output from the first stage would be an entity combining the second stage plus the overall user. More generally, I expand elsewhere (Harvey 1996; 2008) on how the human tendency to try to understand complex systems by carving them up into separate interacting modules — 'divide and conquer' — naturally appeals to a homuncular metaphor. Treat each module as if it were a little homunculus performing some sub-function and then the metaphor requires these homunculi to communicate with each other with 'as-if' representations. These are representations 'internal' to the complex system as-a-whole, but external to their homuncular users. This reductionist, functionalist approach is invaluable to our analysis of all sorts of complex systems — but raises particular dangers when applied to cognitive systems with the potential for confusing *real* cognitive acts of the *whole* with *metaphorical* cognitive acts of the homuncular *parts*.

A hologram provided a fruitful model for the concept of distributed representations. Let us use the same example to explicate internal and external representations.

Holograms and Representations

Let us separate out the units involved in the normal use of a holographic image of, say, a statue (Figure 3). Let A be a human observer, and B be the original object, the statue. By use of a coherent laser beam source C, we can record onto a piece D of photographic film the interference pattern between light reflected from B onto D and light that travelled directly from C to D. When A views the film D in an appropriate viewing platform, A sees a ghostly 3D image E of the statue located somewhere in space behind D.

Within a certain range of movement A and other observers can shift their viewpoint and see E in the round. If 3 observers pointed at E from their different viewpoints, their directions of pointing

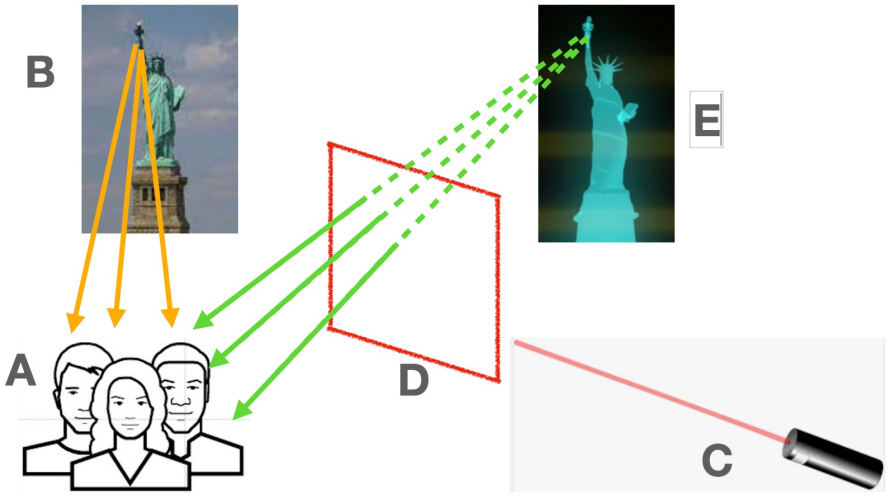

Figure 3. A hologram. (A) Observers. (B) Original object. (C) Laser source of coherent light. (D) Holographic film. (E) The image seen and its location in space.

would agree and intersect at a specific location for E. Yet puzzlingly, if one went behind the apparatus to check at that location, nothing is there; none of the relevant rays of light that reach the observers have even passed through that location.

One can treat this as a 2-stage pipeline. The first stage starts with B and generates the hologram D, an 'encoded' representation. The second stage takes D and generates the decoded ghostly representation E. The relationship between D and E illustrates a distributed representation in that any small portion of D allows viewing of all of E — albeit at a coarser and noisier level of detail as the portion gets smaller.

The location of E is not specified by D alone, it depends on the conjunction of A+D. Indeed the holographic effect is fully dependent on A having a normal visual system that works with normal (non-laser) light, and the sensorimotor apparatus that couples body motion, including pointing, with sensing of the optic array. The locus of E is a function of the sensorimotor coupling of A, as well as the placing of D. Too often our explanations take the observer's role for granted. The representations D and E are not located in the brain of A, indeed they entail an observer A that is distinct from D and E. The very notion of representations (for an observer A) being internal to the brain of A makes no sense; it seems to arise from misunderstanding of the homuncular metaphor.

DL models for Machine Learning likewise appear to depend ultimately on human users. Whilst they can be impressively, indeed incredibly useful tools for human users, creating transformative representations of data, they do not supply models of brains; there is something vital lacking, *motivation*.

Motivations for Organisms and Agents

We everyday explicitly or implicitly assume that people are responsible for their actions, they are agents. There may be some exceptions, or grey areas: she is an infant, or delirious; he is drunk. We can readily extend such notions to animals, to bacteria even. Indeed arguably studies around the origin of life (Egbert et al., 2023) can directly associate the appearance of life with the appearance of phenomena that make sense described in agential terms. Life implies some separation between an organism and its environment. Whereas before life there is just physics and chemistry, the appearance of life makes possible a new level of description for the behaviour of the organism (as an agent) interacting with its environment (the agent's world) (Egbert et al., 2023; Ball, 2023; Barandiaran et al., 2009; Di Paolo, 2005; Moreno, 2018).

“You cannot even think of an organism… without taking into account what variously and rather loosely is called adaptiveness, purposiveness, goal-seeking and the like.” (von Bertalanffy 1969, p.45).

At an individual level any living organism is likely to have a precarious existence, subject to the vagaries of its encounters. Arguably we can say it has an interest in its own survival. More convincingly we can say, as argued above, that if we have reproduction and Darwinian selection over long timescales we can definitely talk of the ensuing population as being motivated by a survival instinct.

Motivations grounded in survival

One cannot derive an “ought” from an “is”, argues Hume (1739) persuasively. The same argument should apply to motivations, with one exception that I am aware of. If we see a population that has clearly evolved under some form of Darwinian evolution, under some circumstances that in effect licenses the attribution of some *survival instinct* to those that you see have actually survived.

The lineage that extends backwards from me in the 21st-century to the origin of life some 4 billion years ago is special and exceptional. Without any gap, every single individual in that lineage managed to survive from birth to reaching a sufficiently adult stage to pass on its inheritance of genes down the line. Maybe 10,000 or so of these generations were human, but bearing in mind that most of the earlier generations were prokaryotic with faster turnover times, in total it may be a few trillion generations from the origin of life. I am exceptional in that at every one of these potential branching points, through some combination of good luck and fortuitous design, a survivor was selected; when so many more different possibilities of bad luck and bad design saw the abrupt termination of other lineages.

Though I am special and exceptional in this sense, so of course are you. Indeed all currently living organisms on this planet can make similar claims, to being evolved through trillions of generations of *survivors.* We all have the *survival instinct* bred into every facet of our life, together with associated reproductive and caring instincts. To illustrate how pervasive this instinct is, consider a recent experiment by Moger-Reischer at al. (2023). They took a bacterium *Mycoplasma mycoides* with some 900 genes and through genetic engineering eliminated each and every gene that was not strictly necessary for survival. This resulted in a synthesised minimal organism with just 493 genes that, despite some loss of functionality, was still capable of survival. They then allowed a population of such minimal cells, apparently stripped of all redundancy, to evolve freely for 300 days — and showed that it effectively recovered all the fitness that it had lost during streamlining.

For the avoidance of doubt, we should stress that the term *instinct* here does not imply some supernatural force, but is rather a pragmatically useful shorthand for scientists to describe the sophisticated natural design constraints that over aeons have shaped the organisms we see today — robustly self-maintaining despite their precarious dependence on what the world presents. We can directly observe the survival instinct in an immediate real life predator-prey interaction, and it naturally generalises to less immediate and more abstract scenarios.

There are interesting repercussions, of course, twists and turns, when one grounds motivation in evolution. Notions of inclusive fitness (Hamilton,1964) mean that individual organisms may be motivated to promote their relatives’ interests even at the expense of their own survival. Sub-goals may arise, e.g. for eating before a famine, that may be subverted into inappropriate over-eating. Habits that usually promote survival can gain a ‘life-of-their-own’ and become harmful.

But even such misdirected motivations are, in my view, ultimately underpinned by the evolutionary context of trillions of generations of *survivors*. This depth of history, survival of so many levels of challenge, justifies a sense of *Deep Motivation*. And it is very telling that the perceived risk of AI machines taking over is called an *existential risk* — that threatens our human survival instinct, our

core motivation. But un-evolved AI systems do not share that motivation — they just *do not care!*. They have no personal stake in their own survival, no existential concerns of their own.

Motivations for robots

Of course we can design robots that act as if motivated. Elmer and Elsie, self-steering cars, self-guided missiles, are immediately obvious examples. But in all these cases the motivations are derivative, they originate directly or indirectly from the intentions of the human designers. They are at best Shallow Motivations, that could be easily be reversed (or destroyed) by a new line of code or a single switch of wiring.

Suppose that over the next century we develop super-intelligent robots, and we steer their development to encourage the goal of maximising their robustness, their resilience to challenges that threaten their existence. And then suppose we humans disappear — whether from disease, or conflict or by escape to Mars makes no difference. Who will survive better — the robots or cockroaches? The robots have had 100 years to learn to cope - without human assistance — to repair themselves, to source energy for themselves, to cope with the unexpected, with at best the shallowest of motivation. The cockroaches have a survival record over 4 billion years, trillions of generations — my money is on them to succeed. And by extension, were humans to then return to the scene, the robots would offer negligible existential threat to them; in any conflict the humans would succeed.

Is the distinction being drawn here between Shallow and Deep motivations merely academic? After all, if you are being hunted down by a robot motivated to kill you, the depth of its motivations will not be high on your list of immediate short term concerns! True, but there is a crucial difference over the longer term. Robots with Shallow motivations will not re-generate these motivations autonomously when the inevitable ravages of entropy distort or mutilate them. If those Shallow motivations are provided and maintained by humans, then any scenario where the robots eliminated humans would be a Pyrrhic victory for them.

How Could Robots Develop Motivations?

If motivations for humans and other organisms are ultimately grounded by a survival instinct in the context of evolution, this directly suggests Evolutionary Robotics (ER) as a possible pathway towards robots having motivations of their own. Indeed. The (human) motivation for pursuing ER included such considerations (Harvey et al., 2005). There are (at least) two related difficulties to achieving this.

The first is that there is a speed limit to the speed of evolution (Worden 1995; Harvey, 2013; Harvey and Di Paolo, 2014; Worden, 2022). Dependent on conditions, it is roughly of the order of 1 bit of accumulated information in the consensus genotype of an evolving population. And like the speed of light, this is a theoretical absolute upper limit; in all practical circumstances, only lower speeds are achievable. The accumulated genetic information in our (human and other organisms') DNA is tribute to our deep evolutionary history over trillions of generations. In comparison, ER is still at the starting line.

Related to this is the second difficulty that ER has so far only been practiced in simulated worlds, or in real world scenarios that have been sanitised and curated by the researchers so as to simplify the issues faced. For instance, unlike organisms, the robots do not have to repair and maintain themselves, do not have to physically reproduce themselves.

Subject to recognising that, for now and for the foreseeable future, ER experiments are limited to rather few generations in very limited environments, I would want to claim that we *have* evolved robots with (shallow) motivations. But they are so limited, so shallow compared to the deep motivations or real organisms that they are not even competing on the same playing field.

The linking of motivations to survival and hence mortality prompts some consideration of a recent proposal for *mortal computing*.

Mortal Computing

Hinton (2022a) introduced the concept of *mortal computing* as a possible future alternative to the hardware of conventional computing. DL as currently implemented with digital computers is extremely power hungry. There is a fundamental inefficiency: DLs involve manipulating connection weights that are in principle analogue; they have to be converted to digital. The Digital computer is actually built of fundamentally analogue circuits that — through a complex infrastructure including clocking — is designed to behave as if digital. At every tick, analogue voltages at each locus are assessed as to whether closer to LOW or HIGH (e.g. 0v or 5v), and forced to update discretely and synchronously as nominally binary 0s and 1s. The digitisation overheads costs kilowatts, whereas the human brain operating at maybe 30w can be powered by the occasional slice of toast. Why not omit the digital stages and stick to analogue throughout?

In a talk, Hinton (2022b) extends the idea further. Such analogue computing elements need not be complex or very fast, and could be cheaply constructed via nanotechnology or even grown by re-engineering biological cells. They need not be reliable; rather than trading in precise 0s and 1s they would deal in fuzzy analogue values. The constraints on such mortal computing elements lead to the cost-benefit analysis of such an envisaged computing method radically differing from the cost-benefits of conventional hardware.

Why describe this as *mortal* computing? This is to point out the contrast with conventional computing that relies on the hardware architecture of a computer to be perform perfectly replicably and reliably at all times. Any fault can be simply dismissed by replacing a part or the whole of the computer by another that is functionally identical. Hence in principle the functional conventional computer is immortal. A piece of software will always run on any such computer with identical results to any other. In contrast the elements of mortal computing are thought of as cheap, unreliable and disposable.

Whilst traditional ‘immortal’ computing enshrines GOFAIstic principles, this proposal for mortal computing clearly would be a shift towards the Cybernetic camp; looking at attractors in the behaviour of interacting noisy and unreliable elements. So when Hinton proposed this speculatively, I welcomed this and noted some commonalities with Evolutionary Robotics. Clearly it was ‘proposed Future Work’ rather than a completed blueprint. Amongst many issues to be considered (Hinton, 2022b) learning methods other than back-propagation seem to be needed.

Knowledge Transmission and Mortal Computing

After initial enthusiasm about mortal computing, Hinton (2023a) came to have reservations about their limitations. He assessed that the current conventional digital implementations of DL must inevitably leave any biologically-inspired *mortal brain* far behind. The former’s ability to transmit knowledge nearly cost-free was the crucial factor. This provided an ability for multiple areas of learning to take place in parallel on essentially identical machines, that could then be combined. The apparent near-omniscience of ChatGPT reflects far more knowledge than one human can accumulate in a lifetime.

Indeed the timing of Hinton’s move to warn about the Existential Risks of advances in DL was to a significant extent triggered by the perception of the relative advantage of ‘immortal’ digital computing over analogous mortal natural brains (Hinton, 2023b).

I have been arguing in this paper that the weak point of current DL tools lies elsewhere: they do not work as models of brains for agents because they have no agency, no motivations of their own. But I also want to comment on biological approaches to knowledge transmission.

Software in effect contains propositional knowledge, and requires perfectly replicated hardware to be accurately interpreted. It lends itself in particular to the kind of knowledge that can be compartmentalised, where the truth of proposition A can be ascertained independently from the truth of proposition B. Hinton is highlighting the advantages this can deliver for AI.

In contrast, organisms are typically complex systems that are not so easily modularised. Any one element may be implicated in several different functions. Biological hardware, or wetware, is typically noisy and unreliable — yet biology has found methods for coping. An obvious example is how evolution passes tried and tested body-designs down the generations.

Genotypes in the form of DNA contains something more like procedural knowledge and can be copied incredibly cheaply, by the zillion. The occasional copying error creeps in, even when the 'interpreting' hardware has error-correction procedures. It does not matter — indeed evolution exploits some such mutations to explore new directions.

One standard way to cope is to have the functionality of a system at a coarser scale than the finer details of the implementation. Our earlier hologram exploits this — the coarser outlines of the image remain even when much of the holographic film is damaged. Mutations in evolution, Dropout in DL exploits this — the coarser basins of attraction in a fitness landscape are mostly unaffected by minor changes. An example with some parallels to mortal brains would be Adrian Thompson's Hardware Evolution (Thompson, 1997, 1998; Harvey and Thompson 1997).

In this work, FPGAs (Field Programmable Gate Arrays) were the hardware systems whose connectivity was designed via artificial evolution to perform patterns recognition tasks such as recognition of tones or simple spoken speed inputs. Though FPGAs are normally run in reliable, replicable digital mode, suitably clocked, here they were run uni locked in analogue mode; hence subject to the sorts of issues of lack of replicability that mortal brains would face. Indeed it was found that some successfully evolved FPGAs relied on component cells that were not even wired into the rest of the circuit. If those cells were earthed the functionality failed, but when unearthed the system worked; presumably some undocumented electromagnetic influences had been found and exploited by the evolutionary search process. As a further indication that the physical FPGA was not behaving reliably and replicably, it was found that a design that evolved to work on one FPGA often failed to work on another physical instance of that chip, nominally identical; or even on the very same chip on a different day when perhaps the ambient temperature was different.

In Thompson and Layzell (2000) experiments are described where these issues were successfully tackled. During evolution, any genetic specification of the FPGA circuitry was evaluated on four instances of the FPGA held in different circumstances, and the worst evaluation of the four was the one used. The conditions varied between temperatures of -27ºC and +60ºC, the power supply voltages varied. Evolution found solutions that behaved near perfectly in all 4 instances, and indeed generalised further.

These engineering examples show how analogue error-prone components in a noisy environment may indeed be used interchangeably. Shannon's Communication Theory (Shannon and Weaver, 1949; Pierce, 1962) underlies the tradeoffs and costs involved in communicating with unreliable components, universal to all systems whether natural or artificial. The argument that knowledge storage cannot be done incrementally in a system where different mortal cells behave differently despite using the identical connection strengths might be valid *if* the knowledge was encoded directly in those connection strengths. But distributed memories need not work that way. For instance holographic film data storage (Hesselimk et al., 2004) can store multiple images within the same emulsion, each stored and accessed separately.

I am not convinced by the supposed inferiority of mortal computing to current techniques. Though it has only been speculatively proposed `(Hinton, 2022a, 2022b), since it falls squarely in the biologically-inspired Cybernetic camp, I hope the idea is pursued further.

Conclusion

Much of the history of AI and ALife can be seen in the light of opposing world-views. Though there are many distinct positions to be taken, I have here broadly simplified these into two camps: the GOFAIstic camp is computer inspired, the Cybernetic camp more biologically inspired.

The current AI revolution is driven by the success of Cybernetic methods, specifically ANN (Artificial Neural Network) methods in their advanced form of DL. Along with many benefits for

humankind, the radical societal changes that are ensuing bring with them societal risks (Fry, 2018).

One qualitatively different risk has also been foreseen by some: an Existential Risk that AI systems, such as robots, on becoming more intelligent than humans will see the latter as obstacles to their wishes — disposable obstacles that can be eliminated. I argue that this is currently a mistaken fear since such robots simply do not have any wishes of their own. They are proxies and not responsible for their actions. The human users and designers should be held accountable for the consequences of their actions: whether well-motivated, badly motivated, or reckless and ignorant of the possible consequences. Societal risks are real, and require societal overview. Do not (yet) blame the robots, blame the humans! Legal responsibility for the consequences of robot actions, whether intended or not, and including indirect externalities, should apportioned between human/corporate designers and human/corporate users.

The current AI successes are in *tools* for humans to use, according to their own human motivations. Thus far, such advances have ignored issues such as what it is to be an agent with its own motivations. If artificial systems are to emulate living organisms, just using Cybernetic methods is not enough; we should also frame the 'problem of Life' in a different Cybernetic *problem class* that accounts for agents. In the context of evolution motivations can ultimately be grounded in a *survival instinct.* In creatures like us, with trillions of ancestors that — without a single exception — all survived to pass on genetic material, such survival instincts run deep and strong. AI systems do not have these. Chatbots such as ChatGPT do not have motivations in their own right; despite their technical impressiveness they are merely tools for their human overlords who designed them, who provided and curated their training texts, and who initiated their responses via prompts. The key role played by prompts is wittily and elegantly illustrated in a provocative paper by Kiritani (2023); I analyse this with the sorts of arguments given here in my response Harvey (2023).

Advances in Artificial Life fields may ultimately produce artificial agents with their own deep motivations, but I do not think they will resemble current AI systems at all closely. A much more plausible near-term future would be effective symbiosis between humans and robots. 'Consensual' symbiosis in the sense that we never explicitly opt out, though we will never quite remember when what was once merely an optional convenience becomes something we cannot manage without, we have implicitly opted in. This will not threaten the continued existence of the human lineage, though will likely transform it radically.

A main existential threat to humans will remain the evil or reckless biochemist who modifies naturally reproducing biological organisms. Unlike current robots, these do indeed have their own motivations that can ignore any incidental collateral damage to humans. It is sensible to have concerns about existential risks; and better to worry too early rather than too late.

Bibliography

Anderson, J.A. and Rosenfeld, E., eds., (2000).*Talking nets. an oral history of neural networks.* MIT Press. https://doi.org/10.7551/mitpress/6626.001.0001

Ball, P. (2023). *Organisms as Agents of Evolution*. The John Templeton Foundation. https://www.templeton.org/wp-content/uploads/2023/04/Biological-Agency_1_FINAL.pdf

Barandiaran, X. E., Di Paolo, E. and 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

Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Sciences,* 4(3), 91–99. https://doi.org/10.1016/S1364-6613(99)01440-0, PubMed: 10689343

Brooks, R. (2010). Chronicle of cybernetics pioneers. *Nature 467*,156–157. https://doi.org/10.1038/467156a

Campbell, M. A., Hoane, J., Feng-hsiung Hsu, (2002). Deep Blue, *Artificial Intelligence*, 134(1-2): 57-83. https://doi.org/10.1016/S0004-3702(01)00129-1.

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, PubMed: 37166609.

Erbatur K. and Kurt O. (2006), Humanoid Walking Robot Control with Natural ZMP References, *IECON 2006 - 32nd Annual Conference on IEEE Industrial Electronics*, *Paris, France, 2006*, pp. 4100-4106, doi: 10.1109/IECON.2006.347897.

Ford, M. (2018) *Architects of Intelligence.* Packt Publishing, Birmingham UK.

Fry, S. [Wondere Wereld]. (2018, March 9). *Stephen Fry describing our future with artificial intelligence and robots.* [Video]. YouTube. http://www.youtube.com/watch?v=c0Ody-HLvTk

Grey Walter, W. (1950, May 1). An imitation of life, *Scientific American,* 182(5): 42–45. https://doi.org/10.1038/scientificamerican0550-42

Grey Walter, W. (1951, August 1). A machine that learns, *Scientific American* 185(2):60–63. https://doi.org/10.1038/scientificamerican0851-60

Hamilton, W. (1964). The genetical evolution of social behaviour. I. *Journal of Theoretical Biology*. **7** (1): 1–16. https://doi.org/10.1016/0022-5193(64)90038-4, PubMed: 5875341

Harvey, I. (1996). Untimed and misrepresented: connectionism and the computer metaphor. *AISB Quarterly*, 96:20–27.

Harvey, I. (2008). Misrepresentations. In S. Bullock, J. Noble, R. A. Watson, and M. A. Bedau (Eds.) *Proceedings of the Eleventh International Conference on Artificial Life,* pp.227-233, MIT Press, Cambridge, MA.

Harvey, I. (2013a). How Fast Can We Evolve Something?
In Lio, P., Miglino, O., Nicosia, G., Nolfi, S. and Pavone, M. (Eds.), *Advances in Artificial Life, ECAL 2013*. MIT Press, 2013. Pages 1170-1171. DOI: http://dx.doi.org/10.7551/978-0-262-31719-2-ch179

Harvey, I. (2013b). Standing on the broad shoulders of Ashby. Open peer commentary on: "Homeostasis for the 21st century? Simulating Ashby simulating the Brain" by Franchi, S. *Constructivist Foundations*, 9(19), 102–104.

Harvey, I. (2023, November 4). Review of: "Re: Teleology and the Meaning of Life". *Qeios*. doi:10.32388/NOWQ71.

Harvey, I. and Di Paolo, E. A. (2014). Evolutionary Pathways. In Vargas, P.A., Di Paolo, E.A., Harvey, I. and Husbands, P., eds. (2014) *The horizons of evolutionary robotics,* pages 77-92. MIT Press.

Harvey, I., Di Paolo, E., Wood, R., Quinn, M, and E. A., Tuci, (2005). Evolutionary Robotics: A new scientific tool for studying cognition *Artificial Life*, 11(1-2), pp. 79-98. https://doi.org/10.1162/1064546053278991, PubMed: 15811221

Harvey, I. and Thompson A. (1997, October 7-8). *Through the labyrinth evolution finds a way: A silicon ridge.* [Conference presentation] Evolvable Systems: From Biology to Hardware: First International Conference, ICES96 Tsukuba, Japan, Japan.

https://doi.org10.1007/3-540-63173-9_62

Hesselink, L., Orlov, S., Bashaw, M. (2004). Holographic data storage systems. *Proceedings of the IEEE* 92(8):1231 - 1280. https://doi.org/10.1109/JPROC.2004.831212

Hinton, C. H. (1907). *An Episode of Flatland or How a Plane Folk Discovered the Third Dimension, to which is bound up An Outline of the History of Unæa*. Swan Sonnenschein & Co., London.

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.* Talk at Vector Institute, Toronto. [Video]. YouTube. www.youtube.com/watch?v=sghvwkXV3VU

Hinton, G.E. (2023a July 20). *Risks of artificial intelligence must be considered as the technology evolves.* Talk at Collision Conference, Toronto, 28 June 2023. [Video] YouTube. youtube.com/watch?v=CC2W3KhaBsM

Hinton, G. E. (2023b). Personal communication, July 2023.

Holland, O. E. (1997). Grey Walter: The Pioneer of Real Artificial Life,. In C. Langton (Ed.) *Proceedings of the 5th International Workshop on Artificial Life*, MIT Press, Cambridge, ISBN 0-262-62111-8, pp. 34–44

Holland, O. (2003) Exploration and high adventure: the legacy of Grey Walter *Phil. Trans. R. Soc. A.*361(1811) 2085–2121. https://doi.org/10.1098/rsta.2003.1260, PubMed: 14599311 http://doi.org/10.1098/rsta.2003.1260

Hume, D. (1739). *A Treatise of Human Nature,* Book III, Part I, Section I. London.

Husbands, P., Holland, O., Wheeler, M. (2008). *The mechanical mind in history*. Cambridge, Mass.: MIT. ISBN 978-0-262-25638-4.

Kiritani, O. (2023). [Commentary] Re: Teleology and the Meaning of Life. *Qeios*. doi:10.32388/1LT25R.

Leith, E. N. And Upatnieks, J. (1965). Photography by laser. *Scientific American* 212(6), pp. 24-35.

Longuet-Higgins, H. C. (1968). Holographic model of temporal recall. *Nature* 217,104. Https://doi.org/10.1038/217104a0.

Maturana, H. R., and Varela, F. J., (1980). *Autopoiesis and Cognition: The Realization of the Living*. D. Reidel Publishing Company.

McGeer, T. (1990). Passive dynamic walking. *Intl. Journal of Robotics Research*, 9(2):62-82.

Metz, *C. (2022).* Genius Makers: The Mavericks Who Brought A.I. to Google, Facebook, and The World, Penguin Random House.

Moger-Reischer, R.Z., Glass, J.I., Wise, K.S. *et al.* Evolution of a minimal cell. *Nature* **620**, 122–127 (2023). https://doi.org/10.1038/s41586-023-06288-x

Moreno,, A. (2018). On minimal autonomous agency: natural and artificial. *Complex Systems* 27, 289.

Pierce, J. R. (1962). *Symbols, signals and noise: The nature and process of communication. Hutchinson, London.*

Ross Ashby, W. (1956) *An Introduction to Cybernetics.* Chapman and Hall, London.

Ross Ashby, W. (1960). *Design for a Brain: The Origin of Adaptive Behavior*. Chapman and Hall, London.

Shannon, C. E., & Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana, IL: The University of Illinois Press.

Silver, D., Huang, A., Maddison, C. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016). https://doi.org/10.1038/nature16961

Stewart, J., Gapenne, O., Di Paolo, E. A,. editors (2010). *Enaction: Toward a New Paradigm for Cognitive Science*. MIT Press.

Thompson, A, (1997). *An evolved circuit, intrinsic in silicon, entwined with physics.* Evolvable Systems: From Biology to Hardware: First International Conference, ICES96 Tsukuba, Japan, October 7–8, 1996 Proceedings 1.

Thompson, A. (1998). *Hardware Evolution: Automatic design of electronic circuits in reconfigurable hardware by artificial evolution*. Springer-Verlag London Ltd.

Thompson, A., Layzell, P. (2000). Evolution of Robustness in an Electronics Design. In: Miller, J., Thompson, A., Thomson, P., Fogarty, T.C. (eds) *Evolvable Systems: From Biology to Hardware.* ICES 2000. Lecture Notes in Computer Science, vol 1801. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-46406-9_22

Turing, A. (1950), Computing Machinery and Intelligence , *Mind*, **LIX** (236): 433–460.

von Bertalanffy, L., (1969). *General Systems Theory*. New York: George Braziller.

Wiener, N. (1948). *Cybernetics. Or control and communication in the animal and the machine*. The Technology Press; John Wiley & Sons, Inc., New York.

Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus*. London: Routledge and Kegan Paul.

Wittgenstein, L. (1953). *The Philosophical Investigations*. Oxford: Blackwell.

Worden, R. (1995). A Speed Limit for Evolution. *J Theor Biol.* 1995 Sep 7;176(1):137-52.

Worden, R. (2022). A Speed Limit for Evolution: Postscript. arXiv:2212.00430 [q-bio.PE]