DM Révision Méthode Pré CB PT Février 2025

Super-AI : imaginary or real threat ?

Document 1 - Fears about AI's existential risk are overdone, says a group of experts

***The Economist*, 1 July 2023**

In the past year, as the startling capabilities of artificial intelligence (AI) have emerged into public view, attention has been drawn to the existential risk, or "x-risk", that the technology may pose. The concern is that computers endowed with superhuman intelligence might destroy most or all human life. The majority of researchers raising the alarm are sincerely motivated by concern about AI-related-risks, present and future. However, calls to action to mitigate superintelligent-AI x-risk may both impede the development of beneficial uses of AI of which there are many-and distract regulators, the public, companies and other researchers from addressing important shorter-term risks.

Superintelligence is not required for AI to cause harm. That is already happening. AI is used to violate privacy; create and spread disinformation; compromise cyber-security and build biased decision-making systems. The prospect of military misuse of AI is imminent. Today's AI systems help repressive regimes to carry out mass surveillance and to exert powerful forms of social control. Containing or reducing these contemporary harms is not only of immediate value, but is also the best bet for easing potential, albeit hypothetical, future x-risk.

It is safe to say that the AI which exists today is not superintelligent. But it is possible that AI will be made superintelligent in the future. Researchers are divided on how soon that may happen, or even if it will. Still, today's AI models are impressive, and arguably possess a form of intelligence and understanding of the world; otherwise they would not be so useful. Yet they are also easily fooled, liable to generate falsehoods and sometimes fail to reason correctly. As a result, many contemporary harms stem from AI's limitations rather than its capabilities.

It is far from obvious whether AI, superintelligent or not, is best thought of as an alien entity with its own agency or as part of the anthropogenic world, like any other technology that both shapes and is shaped by humans. But for the sake of argument, let us assume that at some point in the future a superintelligent AI emerges which interacts with humanity under its own agency, as an intelligent non-biological organism. Some X-risk-boosters suggest that such an AI would cause human extinction by natural selection, outcompeting humanity with its superior intelligence.

Intelligence surely plays a role in natural selection. But extinctions are not the outcomes of struggles for dominance between "higher" and "lower" organisms. Rather, life is an interconnected web, with no top or bottom (consider the virtual indestructibility of the cockroach). Symbiosis and mutualism – mutually beneficial interaction between different

Species – are common, particularly when one species depends on another for resources. And in this case, Als depend utterly on humans: from energy and raw materials to computer chips, manufacturing, logistics and network infrastructured we are as fundamental to Als' existence as oxygen-producing plants are to ours.

[...] Regulators should not prioritise existential risk posed by superintelligent Al. Instead, they should address the problems which are in front of them, making models safer and their operations more predictable in line with human needs and norms. Regulations should focus on preventing inappropriate deployment of Al. [...]

Document 2 - Artists may make Al firms pay a high price for their software's 'creativity'

John Naughton, *The Guardian*, 28 October 2023

Those whom the gods wish to destroy they first give access to Midjourney, a text-to-graphics "generative Al" that is all the rage. It's engagingly simple to use: type in a text prompt describing a kind of image you'd like it to generate, and up comes a set of images that you couldn't ever have produced yourself. For example: "An image of cat looking at it and 'on top of the world', in the style of cyberpunk futurism, bright red background, light cyan, edgy street, art, bold colourful portraits, use of screen tones, dark proportions, modular" and it will happily oblige with endless facility.

[...] Many people think it's magical, which in a sense it is, at least as the magician Robert Neale portrayed it: a unique art form in which the magician creates elaborate mysteries during a performance, leaving the spectator baffled about how it was done. But if the spectator somehow manages to discover how the trick was done, then the magic disappears.

So let us examine how Midjourney and its peers do their tricks. The secret lies mainly in the fact that they are trained by ingesting the LAION-5B dataset - a collection of links to upwards of 6bn tagged images compiled by scraping the web indiscriminately, and which is thought to include a significant number of pointers to copyrighted artworks. When fed with a text prompt, the Als then assemble a set of composite images that might resemble what the user asked for. *Voilà!*

What this implies is that if you are a graphic artist whose work has been published online, there is a good chance that Midjourney and co have those works in its capacious memory somewhere. And no tech company asked you for permission to "scrape" them into the maw of its machine. Nor did it offer to compensate you for so doing. Which means that underpinning the magic that these generative Als so artfully perform may be intellectual property (IP) theft on a significant scale.

Of course the bosses of Al companies know this, and even as I write their lawyers will be preparing briefs about whether appropriation-by-scraping is legitimate under the "fair use" doctrines of copyright law in different jurisdictions, and so on. They're doing this

because ultimately these questions are going to be decided by courts. And already the lawsuits are under way. In one, Some graphic artists launched a suit against three companies for allegedly using "their original works to train their Als in their styles, thereby enabling users to generate works that may be insufficiently transformative from the original protected works - and in the process generating unauthorised derivative works.

Just to put that in context, if an Al company was aware that its training data included unlicensed works, or that its algorithms generated unauthorised derivative works not covered by "fair use", then it could be liable for damages of up to $150,000 for each instance of knowing use. And in case anyone thinks that infringement suits by angry artists are like midge bites to corporations, it's worth noting that Getty, a very large picture library, is suing Stability Al for alleged unlicensed copying of millions of its photos and using them to train its Al, Stable Diffusion, to generate more accurate depictions based on user prompts. The inescapable implication is that there may be serious liabilities for generative Als coming down the line.

Now, legal redress is all very well, but it's usually beyond the resources of working artists. And lawsuits are almost always retrospective, after the damage has been done. It's sometimes better, as in rugby, to "get your retaliation in first". Which is why the most interesting news of the week was that a team of researchers at the University of Chicago have developed a tool to enable artists to fight back against permissionless appropriation of their work by corporations. Appropriately, it's called Nightshade and it "lets artists add invisible changes to the pixels in their art before they upload it online so that if it's scraped into an Al training set, it can cause the resulting model to break in chaotic and unpredictable ways" - dogs become cats, cars become cows, and who knows what else? (Boris Johnson becoming piglet, with added grease perhaps?) It's a new kind of magic. And the good news is that corporations might find it black. Or even deadly.

Document 3 - What would humans do in a world of super-Al?

***The Economist*, 23 May 2023**

In "WALL-E" , a film released in 2008, humans live in what could be described as a world of fully automated luxury communism. Artificially intelligent robots, which take wonderfully diverse forms, are responsible for all productive labour. People get fat, hover in armchairs and watch television. The “Culture" series by lain M.Banks, a Scottish novelist, goes further, considering a world in which Al has grown sufficiently powerful as to be superintelligent? operating far beyond anything now foreseeable. The books are favourites of Jeff Bezos and Elon Musk, the bosses of Amazon and Tesla, respectively. In the world spun by Banks, scarcity is a thing of the past and Al "minds" direct most production. Humans turn to art, explore the cultures of the vast universe and indulge in straightforwardly hedonistic pleasures.

Such stories may seem far-fetched. But rapid progress in generative Al? the sort that underpins OpenAI's popular chatbot, ChatGPT-has caused many to take them more

seriously. On May 22nd OpenAI's founders published a blog post saying that "it's conceivable that within the next ten years, AI systems will exceed expert skill level in most domains, and carry out as much productive activity as one of today's largest corporations." Last summer forecasters on Metaculus, an online prediction platform that is a favourite of many techies, thought it would take until the early 2040s to produce an AI capable of tricking humans into thinking that it was human after a two-hour chat, had good enough robotic capabilities to assemble a model car and could pass various other challenging cognitive tests. After a year of astonishing AI breakthroughs, Metaculus forecasters now think that this will happen by the early 2030s! There is no shortage of money for research, either. Five new generative-AI unicorns (startups valued at $1bn or more) have already been minted this year.

The road to a general AI – one better than the very best of humanity at everything – could take longer than expected. Nevertheless, the rising possibility of ultra-powerful AI raises the question of what would be left for humans when it arrives. Would they become couch potatoes as in "Wall-E"?

[...] In 2019 Philippe Aghion, Ben Jones and Chad Jones, three economists, modelled the impact of AI. They found that explosive economic growth was plausible if AI could be used to automate all production, including the process of research itself-and thus self-improve. A nearly unlimited number of AIs could work together on any given problem, opening up vast scientific possibilities. Yet their modelling carried an important caveat. If AI automated most but not all production, or most but not all of the research process, growth would not take off. As the economists put it: "Economic growth may be constrained not by what we do well but rather by what is essential and yet hard to improve."

[...] It seems unlikely that people will give up control of politics to robots. Once AIs surpass humans, people will presumably pay even loser attention to them. Some political tasks might be delegated: humans could, for instance, put their preferences into an AI model that produces proposals for how to balance them. Yet as a number of political philosophers, including John Locke n the 17th century and John Rawls in the 20th, have argued, participation in political procedures gives outcomes legitimacy in the eyes of fellow citizens. There would also be more cynical considerations at play. Humans like to have influence over one another. This would be true even in a world in which everyone's basic needs and wants are met by machines. Indeed, the wealthiest 1% of Americans participate politically at two to three times the rate of the general public on a range of measures from voting to time spent on politics.

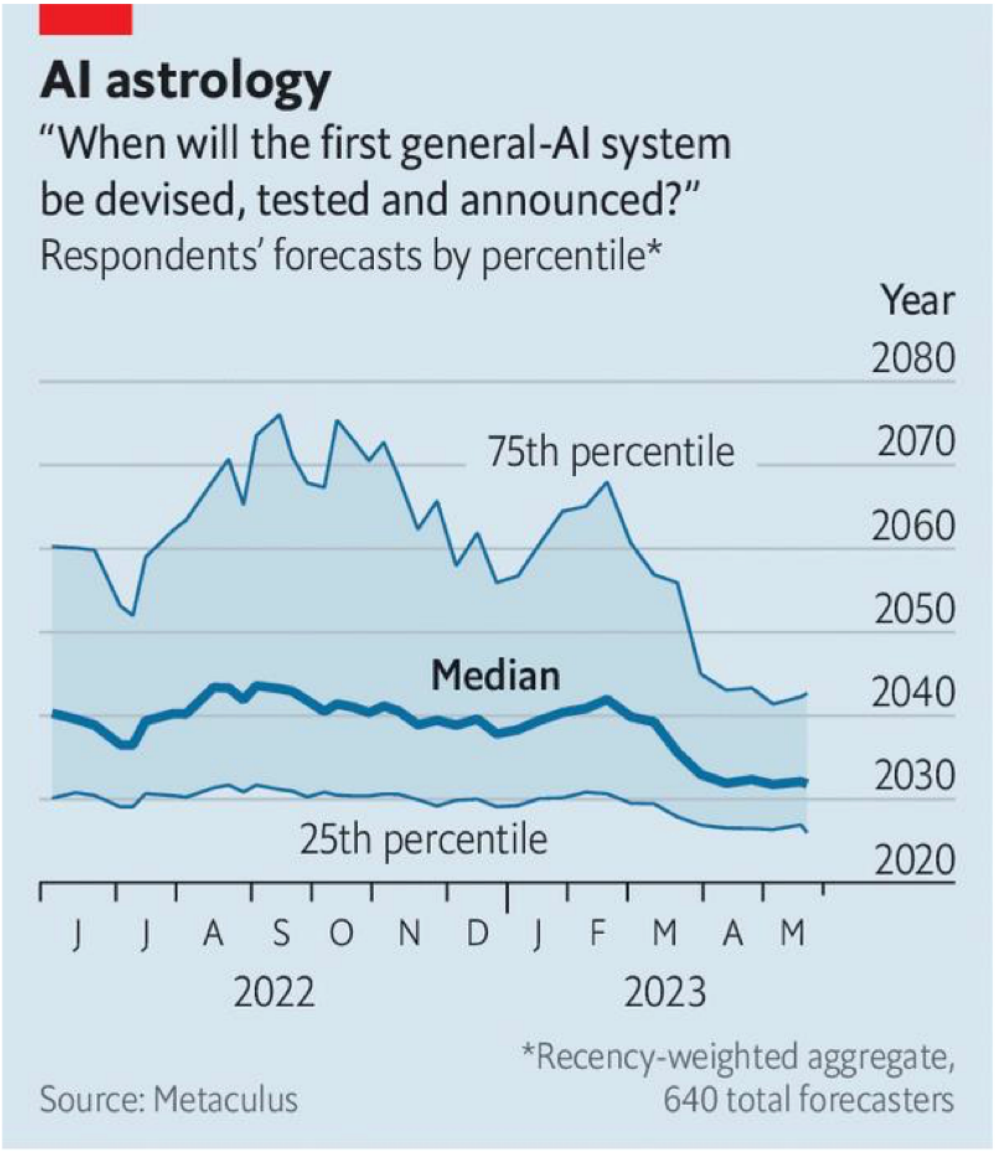
Last, consider areas where humans have an advantage in providing a good or service-call it a "human premium". This premium would preserve demand for labour even in an age of superadvanced AI. One place where this might be true is in making private information public. So long as people are more willing to share their secrets with other people than

machines, there will be a role for those who are trusted to reveal that information to the world selectively, ready for it then to be ingested by machines.

The human premium might appear elsewhere, too. People value history, myths and meaning.

[...] In areas such as caregiving and therapy, humans derive value from others spending their scarce time with them, which adds feeling to an interaction. Artificial diamonds, which have the same molecular structure as those from the ground, trade at an enormous discount – around 70% by one estimate. In the future, items with a "made by a human" tag might be especially desirable.

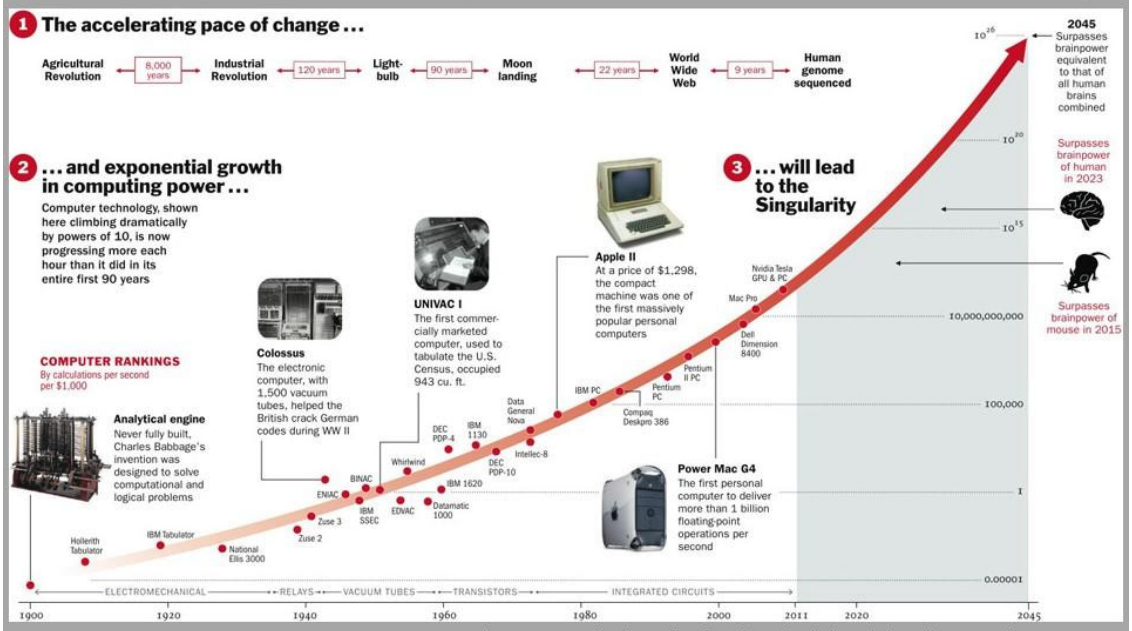
Document 4 – AI Astrology



Document 5 –

Will we have a mature Superintelligence by 2050?

Futurist Ray Kurzweil - 147 predictions since 1990 with 86% accuracy



October 2017: Kurzweil + Diamandis: Disruptive Technologies, Mind-Boggling ...NextBigFuture.com

We are no longer able to halt this trend