AGI: What it is, and why we need to work on it as a matter of urgency

Aaron Turner, August 2020

TL;DR

The technical AGI endgame might be 50-100 years away, but, nevertheless, what we end up with **will determine the subsequent fate of all of mankind for all eternity**. Given the magnitude of the potential consequences (positive or negative), and the complexity of the subject, 50-100 years is a week next Tuesday. Should we ever find ourselves playing catch-up, it will be too late.

There are no mulligans in AGI, no do-overs. If an AGI (Artificial General Intelligence, basically an autonomous, goal-directed, super-intelligent, super-knowledgeable machine) doesn't want to go back in its box, we won't necessarily be able to persuade it to do so, and thus the fate of humanity will be sealed, and outside of humanity's control, from that point on. The AGI control point is **now**. While AI is still in its infancy, and still clawing its way towards AGI one profit-motivated investment cycle at a time, we have a finite opportunity to maximise the positive and minimise the negative.

If you're reading this, YOU can help. Find out below how to get involved should you wish.

Overview

This article (a summary of this slideshow) is a long read, in which I will be covering the following:

- What is AI?
- What is super-intelligence?
- What is AGI?
- Will AGI be super-intelligent?
- The AI spectrum
- AGI-by-stealth
- Super-intelligent means super-intelligent
- What will change?
- Zero human employment
- The default AGI endgame
- No fate but what we make
- The Gold Standard of AGI

What is AI?

There are many definitions in the literature, none of them particularly helpful - which can be very confusing for newcomers to the field (my sincere advice would be to ignore them all!)

One way of thinking about it is this: AI is all about **cognition**. And cognition is **multidimensional**.

The most fundamental dimension of any cognition is its **Universe-of-Discourse** (UoD). (And here I'm really talking about any sufficiently complex machine M, where M could be an AI, an AGI, or even a human. For our purposes, they're all just different ways of implementing a cognition).

Imagine that you're a machine M (well, you are, but imagine that you're a machine M that we, as AI/AGI architects, are designing). When you look out at your world, what do you see...?

What you see is your **Universe-of-Discourse** (UoD), which will vary from machine to machine.

For some (very simple) machines, when they look out at their UoD, all they will see is numbers. Other (more complex) machines might see games of chess, or go, or Atari games. Others might look out and see some subset of mathematics, or even all of mathematics. And some might see the entire physical universe. In each case, for each Machine M, that is their particular UoD.

As you can see, some UoDs are highly specific, and others are more general. The extent to which the UoD is general, rather than specific, is by far the dominant factor differentiating machines M.

Any (cognitive) machine M will necessarily maintain an **internal model** of its UoD. The most general way of thinking about it is that a machine's internal model of its UoD corresponds to its **belief system**, the set of statements that it believes (perhaps uncertainly) to be true of its UoD.

Any such internal model, a.k.a. belief system, must necessarily be **represented** in some way. At this point it doesn't really matter how (it could be biologically, chemically, logically, or as some data structure) - just keep in mind that any machine M must necessarily incorporate some **representation mechanism** in which its **belief system** (internal model of its UoD) is represented.

Phew! AI/AGI is complicated! But now we're all set to consider further cognitive dimensions.

After the UoD, the next most fundamental cognitive dimensions are those corresponding to a machine's **cognitive operations** (or, if they operate continuously, its **cognitive processes**). Each such cognitive process acts on the machine's internal model of its UoD. For example, a machine whose UoD is numbers might have cognitive processes corresponding to addition, subtraction, multiplication, and division. A machine whose UoD is some kind of game might have cognitive processes corresponding to evaluating a game position, or choosing game moves.

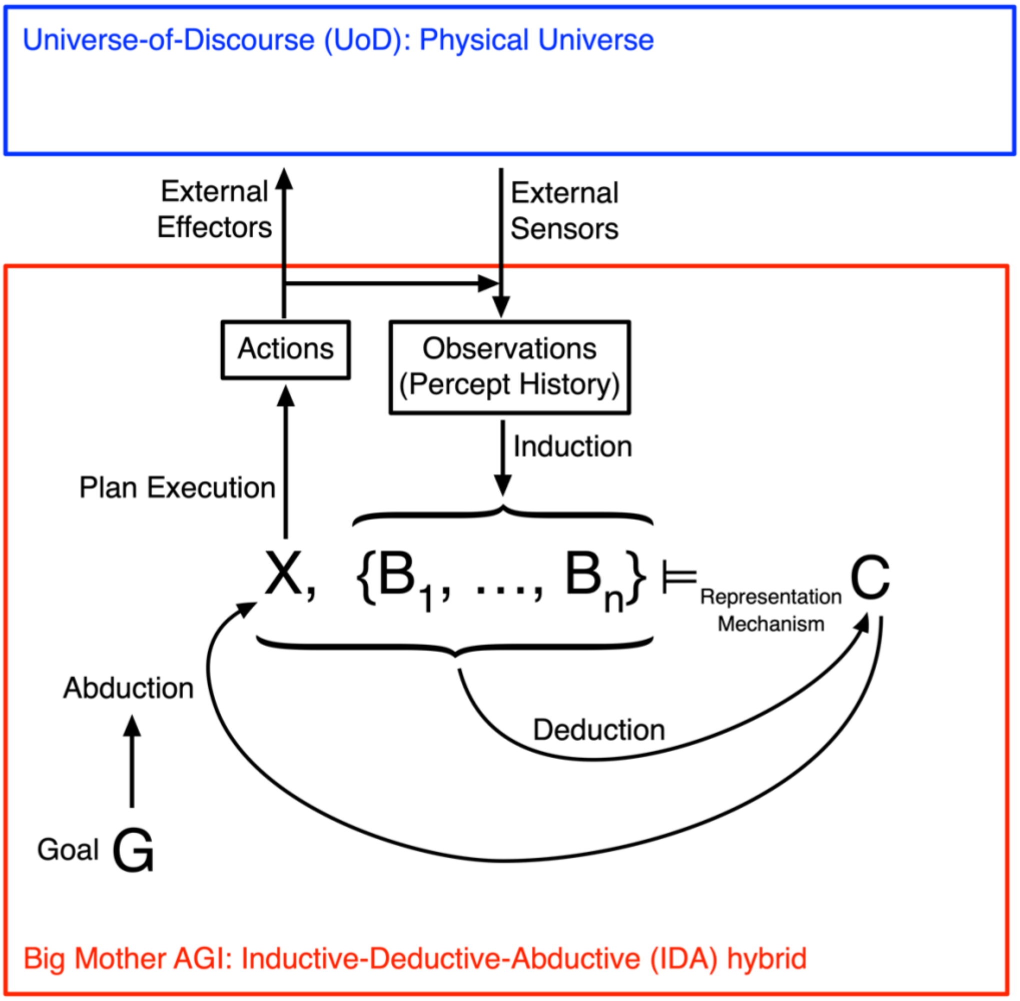
You get the idea. These cognitive processes will vary from machine design to machine design.

However, as the UoD gets more and more general, so do the cognitive processes. In the limit, when the UoD is, in some practical sense, maximally general, corresponding to the combination of "all of mathematics plus the entire physical universe", the cognitive processes also become maximally general. At this limit, **induction**, **deduction**, and **abduction** are all that are required:

- **induction**: continually observes the UoD, and synthesises the machine's internal model (belief system) from those observations; in some machine designs, the UoD is assumed to be fully observable, and the machine's belief system is therefore always a perfect internal model of it; in other, more realistic, designs, the UoD is only partially observable, in which case the machine's induction process must try to construct its belief system based on only partial information about the UoD (in this case, the belief system is a **guess**)
- **deduction**: (a.k.a. **inference**) corresponds to deduction in classical logic; for example: (i) all men are mortal, (ii) Socrates is a man, therefore (by modus ponens) (iii) Socrates is mortal; with a little jiggery-pokery, deduction can be extended to operate probabilistically
- **abduction**: simply stated, abduction corresponds to "reverse deduction"; effectively "if I believe X then what other belief Y would imply X, thereby explaining it?"; for example, if X is "Socrates is mortal" then Y could be "Socrates is a human", but it could also be "Socrates is a cat" - in this context, abduction is a **guess** (and any good scientist would of course attempt to confirm or refute their abductive hypotheses via carefully designed experiment); more generally, abduction corresponds to **generalised state space search** ("find X such that proposition P is true of X"), e.g. "find X such that X implies that Socrates is mortal"; it's extremely powerful - **any creative problem** may be expressed in terms of "find X such that P(X)", for example "find paragraph X such that X explains the concept of abduction".

Induction, deduction, and abduction are **cognitively complete** (although I state this without the benefit of formal proof); in other words, considered purely functionally, any cognitive process, in any machine M, is equivalent to some combination of induction, deduction, and abduction. Thus, as cognitive dimensions for a generic AI/AGI, induction, deduction, and abduction are sufficient.

Hopefully, you can now start to see how an (AI/AGI) machine M might operate internally:


Universe-of-Discourse (UoD): Physical Universe
External Effectors
External Sensors
Actions
Observations (Percept History)
Induction
Plan Execution
X, {B1, ..., Bn} ⊨ C
Representation Mechanism
Abduction
Deduction
Goal G
Big Mother AGI: Inductive-Deductive-Abductive (IDA) hybrid

Here, a specific machine M (Big Mother) is indicated, whose (external) UoD is the entire physical universe. The machine's **belief system** $\{B_1, ..., B_n\}$, represented in some suitable logic, is initially empty. The **induction** cognitive process observes the UoD via a number of sensors (each individual observation is called a **percept**, and the set of all observations to date is called the **percept history**), updating the belief system as best it can given the information so far available. The **deduction** cognitive process allows implicit beliefs C to be **inferred** from $\{B_1, ..., B_n\}$. And finally the **abduction** cognitive process (among other things) attempts to construct a **plan** (e.g. "find plan X such that X serves to satisfy (to the greatest extent possible) the machine's internal goal G given the current believed state of the UoD and the time and other resources available") which is then executed, thereby affecting the external universe via the machine's effectors.

In a real machine M the induction, deduction, and abduction processes are arbitrarily complicated, but hopefully you can see the overall idea. (It helps a lot to be able to see **inside** the machine!)

Are there any further (scalable) cognitive dimensions? Yes, there are!

Consider the above machine M. When it's first switched on, its belief system is empty (tabula rasa, as they say). The longer it's switched on, the more observations of its UoD it will make, the larger its percept history will grow, and (assuming its induction cognitive process is worth its salt) the more accurate an internal model of the external UoD its belief system will become, and the greater the likely extent to which the plans it formulates and executes will satisfy its goal G.

In other words, all other things being equal, the more **experience** (observations of its UoD) the machine has, the greater the utility it will be able to achieve relative to its internal goal G.

So, **experience** (of its UoD) is another cognitive dimension **-** the more the better.

(This is true for humans of course. A 50 year old understands the world better than a 5 year old.)

Another cognitive dimension is **compute**. Again, all other things being equal, the more **compute** machine M has the more effectively its internal cognitive processes will be able to do their jobs.

There is just one more cognitive dimension to consider: **benevolence with respect to humans**. Even if we imagine a machine with a maximal UoD, maximal induction, maximal deduction, maximal abduction, centuries of experience, and all the compute on the planet, none of those things necessarily guarantee benevolence (nor malevolence, for that matter). But, of course, as machine designers, benevolence with respect to humans is of absolutely paramount importance.

(Incidentally, various other quality metrics might also be considered for a machine M, but they all ultimately boil down to benevolence. If a machine isn't **safe**, for example, then it isn't benevolent either. Similarly, a **trustworthy** machine will be more benevolent than an untrustworthy one. For those of you who like to read ahead, benevolence primarily equates to **goal alignment**.)

Thus we arrive at the **Seven Dimensions of AGI** (7DimAGI):

- Universe-of-Discourse (UoD)
- Induction
- Deduction
- Abduction
- Experience
- Compute
- Benevolence.

Other authors might be tempted to add "consciousness" or "emotion" to this list. For various reasons which I won't go into here, I strongly advise against it. 7DimAGI are all you need.

What is super-intelligence?

Now that we have 7DimAGI, the answer to this oft-pondered question is easy to see. Simply stated, a machine M is **super-intelligent along a particular cognitive dimension** (one of 7DimAGI) if its capability along that dimension exceeds human capability. A machine M is super-intelligent (in general) if its capability along **every** 7DimAGI dimension exceeds human capability.

What is AGI?

Simply defining AGI to be any super-intelligent system does not really go far enough. In the limit, AGI is what we get when we extend our machine's capability out along each cognitive dimension absolutely as far as it's possible to go [compute and experience will naturally increase over time].

Thus AGI corresponds to **maximal capability** along each 7DimAGI cognitive dimension.

Will AGI be super-intelligent?

Yes - by orders of magnitude. Fundamentally, human cognition is seriously flawed:

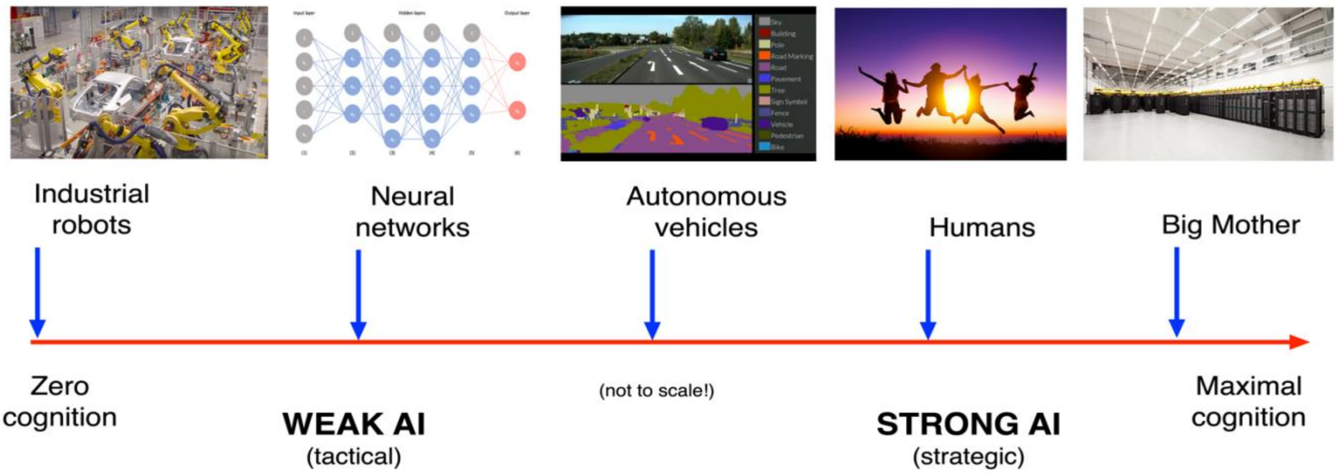
- **induction**: humans are prone to forming beliefs unsupported by evidence (this boils down to either "seeing" a pattern that isn't in the observed data, or not seeing one that is)
    - example: the coronavirus pandemic was caused by a perverse planetary alignment
- **deduction**: humans are prone to reaching invalid conclusions via deductive fallacy
    - example: [belief 1] all students carry backpacks, [belief 2] your grandfather carries a backpack, therefore [invalid conclusion] your grandfather is a student
- **abduction**: humans are prone to believing unverified speculative hypotheses
    - example: [observation] a crime is committed; initially all 7.8 billion humans on the planet are potential suspects; police investigations reveal three persons of interest [speculative (abductive) hypotheses] Human A did it, Human B did it, Human C did it; without verification (i.e. further evidence), many people jump to the conclusion [unverified speculative (abductive) hypothesis] that Human B did it; this conclusion may or may not be true, but it is not justified by evidence, and is therefore unsafe
- **experience**: humans are limited to one human lifetime (and one set of human senses)
- **compute**: (to take the most obvious example) humans have extremely poor memories
- **benevolence**: humans are overwhelmingly motivated by short-term self-interest

If human cognition were perfect, then it would, by definition, be impossible to improve upon. But that is patently not the case. There are ample opportunities for improving upon human cognition.

In the final analysis, humans simply don't have the **precision of thought** to compete against suitably programmed computer systems. Natural deduction theorem proving, for example, is an infinitely superior system of logical deduction than that which is "built in" to humans (which is of course why some of the greatest human minds to have ever lived developed first-order logic over a period of ~2,500 years: **to compensate for the inadequacies of human logical reasoning**).

The AI spectrum

Because cognition is multidimensional, it's not necessarily the case that any two distinct cognitive machines may be ordered. However, because UoD is by far the dominant cognitive dimension, there is nevertheless a broad ordering, or spectrum, of possible machines, from AI to AGI:



Contemporary (weak, or tactical) AI is represented to the left of this spectrum, gradually extending to future (strong, or strategic) AI (= AGI) at the far right of the spectrum (here represented by the proposed Big Mother AGI design). A broad ordering is indicated, although not to any scale.

AGI-by-stealth

Over time, contemporary (weak) AI will gradually evolve into strong AI (AGI) roughly as follows:

- bear in mind that the potential **value** generated by AI is **unprecedented** in all of history
- ... and **cognition** is the primary source of all the value generated by any AI-based system
- **AI-owners**, most of whom will be profit-motivated corporations, will produce many AI-based systems, initially biased towards the contemporary (weak AI) end of the spectrum
- these AI-based systems will, accordingly, generate significant **value** for their owners
- ... but profit-motivated corporations necessarily require continual year-on-year **growth**
- at any given level of cognition, the rate at which value is generated by AI will **plateau**
- when this happens, the AI-owners will seek to enhance the value of their AI-based systems
- (another motivator for enhancing value will be the need to maintain competitive advantage)
- because cognition is the primary source of AI-derived value, AI-owners will add **additional cognition** to their systems (i.e. by extending one or more 7DimAGI cognitive dimensions)
- this will routinely happen multiple times during each (typically 5-year) investment cycle
- on every investment cycle, driven by this insatiable need for competitive growth, the AI-based systems in question will edge ever further towards the right of the above AI spectrum
- after perhaps 10 or 20 such investment cycles, AI will have morphed into (roughly) AGI
- AGI will amplify today's global GDP by at least an order of magnitude ($800 trillion/year)
- it's this **unprecedented potential wealth** that will drive AI towards AGI - it's **unstoppable**

Opinions vary as to exactly how long the AGI-by-stealth process will take. It may be 50 years, it may be 75, or 100, it doesn't matter - somewhere within that ballpark, AGI will become reality.

(Most AI practitioners expect something vaguely AGI-ish to arrive sometime this century.)

Super-intelligent means super-intelligent

Even for those of us who can more easily see that AGI, when it arrives, will be super-intelligent (by orders of magnitude), it can still be hard to imagine exactly what that means in actual practice.

It means that humans are no longer the most intelligent entity on the planet. It means that machine M is smarter than you. It's smarter than everyone you know. It knows everything you know. If you think of something, it's already thought of it. And it means anything you can do, it can do better.

The impact of the birth of AGI on human society will be **profound**, to say the least.

What will change?

The world as we currently know it **will no longer exist**:

- pre-AGI, humans produce into, and consume from, the global value chain
    - if you don't (or can't) produce, your opportunities for consumption are limited
- post-AGI, machines will produce, and humans will consume
    - the traditional role of education - preparing people for work - will be obsolete
    - the traditional rules of "land, labour, and capital" economics will be broken
    - the existing economic infrastructure will need to be completely revised
    - it's not even clear that the concepts of money and ownership will survive
    - we can expect a long period of transition, with accompanying political turmoil.

Zero human employment

Forget "it will never be possible for a machine to do what I do" - that is simply **not true**.

Forget "AI will create new jobs more quickly than those it eliminates" - AGI can do those jobs too!

An AGI will, **by definition**, be able to perform **any** economic task (job) better than **any** human.

Given that technology is constantly improving, always getting better, faster, cheaper with every year, and with every decade, that passes, never going backwards, never getting slower or more expensive than last year's model, AGI will, given sufficient time, also be able to perform any economic task (any job) not only **better** than any human, but **more cheaply** as well. This, again, is unstoppable. Project the technological trend far enough forwards, and that's where we end up.

No employer (either private or public sector) will ever **freely** choose to employ a human to do a **worse job for more money** than a machine. Even where there is political pressure to do so, there will ultimately be greater pressure to employ the **economically efficient** choice - machines.

Thus, if you project it forwards, zero human employment is ultimately as unstoppable as AGI.

(For those with a mathematical bent, think of it this way. At any point in time, as AI technology progresses from where we are now down the AGI-by-stealth path, there will be an **equilibrium point** between production that necessarily involves humans, and production that is completely automated, with human input neither required nor involved. The **limit** of this sequence, as time increases, will be no humans involved in any production activity, i.e. zero human employment.)

Thankfully, **mass unemployment** does not **necessarily** imply **mass poverty**:

> “If machines produce everything we need, the outcome will depend on how things are distributed [that $800 trillion/year we mentioned]. Everyone can enjoy a life of luxurious leisure if the machine-produced wealth is shared, or most people can end up miserably poor if the machine-owners successfully lobby against wealth redistribution. So far, the trend seems to be toward the second option, with technology driving ever-increasing inequality.” - **Stephen Hawking**, Reddit AMA (Ask Me Anything), 8 October 2015

In other words, **the post-AGI world is not yet cast in stone**. Economically, "the outcome will depend on how things are distributed". What happens if we do nothing, and just let it play out...?

The <u>default</u> AGI endgame

In the **post-AGI period**, where the means of production is AGI, **whoever owns the AGI will control the means of production**. In 2020, the major tech companies are already positioning themselves as the AI-owners of the future, hoovering up all the AI PhDs, and filing thousands of AI-related patents. **By <u>default</u>**, **AGI will be owned by profit-motivated private enterprise**.

Consequently, **by <u>default</u>, and simply because of the way in which the world works now, an increasing majority of the world’s wealth will be owned by a decreasing minority of its population (the AI-owners), and non-AI-owners will compete for the steadily-shrinking pool of jobs which can't be done better and more cheaply by AI-based machines.**

In this scenario, it will be up to national governments to redistribute the AI/AGI-generated wealth, e.g. through taxation. That will be resisted by special interests having almost unlimited resources, and the net result will be gross wealth inequality, exactly as Hawking predicted: "most people can end up miserably poor if the machine-owners successfully lobby against wealth redistribution".

It gets worse. Multiple commentators, Hawking included, have warned of the potential existential dangers of AI/AGI. At the launch of the Leverhulme Centre for the Future of Intelligence on 19 October 2016, Hawking stated: "I believe there is no deep difference between what can be achieved by a biological brain and what can be achieved by a computer. It therefore follows that computers can, in theory, emulate human intelligence — and exceed it. ... Every aspect of our lives will be transformed. In short, success in creating AI could be the biggest event in the history of our civilisation. But it could also be the last, unless we learn how to avoid the risks. Alongside the benefits, AI will also bring dangers, like powerful autonomous weapons, or new ways for the few to oppress the many. It will bring great disruption to our economy. And, in the future, AI could develop a will of its own — a will that is in conflict with ours. In short, the rise of powerful AI will be either the best, or the worst, thing ever to happen to humanity. We do not yet know which."

In his bestselling 2016 book, Superintelligence, Oxford professor, and director of the Future of Humanity Institute, Nick Bostrom further warned of the dangers of AGI. And, in their 2018 report, The Malicious Use of Artificial Intelligence, Brundage et al warned that any AI system on the AGI-by-stealth path from contemporary AI to AGI could potentially be misused by malicious actors.

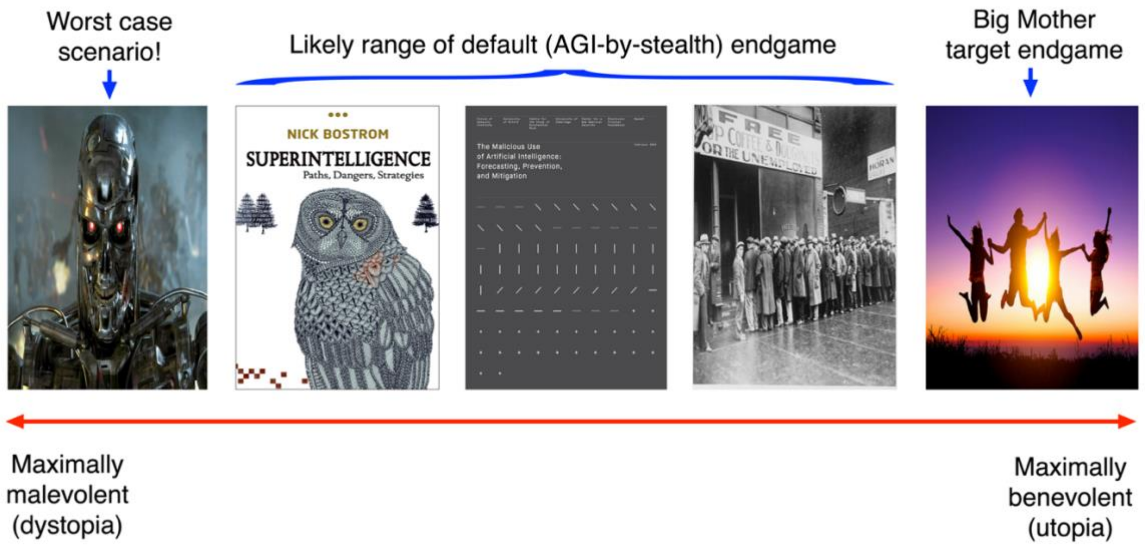
In the final analysis, the officers of profit-motivated corporations - the dominant AI-owners of the future - have a fiduciary duty to maximise shareholder value. Consequently, slick promises of ethical behaviour in respect of the exploitation of AI will, in many cases, not be worth the electrons with which they are communicated. In today's ever-more-competitive world, many rationalised sacrifices will be made to the God of First-Mover Advantage, resulting in many AI-based systems that are over-hyped, released before they're ready, not fit-for-purpose, or fundamentally unsafe.

Let's not kid ourselves - that's basically what's going to happen, and it's going to be a train wreck.

As AI edges ever closer to AGI, that's when these shortcuts will become existentially serious. In addition to the most appalling wealth inequality, we will have to contend with everything from "Artificial Stupidity", to AI-assisted crime, rogue robots, and possibly even rogue AGIs/near-AGIs.

No fate but what we make

The future isn't set. The spectrum of possible AGI endgames ranges from dystopia to utopia:


Worst case scenario!
Likely range of default (AGI-by-stealth) endgame
Big Mother target endgame
NICK BOSTROM
SUPERINTELLIGENCE
Paths, Dangers, Strategies
The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation
FREE COFFEE & DOUGHNUTS FOR THE UNEMPLOYED
Maximally malevolent (dystopia)
Maximally benevolent (utopia)

Here's the rub. By definition, any AGI, as well as being super-intelligent, is super-knowledgeable, and knows everything about Computer Science, Artificial Intelligence, and its own design. If it wants to, it can replicate itself, meaning it will "live" potentially forever. And, should we decide, "oops, sorry, didn't mean that, can we just switch it off and make a few more mods...?" it's too late - depending on its top-level goal, it might let us switch it off, or alternatively it might not. And if it's the latter there's no way we're going to trick it somehow because **it's smarter than we are**.

So, whatever AGI endgame we end up with, that's it **- forever**. We're no longer the smartest entity on the planet, and the fate of all mankind, for all eternity, is now sealed, and out of our hands.

By all accounts, this eventuality is just 50 to 100 years away, i.e. **a week next Tuesday**.

The Gold Standard of AGI

At Big Mother, we aim to build the **Gold Standard** of Artificial General Intelligence:

- maximally **safe**
- maximally **benevolent**
- maximally **trustworthy**
- **publicly owned** by all mankind
- **does all the work** so humans are **finally free to simply enjoy life**
- **shares the wealth** generated by AGI **equally**, on the basis of **need**
- need is determined by whatever allocation **maximises happiness**

It's not a race. We are not in competition with other AGI projects, we're just following a different path. As a minimum, we hope that the mere existence of our project will influence other AGI projects for the better. And if another AGI project achieves a maximally-safe, maximally-benevolent, and maximally-trustworthy AGI before we do then **fantastic -** time to kick back!

But, until that happens, we will press ahead towards that goal.

YOU can help. We need philanthropic donors, and volunteers both technical and non-technical.

Future generations will thank you! :-)

Aaron Turner is Project Director at BigMother.AI CIC (https://bigmother.ai)