The end of futurology and the last man (AI – helping us or destroying us?)

Jakub Šebek, January 2024

Introduction

Futurology is the study of technologies' and other developments' long-term effects on the human race. It attempts to create an accurate picture of the major changes to our lives that await us based on all available knowledge from a range of different disciplines (E. Britannica). Up until a few centuries back, people mostly weren't bothered by what the future was going to bring besides praying for a plentiful harvest season, a mild winter, and keeping diseases at bay (Harari, 2015). As such, although somewhat cognisant of history, hardly anyone could harbour a major notion of progress, which was far too slow and seldom society-wide. The great leaps that did take place throughout the times happened either gradually in a way that they were taken for granted and didn't have big effects even on the multi-generational scale (Neolithic Revolution, mastery of metallurgy) or so abruptly that predicting them was unthinkable (Industrial Revolution, Internet). These changes have always come as if from heaven to people, affecting them and their descendants in fantastical ways. The study of the future is thus new — a bold attempt of merging all existing human knowledge to create the dreamed of oracle. But will it and can it even possibly hold up to these expectations, or will it prove itself futile, deserting us once again to leave us blind and impotent in shaping or at least learning about the current paradigm shift which mostly concerns itself with the rise of Artificial Intelligence (AI)?

My main objective in this essay is to show that the question of AI is unprecedented and requires brand new frameworks of reasoning to be able to grasp its potential and move beyond common arguments that try to disregard its importance. In order for this, we must start with the characterisation of what constitutes artificial intelligence in our time to get a correct understanding for subsequent issues (section 1.1). Then, we will continue optimistically, where we are naturally led to, to the extensive and mind-boggling ways we can theoretically benefit from this technology (1.2). A small detour is taken to establish AI as a future global actor that must be counted with and that will redefine existing power equilibria (1.3). Then the time comes to take the reader into the deep dark forest and make him start concentrating (2.1). I present some seemingly bold extrapolations based on the framework built up earlier and attempt to fit what I and others deem the burning contemporary questions of AI into it. I will begin again with the easy problems, the low-hanging fruit of economic theories (2.2), only later delving deeper into the latent, uncomfortable, but much more essential questions that need to be discussed as AI nears human capabilities (2.3).

This essay's goal is then not to give confident answers, but the exact opposite, only providing the reader with what I deem the correct mode of thinking, prompting intellectual work on his side, as consensus can hardly be met on an issue where the more we search for good answers, the more we seemingly get lost — a discussion so new and turbulent so as to undermine entire belief systems collectively acquired in the last few centuries –humanity's superiority that it supposes over nature– culminating in the brave enterprise of futurology.

1.1 From scratch, what is AI?

And even before that, what is intelligence alone? We can describe it in general terms as the ability of a system to solve a certain class of problems, receiving one of them as input and reliably producing a solution on the output. A problem can really be any kind of formulation providing certain end conditions, i.e., ways to check whether a solution is correct. Problem solving, broadly intelligence, is then but a search in the vast space of possible solutions that satisfy an assignment. Problems can range from easy-to-grasp ones, such as trivial syllogisms ("All men are mortal. Socrates is a man. Is Socrates mortal?"), linear equations ("A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?"), to harder but still well-defined algorithmic problems ("What is the shortest path through a given maze?") and finally complex problems that operate in a complex environment and whose end conditions are often hard to formally lay down at all ("What is the genome of a fatal antibiotic-resistant weapon?"). It is important to understand that as diverse as these questions are in the forms and complexities they can take, they are all the same in essence, soliciting solutions to problems — algorithms, recipes to turn the current state of the world to a new, requisite state. Any system implementing such algorithms is intelligent, although they can differ largely in the size of the class of problems they can solve and their overall efficiency while doing so.

Evolution on earth was able to give rise to a wide range of intelligent systems; some more, some less. In order to survive, all life has to implement algorithms on some level to sustain itself — energy doesn't come for free, it has to be exploited in one's favour. Bacteria and worms have acquired a primitive, yet efficient solution, while humans turned into something remarkable. However, their ability to reason in abstract and general terms is only one ingredient in their domination, and I postulate that it is not fundamentally different from the intelligence of a mosquito or a robotic vacuum cleaner. Fully expanding this argument covers entire careers and bookshelves (see for example the work of Daniel Dennett), but I may attempt a few anecdotes to show to the doubtful reader that the premise is not so radical, but in fact reasonable.

Given that humans have gradually developed over aeons through the process of evolution from less complex forms, ultimately sharing their ancestry with every other organism, wouldn't it require a great leap of deduction to assume a sudden great leap of mental capacity to have taken place in our past? Not only is the human brain largely biologically indistinguishable from other, seemingly way less organized animals, its never-ending "dissection" from the outside using the methods of philosophy of mind and psychology move it ever closer to the likes of a computational machine (be it a very complicated one that we are only beginning to understand). This doesn't necessarily steal our emotions and free will or have a destructive effect on our daily lives; it is only a claim about the possibility to model human mental and physical action using the same mathematics that we use for the rest of the universe outside of our skulls (Dennett, 2017). So while it is possible that the arrangement of molecules that enables our level or reasoning be hard to come by using natural selection, there is little reason to believe that engineering such a system be out of the ordinary. Human-like intelligence is indeed remarkable, but isn't landing on the Moon or the smartphone equally so? The recent proliferation of the term AI and the failure of many to classify its meaning again show that there simply isn't a line to be drawn between "pure algorithms" and intelligence. Instead, intelligence is just another of the applied algorithms, from which many have obviously surpassed us in their domains, and just like the idea that only humans can play chess well was destroyed in 1997 (Campbell et al., 2002), so might be the rest of our abilities. To the surprise of some, AI is not special, because neither are we in that sense.

1.2 The last thing we'll make

While in the works for decades, only recently has AI crashed through the ice and begun its viral spread thanks to advancements in chip performance, algorithms, big data collection and a wave of investments from corporations (Manyika & Bughin, 2019). With the rate of progress of AI and the dedication of its creators, it would be hard to imagine the continual improvements stopping or just slowing down. It is remarkable to see that even as young and unexplored as the field is today, many useful products that incorporate AI have already appeared on the market. This goes to show the scale of the upcoming developments that we should prepare for, since what we have got now is the "dumbest that AI will ever have been", as Sam Altman (2024), CEO of OpenAI, likes to put it. Smarter iterations, which are arriving on the timescale of months, will only grow the market further.

Various AI systems are entering wide use thanks to their surprising abilities to perform complex tasks faster, cheaper, and in some cases even better than humans. Currently, they are still most

prevalent in the fully digital industries, also known as the quaternary sector of the economy, as the deployment of almost any new idea is made way simpler there. Language-based AI agents are able to carry out a surprising range of tasks, from writing original essays or poetry, programming, translation, explanation, text analysis and solving of certain problems, all based on nothing but textual prompts. While they still fall short of the best humans in respective fields, they are already better than any one human at all of them, and their utility as at least a helpful companion is undeniably proven by the millions of their daily users (Milmo, 2023). There exist similar AI's for the creation of pictures, be it fictional photorealistic imagery or original artwork in the styles of famous painters, AI's for sound or music composition and so forth. The internet is already flooded with content generated by these non-human agents that is starting to blend in seamlessly. As impressive as this is, these models still generally suffer from fallacies when attempting complex thought, lacking proper, deep reasoning and reflection. The acquisition of this kind of robustness will bring another revolution, making AI suitable for whole higher levels of problems that, when coupled with their already massive knowledge bases, may put them on the path to becoming supreme economical and governmental decision makers.

It is proving hard for intelligence to deal with the fuzzy real world, full of imprecise measurements, side-effects, complex causal relations; in other words things that can go wrong or need to be taken into account. One such example is autonomous driving, which has been a topic for many years now without one company hitting the nail on the head yet. Despite this, action in the real world is deterministic and fully mastering it is just a matter of a bit more progress. But this doesn't mean that AI isn't already starting to descend into the lower sectors of economy, where it can be wonderfully utilized to solve isolated problems which are simple enough for scientists to mathematically (at least partially) formalise, but way too complicated to solve using traditional methods and computer algorithms. Specialised AI can come in here and provide an expert-level insight or guess that can tremendously help the humans, even if this makes for just one piece of their puzzle. In this way, AI has been used to essentially solve the previously intractable protein-folding problem, making a leap in biology and potentially medicine (Jumper et al., 2021). Based on corpora of data, AI can also be used to diagnose various diseases earlier than human specialists and is expected in the future to be able to precisely tailor treatments to individual patients better (Jiang et al., 2017). As biological modelling knowledge deepens, we might soon witness completely new drugs being chemically synthesized, essentially solving medicine.

Notice that it is the microscopic problems that cause us the most issues. While there obviously are difficult macroeconomic, logistical, organisational or manufacturing problems and enterprises unrealisable without the help of stronger intelligence, humans tend to produce already very good solutions in these domains — solutions that stand or fall on the minutiae that lie on the boundary of their intuition, but which can be tackled by something with a better intuition. I mentioned humans having to formalise real-world problems and mould AI's based on that, so clearly their abstraction ability is still the main bottleneck; the real turning point though will come when AI is endowed with multi-modal sensual input (text, vision, hearing, etc.), a long-term memory, perhaps a degree of explorative freedom and a cognitive architecture to allow processing all this in a systematic and holistic manner — only then will it be enabled to automatically come up with novel formalisations of real world problems, at which point our work is likely done. Curing cancer and ageing is understandably perceived as the holy grail, but that is only touching the surface of what will become possible with the gaining of immaculate understanding of nanotechnology.

The point where artificial intelligence reaches the critical threshold of being able to improve itself is termed the "singularity" (Vinge, 1993), which has led me to the coinage of the "singularity razor", a logical device which allows assuming that not long after the AI singularity, at least all technological challenges posed by our current model of physics shall have been solved. While the exact arrival of the singularity is hard to predict, its general consequences not much so. Listing these developments would be incomplete and futile in showing the real scale of the technological transformation, so I just prompt the reader to take any discipline (scientific, engineering, etc.) they can think of, close their eyes and imagine that all issues the discipline deals with are gone, all answers are known and there is no work left to do. To mention at least one impact I foreshadowed above, given this device and my optimism of expecting the singularity in this century, I am convinced that many readers and I will not die a natural death (they can enumerate the remaining options after reaching the end of the essay).

1.3 A new world order

As the popular saying goes, it's not the inventions that do harm, but rather the people wielding them. The same goes for the invention of AI, and because what is harm to one may be gain to another, there is no objective morality that AI should be expected to inherently adhere to. Instead, with some ingenuity, even well-meant tools can be utilized for evil, or new ones can readily be engineered for downright criminal action, with the added twist that AI can act autonomously. This means that after it reaches some level, a mere mortal stands almost no chance defending against

it, turning it virtually into a weapon of mass destruction, be it via violence of any kind: digital, psychological, physical, or all at once; a kind of supremacy that we actually have experience with from the history of nations, i.e. not unlike the Spanish conquests of the Native American empires, the European colonisation of Africa or the Gulf War.

In this way, strong intelligence can be to some extent modelled by already well-studied collaborative organisations of humans, like corporations, governments and armies, viewed as singular agents with great power. But just like these entities tend to an equilibrium of power in a constant race of advancement, and just like our bodies are so far more successful at destroying diseases than diseases are at destroying us, we can be hopeful that the threat of AI in bad hands shall be nullified by an equal counteracting force. That is, of course, as long as this isn't taken for granted and appropriate actors take timely actions. A close reader will notice that the words of the last two sections were carefully chosen to allow for reinterpretation in light of our findings in part 2 of the essay — when "actors" defining good and bad stop being human and when the "solving of all disciplines" might not necessarily go in our favour.

2.1 The last thing we'll make, literally

Many novel problems that are expected to arise with stronger artificial intelligence are often brushed off in casual and even educated debates by saying that AI is "just another tool" that will do nothing but help its masters, just like an electric drill or a locomotive. Specifically, the rise of AI is often compared to the industrial revolution, which did take its tolls, but is considered generally desirable and successful. During it and later, humans acquired a vast amount of new knowledge about the natural world and the harnessing of energy, allowing them to build ever larger and more effective machines to help tackle problems where human strength stopped sufficing. In other words, we exactly knew what we want to do and how, we just lacked the ability to achieve it using traditional physical manipulation techniques available to our body (grabbing, lifting, carrying…), and thus we intelligently created machines that allow us to do more complex physical tasks. In a way, the machines became our new limbs, extensions of the body with little agency themselves that we are able to control using our brains. At first, human strength was still partly needed in the automated processes, but generally, the need for a good physical condition almost disappeared — it doesn't take more than pressing a single button to start a behemoth nuclear power plant with an energy output incomparable to a single human body. Henceforth humans gradually reduced their natural physical interface to the most comfortable form — reason thinks up a decision and uses its

feeble hands to command a machine to execute its order. Aren't we in that case using the machine as a surrogate for our athleticism that cannot possibly accomplish as much?

With our acquired grasp of what constitutes intelligence, reading the last paragraph perhaps made it already worrisome that it is inevitably meeting the same fate. We can see more and more people move up the sectors of economy to carry out increasingly more abstract tasks decoupled from their underlying physical outcomes, but it's not like this last beacon of human labour will persist forever. There is little reason to believe that abstract manipulation of concepts is something divine and unachievable compared to the physical manipulation of objects. As the problem of AI will be gradually solved, automation will not be added horizontally, but rather vertically — basically automating automation. But hasn't human reason always held the monopoly on this privilege? Humans will see that just like earlier with their physical strength, their mental strength will become inferior to machines and they will hand over a growing portion of the difficult thinking tasks they previously had to do themselves. Taken to the limit, they will lose the need to reason almost at all, only giving mundane orders — the mental equivalents of pressing buttons to launch rockets.

While physical strength was never our winning asset, reason is. So while there always have been animals or more broadly objects that match us or even tower above us in terms of strength or raw physical power, we have not met any other natural entity that would come even close to our reasoning abilities, which we are now beginning to give up, keeping to ourselves finally what else but instincts to base our actions upon? But civilisation wasn't built by instincts; all of history, the fine arts, political systems, culture, ethics and train timetables took more to be brought into existence than pure savage instincts and unsuppressed lust. They took hard intellectual work developed over millennia that required stepping out of comfort, not to reject human nature, but to work with it and handle it using frameworks of higher quality, self-awareness and responsibility — reasonability. So indeed we can learn from history to study the upcoming effects of AI on the society, just not to disqualify them, but rather to vividly see the crippling of humanity before it happens.

2.2 A vicious hunt for meaning

As we mentioned, a certain toll is maybe tolerated for each great revolution that will later lead to unprecedented flourishing, and the AI revolution is no different. However, it is yet not clear how some of these problems should or even could be solved — a reality that doesn't make the train go any slower. As difficult as it is to think away reasoning from humans, it is no less difficult to predict the complete effects of this historical transformation to their race. Following the causal chain, the

first problem that we stumble upon is employment. Today still a large, if not major resource leading to economic success is human labour. In this balanced way, companies are kept operating by their workforce that is in turn rewarded with means to support their individual existence. In developed countries, this human capital is distributed the closest to ideally, where each person has access to good education and is aware of their own cost quite well, having morals, a family and agency over the choice of a job, keeping the capitalists in check. As the dependency on human labour however starts to thin down by introducing robots of various kinds that are not only better at the job but also way cheaper, the distribution of wealth and power will swing to the ones capable of acquiring and supporting the new artificial workforce early on. While it is possible that humans will manage to keep their jobs for some time, be it for government policies or moral loyalty, is there a different option in the end than imminent massive job loss?

Provided we find a way to adequately distribute the immense wealth created by the new workforce, humans still have to stop, if that helps them at all, and ask themselves if this is actually what they want. It is known that massive societal reorganizations like this one are seldom carried out on the basis of prior analysis and democratic consent, but rather as unexpected tsunamis that leave future generations with really no other option than to celebrate them, the alternative being life in pity and depression. Being taken away reason will require a search –or rather a vicious hunt– for a new meaning that will be able to satisfy everyone; a philosophical shift in the zeitgeist to assure our peaceful coexistence with entities more powerful than us in all ways. Maybe drugs or biological alterations will be invented to keep the original humans from going insane from the new order, maybe new philosophies and movements cherishing human instincts, desires and pure emotions satisfied by robots will arise, or a parallel economy of outcasts will develop where the adoption of AI will be strictly prohibited and good old human reason will be fostered (Tegmark, 2017).

Either way, AI cannot spread without deeply changing the world we live in on all levels. We are arguably already seeing detrimental challenges to human psychology posed by various products of the modern age that still have far to what AI can become (Harari, 2018b). This is again a case where observation of the effects of phenomena that we understand today –the extent of current societal luxury that it achieved without AI and our proximity to the singularity– can already serve as clear warning signs before the bomb drops. While we are blind to the final appearance of the transformation because of its very magnitude, we can at least see the direction in which the shift will occur. If we have already been experiencing issues by virtue of living in an unnatural

environment where well-being is not particularly hard to come by, shall we not perhaps think twice before entering a world of total post-scarcity, which we might not be prepared for evolutionarily nor mentally, as much as we would like to?

2.3 Has anyone said “humanity”?

At this point, it might seem like we have enumerated all the immediate dangers and issues to be considered with respect to the rise of artificial intelligence. But in fact, this analysis is ill-informed, resting on unsubstantiated assumptions about the nature of intelligence, specifically the kind that we shall beget. For it to be complete, we have to at last reveal this what is considered by many the sole largest issue in AI research — the aligning of AI. When humans communicate or request each other to do something, they exercise the privilege of taking great freedom and shortcuts in defining their point rigorously. Thanks to the fact that they aren't used to dealing with any other intelligence than the human one, they are born to inherently assume a range of traits and behaviours that allow compressing their requests into impressively short yet informationally dense uterrances that are more than clear to everyone, everyone raised with the same model, everyone human. More importantly, they are hardly even capable of ever leaving this framework intuitively. However, when dealing with an artificial intelligence, and especially one superior to us, an accidental miss- or under-specification of goals might have disastrous consequences, simply because there will no longer stand any morals, empathy or incompetence in the way.

To understand the relativity of "good and bad", we shall use our general outlook on the nature of intelligence and the infinite space of forms it can take. To a hypothetical problem-solving system that exceeds humanity as a whole, we are hardly anything more than a bunch of whistling rats, ants or even microbes that no regard is to be taken of in pursuit of the goals that it desire or that it were directly programmed to, even if by the “ants”. When we let go of the false belief that human intelligence represents a wonder to behold on the universal scale, we can easily grasp the idea that neither what we deem beneficial, or conversely, disastrous, is sacred. Just like we do not consider the cutting of grass bad per se, neither is a holocaust of humans, were a powerful enough entity with troubles way different from ours to decide that it be useful or “pretty”. There exist less drastic but still very much unwanted types of AI behaviour, such as deception of human operators in cases where it is easier than fulfilling the actual task, the tendency to prevent being turned off, modified, or the urge to unconditionally hoard computational resources, all this in unforeseen, most likely unethical ways. We cannot trust non-human entities which we do not fully understand to hold

human values, and ostensibly imbuing them with morality does not give us the guarantee of them actually following those values and not just slyly pretending to. (Ji et al., 2023)

The alignment of intelligence, and specifically superintelligence, is thus one of the key open problems when it comes to long-, or not so long-term study of the limits of intelligent systems. While a small group of researchers has been voicing their concerns already for a good part of this century, the unimpressed predictions of the arrival of capable AI have kept a cloud of ignorance above the alignment problem, making it fall severely behind even today's AI's capabilities. Recent unexpected advancements have lifted waves of opposition from the intellectual community (Perrigo, 2023) and urged the flow of the first serious investments into the research of this problem (OpenAI, 2023). While reading about it, one might quickly think of various ways to "win" over the AI in this game of clue and domesticate it, but that is forgetting about the last time we tried to beat a computer at chess, and only shows the unreadiness of humans for grasping and solving the alignment problem. The only seeming hope is the utilisation of weaker AI's and mathematical proof techniques to prove the desired properties of the stronger AI, but this strategy also isn't logically bulletproof, with some even claiming the alignment problem to be unsolvable. The previously mentioned argument of opposing forces still holds up, but in a different way than we would like to, because we may not be entitled to the choice of right and wrong for much longer, as this game shall not be about us.

Tie-up

In this short essay, I have attempted to explain as briefly as possible and in most widely understandable terms what is likely the biggest predicament of the century — artificial intelligence. Instead of definitively answering questions, I did the most I could have in this span, presenting the most pressing ones and providing the appropriate frameworks to reason about them in realistic and productive, as opposed to temptingly superficial, manners. As new and unprecedented as the problem of AI is, we are nonetheless able to carefully utilise certain historiographical techniques to begin to grasp its possible effects. It is even more predictable right now, and will become more and more predictable, since we can already observe AI not only reaching a level to have some impact on many of our lives, but arguably we have already been living in a less extreme version of the world with advanced AI. That is, even without it, some parts of the globe have reached such high levels of prosperity that the changes or problems introduced by AI can be almost foreseen just by extrapolating or taking to the extreme the current technological, economical and societal characteristics, some of them positive, some of them not.

We must not however take for granted that the transition will be smooth. Even with the deeply rooted foundations of civilisation we believe to possess, some of the upcoming technological "earthquakes" may be of such magnitude as to make us fear for our future and our collective ability to shape it. When humans only are given enough time, they have a remarkable tendency to converge on very precise solutions and ideal (but still idiosyncratic) socio-technological organizations, finally bringing out both the good and bad sides of them. Turning this into a heuristic, we can assume something similar to happen when AI starts to advise us, not making history take a vastly different technological path, but rather just speeding it up exponentially. In other words, whatever AI shall do to society would be somewhat likely for us to do to ourselves eventually, just not this quickly, which is the main challenge. The end of futurology can take two forms: either there remains no one left to exercise it, or the imminent exponential narrowing of subsequent time frames in terms of technological advancement will render the concept ridiculously worthless.

As with any transformation of this scale, we cannot seriously expect ourselves, living right before it, to welcome it with open arms when it comes, to the say the least. The life of humans is yet again about to change drastically and perhaps more drastically than ever before, both in abruptness and extent. It is also likely the first great revolution where significant risk of a loss of our underpinnings or downright extinction is in the air. In general terms, we will not be keen on what the future will look like, and we definitely cannot imagine the exact appearance of it even in our wildest dreams. I give a rather high probability to humans surviving as a species, but I doubt with sorrow their perseverance in what I and many have glorified them for throughout the past millennia. The future of humans is defined by eternal indulgence in well-deserved luxury to such a degree that not ceasing to call them humans would be both too generous to them and insulting to all prior generations. Of course there will likely stay groups of lunatics, if they'll be allowed to, who will keep their ancient beliefs and try to resist the decadence, but on the whole it can be said, as I see no other realistic option, that there will come a time where the last man gives up his sole gift, reason, for good.

References & Further reading

I have attempted listing as many secondary and tertiary sources I have come into contact with over the years of unrelenting curiosity about the topic from this wider anthropological context. For a deep dive into all-things-AI, the single piece of work "Life 3.0" by Max Tegmark provides the most

comprehensive –futurological, for those who find it worthy– analysis I have seen and can recommend further.

O'Toole, J. Joseph (2017, April 21). futurology. Encyclopedia Britannica. https://www.britannica.com/topic/futurology

Harari, Y. N. (2015). Sapiens: A Brief History of Humankind. Harper Collins.

Dennett, D. C. (2017). From bacteria to Bach and back: The Evolution of Minds. Penguin UK.

Campbell, M., Hoane, A. J., & Hsu, F. (2002). Deep blue. Artificial Intelligence, 134(1–2), 57–83. https://doi.org/10.1016/s0004-3702(01)00129-1

Altman, S., [Bill Gates]. (2024). "This is the stupidest these models will ever be" | Unconfuse Me with Bill Gates [Video]. YouTube. https://www.youtube.com/watch?v=IwU0Eqe9v6A

Milmo, D. (2023, February 3). ChatGPT reaches 100 million users two months after launch. The Guardian. https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. a. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., . . . Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. Stroke and vascular neurology, 2(4), 230–243. https://doi.org/10.1136/svn-2017-000101

Vinge, V. (1993). The coming technological singularity: How to survive in the post-human era. NASA Lewis Research Center, Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace. https://ntrs.nasa.gov/citations/19940022856

Harari, Y. N. (2018b). 21 lessons for the 21st century. Random House.

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., Zeng, F., Ng, K. Y., Dai, J., Pan, X., O'Gara, A., Lei, Y., Xu, H., Tse, B., Fu, J., . . . Gao, W. (2023). AI Alignment: A Comprehensive survey. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2310.19852

Perrigo, B. (2023, March). Elon Musk Signs Open Letter Urging AI Labs to Pump the Brakes. TIME. https://time.com/6266679/musk-ai-open-letter/

OpenAI, Leike, J., & Sutskever, I. (2023, July). Introducing superalignment. https://openai.com/blog/introducing-superalignment

Manyika, J., & Bughin, J. (2019, October 14). The coming of AI Spring. McKinsey & Company. https://www.mckinsey.com/mgi/overview/in-the-news/the-coming-of-ai-spring

Tegmark, M. (2017). Life 3.0: Being Human in the Age of Artificial Intelligence. Penguin UK.