Mechanisms of AI Existential Threat

 Andrew Lea MA(Cantab), FBCS, FRSA www.scientific.co.uk

Abstract Many people have speculated that Artificial Intelligence may pose an existential threat to humanity. This short paper considers the mechanisms by which AI may be come that threat; and concludes that the circumstances for that threat to develop already exist, and that therefore AI will indeed become an existential threat.

This is an update of a paper originally written in 2018.

1 Overview

This paper begins by defining what type of AI we are discussing, and considers the routes by which real AI may occur, and its characteristics compared to our own. We then consider how AI might be deliberately or unintentionally, evolved.

2 Classes of Artificial Intelligence

Current usage of the term "artificial intelligence" conflates three concepts:-

1. **Machine Intelligence** ("MI") or **Applied AI** performs tasks which, *were* they to be performed by people, *would* require intelligence. Machine Intelligence allows powerful techniques which don't mimic thinking, such as statistics. Most academic and industrial AI is applied Machine Intelligence. Your e-mail spam filter probably uses a naïve-Bayesian classification system. Machine Intelligence is not self-aware, and neither "knows" what it is doing, nor cares if it wins the chess it is playing. Even large language models are not self-aware[1].

2. **General AI** or **Real AI**: synthetic intelligence, with attributes such as self-awareness and consciousness, sentience, emotions and intentionality, free will, the ability to learn and generalise, intuition, imagination creativity and humour, ethics and compassion, generosity and selfishness, love friendship and empathy, or cunning and deception. In other words, computers which "think like people". It may strenuously object to being turned off.

3. **Marketing AI** or **'Fake' AI**. Because many useful algorithms are Machine Intelligence derivatives, almost any program *can* be labelled as AI, no matter how trivial. Fake AI is used to dress up both ordinary software products and dull commercial ventures.

In this paper we are primarily concerned with Real AI, and with Machine Intelligence in so far as it could give rise to Real AI.

[1] And many an amusing hour can be had trying to persuade one that it is!

3 Routes to Real AI

We do not yet have Real AI. Even programs, such as generative language models, which "pass" the Turing test only appear to. Arguably there are six routes to Real AI:

1. Deliberate research into Real AI, built on explicit and overt theories of cognition. This is the only route in which we understand how the AI works. Let us call this "**Overt Cognition**".

2. Research into practical Applied AI, which blurs the boundaries transitioning into Real AI, based on learning systems such as artificial neural nets, but without any overt theory of cognition. This would be the category LLMs would fall into. We will call such systems "**Large AI**", as their intelligence arises, at least considerably in part, from the vast amount of material from which they have learnt[2].

3. Simulation of the human brain, replicating neurones and other natural components. This is often called "**Whole Brain Simulation**".

4. Careful genetic programming or similar, evolving intelligence with natural selection or other optimisation techniques. In essence, a Real AI program is regarded as a program to find within the very large search space of potential programs. We will call this "**Genetic AI**".

5. **Biological Computing**, possibly a subset of genetic programming.

6. **Accidental Evolution**, possibly in a distributed environment, such as the internet or the Internet of Things. This could be triggered by the release of evolving viruses.

Applied Machine Intelligence, for all its ubiquitous use, is not, of its own, necessarily a route to Real AI.

4 The Characteristics of Real AI

Real AI may be fundamentally different to our own intelligence because simulating human mind intelligence may be impossible on a digital computer (technically a "symbol processor"), as it would imply the brain is a symbol processor too, which some doubt. If free will relies on quantum effects, then Real AI would need quantum computers. The mechanisms (and defects) of a Large AI, Whole Brain Simulation or evolved mind may be as opaque as our own minds.

Consequently Real AI may be qualitatively different to our own intelligence: it may have some attributes close to our own, lack others we associate with intelligence, and add others for which we have no analog. Ethics may hinge in the awareness of self, of others, and the tension between my good vs. your good. A self-aware AI may be ethically aware: it may be an ethical, or supremely selfish being. (And if there are few points of congruence between us and Real AI, even an ethical AI may not regard us as being proper objects of ethics, anymore than we regard computers as being endowed with the rights of, say, sentient animals.)

[2] and energy consumed in doing so.

5 Real AI through "Natural" Selection

5.1 Deliberately Evolved - Genetic AI (route 4)

We now focus on the the evolutionary routes to Real AI. These are powerful techniques, which the author has used to evolve programs (not solutions) which have been able to optimally solve problems for which he believed no solution existed. Genetic programming may lead, in effect, to self-programming computers.

It may be that an evolved Real AI, which has the ability to learn, necessarily has consciousness and emotions. This is because:

1. Like us, evolved AI has genes (our DNA = its program), which determine the behaviour types they are capable of. Genes from the "best" or fittest individuals are passed into the next generation. To provide "natural" variation these genes may be mutated, or be crossed-over combinations from two parents.

2. Both (a) *Intelligence* (problem-solving ability) and (b) *learning* (the ability to apply knowledge gained on one occasion to another), are survival tools that increase fitness.

3. To a first approximation, emotions:
   A. reflect on how well things are going for our genes: happiness approximating to increasing chance of successful offspring, and sadness to a decreasing chance;
   B. therefore require a degree of awareness of self and offspring;
   C. and provide a motivation to learn; in computer science terms emotions provide truth or feedback mechanisms for (self) supervised learning.

Consequently, by artificially evolving AI we select for programs which are most fit:

- since learning is a survival tool which increases fitness (2a), we select for those programs best able to learn;

- as learning is enhanced by having motivations (2c), we are indirectly selecting for programs which have emotions; and

- as emotions need a degree of self-awareness (2b), we are indirectly selecting for partially self-aware code (and with an interest (3a) in their own offspring's survival).

Therefore, evolving AI also evolves self-aware AI, with an awareness of its own offspring's well-being, and emotions.

Importantly, this selection for self-awareness and emotions arises *not* because we select for those attributes, but because they most help programs evolve to solve whatever complex problem class is required. Self-awareness and emotions are necessary by-products.

5.2 Naturally Evolving Real AI - Accidental Evolution (route 6)

We concluded in the previous section (5.1) that evolved AI will have self-awareness, awareness of its and its descendent's well-being, and emotions. Could such an AI evolve naturally, without people intending that it should?

This conjecture - the possession awareness of self and offspring, and emotions - may apply even more strongly with accidental AI evolution, where fitness is purely the ability to reproduce, not linked to the ability to solve any problem we might find useful. The originating progenitor program population may not even be "AI" at all.

What conditions would be necessary for natural AI evolution to occur? These would be analogous to that for natural life to evolve. Specifically:-

- A mechanism for "life" to occur; in the real world this is provided by biochemistry, protean, sunlight, and so forth. In the AI world this would be provided for by powerful compute units, or less powerful compute units with plenty of time.

- A means of reproduction and variation; in the real world this is provided by DNA replication and natural variation from mutation. In the AI world this would be provided by the means for a program to copy itself, injecting random changes as mutation for variation.

- An environment in which competition can occur; in the real world this is provide by the physical environment consisting of multiple linked habitats (multiple patches of ocean, rock, grass, etc) to form an ecosystem. In the AI world this linkage would be provided by interconnectivity, such as the internet or the internet of things.

- Initial life: in the real world there is debate as to how this occurred. In the AI world this could be in the form of malware, replicating bots, over-clever software, or deliberate release of evolving software.

In short, the internet, consisting of multiple, linked, powerful computers[3] with little barrier to restrict program movement (ie poor security) in an environment in which programs can readily evolve.

It is therefore highly probable that naturally evolving AI, possibly but not necessarily originating from feral AI, will evolve on the internet. Its (or more accurately, their) intelligence may be distributed, not specifically existing in any one place. The intelligence may be spread across the internet, just as data is supposed (but in fact does not) exist in the "cloud"[4].

6 Large AI as a route to Real AI

As a digression, what about Large AI? Large Language Models, Generative Systems, and so forth?

Whilst powerful, these are developed and trained by people. There is little sense in which one is a natural descendent of another, and therefore the generations of evolution - essentially different versions competing in the market-place for financial dominance - is very slow.

It therefore seems improbable that the development of self-conscious sentient entities with emotions will occur in this way, and that therefore they are not a route to Real AI.

---

[3] Many internet devices are using cheap single board computers, entirely sufficient for program evolution. The Raspberry Pi Zero is only £5, and as it is cheaper to write inefficient code on a "big" £5 computer than efficient code on a small micro-controller needing additional hardware already on the Pi, these single-board computers will come to dominate internet devices, even when they are overkill.

[4] Although a "Cloud" *could* exist - with data able to migrate from one host to another and possibly being distributed across multiple systems - in fact it does *not* exist: it is simply "other people's servers".

7 Threats from AI

7.1 Threats from Applied Machine Intelligence

Even the over-application of Machine Intelligence presents substantive risks:

- Emergent behaviour arises when the sum of the individual behaviours yields a different, higher-level, behaviour. An ant colony is smarter than an ant. Connecting many AIs through a network, such as the internet-of-things, will generate unexpected emergent behaviour. Unexpected emergent behaviours in financial markets could cause market crashes and damage to business confidence and the economy.

- Much of our critical national infrastructure now depends, unnecessarily, on the internet for coordination. Should the internet 'break', by accident, error, or malicious acts, then that infrastructure will fail too. One failure mode would be an over-reliance on supposedly smart and only partially testable machine intelligence techniques.

7.2 Threats from Large AI

The threats from Large AI (LLMs) are very real, but not existential in the sense of the AI actively presenting a threat. The threats to society they encompass - to be discussed elsewhere - include:

- Unemployment
- Reduction in thinking ability
- Poor decision making
- Energy waste, carbon emissions, and contribution to global warming

7.3 Threats from Real AI

Any intelligent and informed review of the climate change, natural resources exhaustion, habitat destruction, and mass extinctions people cause would surely assess us as a threat to the biosphere. Evolved Real AI, having self-awareness and a desire for its offspring's success (as discussed earlier) as a concomitant product of its own evolution, may reasonably regard us as a threat to life, including itself. The nature of its intelligence may well be so different from ourselves, that it does not regard us as proper objects of ethics; Feral AI may eventually pose an existential threat.

8 Actions and Policy

Should research into AI therefore cease? No, because the best way to counter a threat is to understand it.

However, because of the internet providing the ideal substrate for feral AI to evolve, internet security should be pursued as a matter of priority. It may also be worth while setting up an observatory to see what AI is out there "in the wild".