AI & SOCIETY (2025) 40:1195–1208
https://doi.org/10.1007/s00146-024-01930-2

MAIN PAPER

The argument for near-term human disempowerment through AI

Leonard Dung[1]

Received: 7 January 2024 / Accepted: 4 March 2024 / Published online: 14 April 2024

Abstract
Many researchers and intellectuals warn about extreme risks from artificial intelligence. However, these warnings typically came without systematic arguments in support. This paper provides an argument that AI will lead to the permanent disempowerment of humanity, e.g. human extinction, by 2100. It rests on four substantive premises which it motivates and defends: first, the speed of advances in AI capability, as well as the capability level current systems have already reached, suggest that it is practically possible to build AI systems capable of disempowering humanity by 2100. Second, due to incentives and coordination problems, if it is possible to build such AI, it will be built. Third, since it appears to be a hard technical problem to build AI which is aligned with the goals of its designers, and many actors might build powerful AI, misaligned powerful AI will be built. Fourth, because disempowering humanity is useful for a large range of misaligned goals, such AI will try to disempower humanity. If AI is capable of disempowering humanity and tries to disempower humanity by 2100, then humanity will be disempowered by 2100. This conclusion has immense moral and prudential significance.

Keywords Existential risk · Scaling hypothesis · AI ethics · AI safety · AI alignment · Instrumental convergence

1 Introduction

A specter is haunting the world—the specter of an existential catastrophe caused by artificial intelligence (AI). While numerous distinguished and influential intellectuals, including Berkeley professor Stuart Russell and machine learning (ML) pioneer Geoffrey Hinton, have voiced dramatic concerns about the risks from advanced AI systems (e.g. Center for AI Safety 2023), there are few works which explicitly lay out, or discuss, why AI might pose a substantial risk to humanity as a whole. Moreover, most of the relevant discussion is contained within blogposts, internet fora, think tank reports and other informal venues. This is not to belittle non-academic work: many of these pieces are insightful and have served as an important source of inspiration and ideas for my argument here. The latter applies particularly to Carlsmith (2022) and Cotra (2021, 2022).

However, the lack of academic discussion has a price: first, arguments for extreme risks from AI are often not presented as explicitly and rigorously as would be necessary to convince critical readers. This makes it hard to satisfactorily evaluate the strength of these arguments. Second, since there is no canonical and detailed exposition of the argument for existential risk from AI, it is more difficult than necessary for critics to engage with these arguments. Critics lack both a clear target to engage with and requisite motivation, since developing detailed and careful objections to blogposts and forum posts is not incentivized in academia. As a result, I perceive the average quality of critical discussions of arguments for existential risk from AI to be low.

My goal is to remedy this unfortunate condition. In the following, I will argue for the conclusion that AI will lead to the permanent disempowerment (e.g. through extinction) of humanity by, at the latest, 2100. I aim to make the argument as explicit and convincing as I can manage. This serves two functions: first, a natural interpretation of my argument is probabilistic. Readers are encouraged to form their own beliefs on how likely the truth of various premises is. Since the argument is deductively valid, the probability of the joint truth of all premises is a lower bound on the probability of the conclusion. Thus, my argument will convince readers that the probability of permanent human disempowerment through AI is higher than they thought or, failing that, support readers to come up with estimates which better reflect their own beliefs. Second, since my argument aims to be as

✉ Leonard Dung
leonard.dung@fau.de

[1] Centre for Philosophy and AI Research, University Erlangen-Nürnberg, 91052 Erlangen, Germany

comprehensive as feasible while also focusing on the main concerns in the (partially non-academic) literature, it can serve as a clear target for future research. By reference to my argument, critics can easily express and defend their view by pointing to the premises of my argument, and the reasons given for them, which they reject. This will enable a more focused, critical and clear discussion of extreme risks from AI in the future.

To preview the overall structure of the argument, I will write down the complete argument here[1]:

The argument for near term human disempowerment through AI

P1: By 2100, humanity is capable of building AI capable of permanently disempowering humanity.

P2: If, by 2100, humanity is capable of building AI capable of permanently disempowering humanity, then AI capable of permanently disempowering humanity will be built by 2100.

P3: If AI capable of permanently disempowering humanity will be built by 2100, then misaligned AI capable of permanently disempowering humanity will be built by 2100.

P4: If misaligned AI capable of permanently disempowering humanity will be built by 2100, then misaligned AI which is capable of permanently disempowering humanity and tries to permanently disempower humanity will be built by 2100.

P5: If misaligned AI which is capable of permanently disempowering humanity and tries to permanently disempower humanity will be built by 2100, then humanity will be permanently disempowered by 2100.

Conclusion: Humanity will be permanently disempowered by 2100.

In the next five sections, I will in turn present, elucidate and motivate each of the five premises of my argument. While I hope to inspire further scrutiny and objections by others, I will also mention, discuss and largely reject some counterarguments myself. I will close by briefly reflecting on the significance of the argument's conclusion: that humanity will be permanently disempowered by 2100.

[1] The overall structure of this argument is similar to, albeit more general than, Carlsmith (2022). However, the reasoning in favor of the premises, and the discussion of putative objections, is much different (e.g., Carlsmith focuses on second-order evidence like surveys for P1, and his argument for P2 discusses only incentives). Also, Carlsmith's paper has been criticized for failing to make all the steps in his reasoning sufficiently explicit, and thus weakening the argument's clarity and persuasiveness (Thorstad 2023). I aim to remedy this putative weakness here.

2 Premise 1

2.1 Explaining premise 1

The first premise of the argument is as follows:

P1: By 2100, humanity is capable of building AI capable of permanently disempowering humanity.

What does this claim mean? Let's go through it, almost word by word. First, by the 'permanent disempowerment' of humanity I refer to any condition where, permanently, humanity is unable to determine its own future. An instance of permanent disempowerment is extinction. If no humans exist, then they cannot determine their own future. However, extinction is not the only possible form of disempowerment. Humans are disempowered by other humans if those other humans have control over them, determining what they can and cannot do. As a further analogy, humanity disempowered chimpanzees. Since humans are more powerful, the continued existence, as well as the living conditions, of chimpanzees depend on human goals. Moreover, chimpanzees have no, or very limited, influence on the goals and plans of humans. Humans form goals which are independent of chimpanzees' wishes. Since humans decide over the conditions of chimpanzee lives and chimpanzees have no influence over these decisions, they are disempowered.

In the form of non-extinction permanent disempowerment referred to in P1, all humans would be in a similar condition with respect to AI as chimpanzees are to humans. That is, AI systems would form goals or have formed goals independently of human wishes and they would decide over the living conditions of humans.[2] However, for reasons explained in Sect. 6, I think that human extinction is the most likely form of permanent disempowerment by AI. Henceforth, for the sake of brevity, I will often drop the qualifier 'permanent' when talking about human disempowerment.

I do not aim to provide an exact definition of 'AI'. There are many prototypical cases of AI systems: transformer models, reinforcement-learning agents, intelligent robots and classical expert systems. We can say that something is an AI if it is sufficiently similar to these and other prototypical cases. As a further elucidation, let us say that something is an AI if it has a sufficient ability to achieve goals or solve tasks (intelligence) and is not biological, i.e. not composed of cells and not the product of biological evolution by natural selection (artificial). However, my argument does not require an exact definition of what counts as an AI system.

[2] Disempowerment, as described here, requires that AI systems actually exert their power over humans to influence their lives, not just that they could do so if they wanted. For an ethical evaluation of the latter form of disempowerment, see Sparrow (2023).

By saying that AI is 'capable' of disempowering humanity I mean the following: if the AI would try to disempower humanity, it would succeed. This is a non-technical notion familiar from ordinary discourse. For instance, we may say that Magnus Carlsen is capable of beating me at chess. He wouldn't necessarily win against me, for he could lose intentionally. But he would win if he tried to. Of note is that the notion of 'capability' I use is quite strong. On a weaker notion, we might say that subject S is capable of achieving X if and only if, when S tries to achieve X, there is a *non-negligible probability* that S achieves X. However, in this paper, I want to explore the case for a strong conclusion: that AI *will* disempower humanity, not only that there is a non-negligible probability of human disempowerment through AI. Thus, I operate with a strong understanding of capability. On this notion, if an AI is capable of achieving X, X will occur if the AI tries to achieve X.

By saying that humanity is capable of building certain AI systems by 2100 I mean that such AI will be built by 2100 given that humanity tries to build such AI. Since humanity is an abstract entity, the notion of 'trying' is less perspicuous. What I mean is: If key actors (states, companies, individuals, etc.) devote a sufficiently high level of effort to building such AI, and key actors devote a sufficiently low level of effort to stopping the creation of such AI, such AI will be built.

Finally, choosing the year 2100 as a reference point for the argument is not obligatory. By choosing a later year, the argument becomes more plausible. However, by choosing an earlier year, the conclusion of the argument becomes more radical and practically important. Choosing the year 2100 achieves a pragmatic balance where the claim that humanity is capable of building sufficiently powerful AI up to this point is plausible while the conclusion implies that children born into the world today are likely to witness human disempowerment through AI in their lifetime.

In conjunction, P1 states that, by 2100, humanity will—if it tries—build AI which—if it tries—permanently disempowers humanity. In the next subsection, I will motivate this premise.

2.2 Motivating premise 1

P1 states that future AI may have more advanced capacities which suffice for being able to disempower humanity. To establish P1, it is not only necessary to show that humanity is capable of building sufficiently capable AI, but also that this is feasible by 2100. To motivate this claim, let us look first at the capacities AI might plausibly need to disempower humanity.

Traditionally, worries about human disempowerment through AI have often focused on so-called 'superintelligence' (Bostrom 2014). Define a 'cognitive capacity' as a mental competence contributing to the ability to achieve goals, broadly construed, such as memory, planning, reasoning, social cognition or mathematical cognition. For our purposes, we can take superintelligence to be AI which is, in all or virtually all cognitive capacities, vastly superior to humans. As an analogy, we can say that the difference in capacities between superintelligence and humans is analogous to the difference between humans and dogs. That is, the thoughts and plans of a superintelligence are beyond our comprehension.

In contrast, we can define AGI (artificial general intelligence) to be AI which is in all or virtually all cognitive capacities superior to humans. This does not require that it is superintelligent. Let PAI (powerful AI) be AI which, in virtue of its cognitive capacities, is superior to humans in some domains which grant significant power in today's world, e.g. scientific research, military strategy, engineering, hacking or social persuasion (Carlsmith 2022). The notion of PAI is logically weaker than AGI which is in turn weaker than superintelligence. Finally, let advanced AI be AI which is significantly more intelligent or powerful than current state-of-the-art AI.

To establish P1, I have to argue that some of these kinds of AI are capable of disempowering humanity and that humanity is capable of building them by 2100. I hold that even many kinds of PAI are capable of disempowering humanity. Plausibly, the reason why humans have disempowered all mammals is that humans have many superior cognitive capacities. Human control over the planet depends on their cognitive capacities. If a kind of being is sufficiently cognitively superior to humans and thus is superior in sufficiently many important domains, then it will, under normal conditions, be capable of disempowering humanity.

There are multiple further factors bolstering the case that PAI would disempower humanity. First, AI has certain intrinsic advantages to humans. Due to limitations of signal transmission in neurons, electrical signals can be transmitted in AI orders of magnitude faster (Luo 2018). AI does not get tired, as humans do. AI systems can communicate and thereby coordinate incessantly and almost instantaneously, while humans have to either be near each other or use devices like phones. Further, AI systems can be multiplied quickly. In the current paradigm, training cutting edge AI models requires a lot more compute than running a few models. For this reason, when there is enough compute available to train a PAI, there will likely also be enough compute to run hundreds of thousands or millions of copies of it (Davidson 2023; Steinhardt 2023). With further availability of compute, the number of AI system can rapidly increase from there. To summarize, we should expect that PAI systems would not only have cognitive advantages in some domains to humans, but process information much faster, function tirelessly, coordinate seamlessly and multiply quickly.

Second, PAI may— alone or in conjunction with humans— be able to innovate AI research and thus make the creation of even more powerful systems possible. This gives rise to a recursive process where smart AI is able to design smarter AI which is in turn able to design even smarter AI and so on. In the most extreme case, the ability of AI to perform AI research might lead to faster than exponential growth in AI capacities, perhaps even to an 'intelligence explosion' (Chalmers 2010; for objections see: Thorstad 2022).

Third, humans might voluntarily give AI resources and control over critical infrastructure. For instance, to increase their usefulness, many large language models (LLMs) have already been connected to the internet. If PAI were connected to the internet and tried to disempower humanity, this would likely make it impossible to simply 'turn it off', since it could distribute copies of itself over the internet. Moreover, for reasons of efficiency, there are calls to give AI access to military equipment. To summarize, both control over important pieces of infrastructure as well as the possibility of AI's improving themselves support the view that even many kinds of PAIs are capable of disempowering humanity. The more powerful AI is, the worse humanity's chances are of resisting attempted disempowerment. Finally, when talking about superintelligence, it is hard to conceive of scenarios where the superintelligence would not be able to disempower humanity.

Assuming that some kinds of PAI, typical AGI and virtually all superintelligences would be capable of disempowering humanity, why should we think that such AI can be built by 2100? First, I have to concede that predicting technological progress and scientific developments in advance is very hard, often impossible. Yet, this epistemic obstacle cuts both ways. While it should discourage us from being certain that PAI and AGI will arise in a particular time period, it should also prevent us from being extremely confident that PAI and AGI won't arise by 2100. Despite this uncertainty, I will now provide my reasoning for thinking that AGI will be built by 2100. This reasoning is based on the hypothesis that the current deep learning paradigm can, perhaps with some moderate adjustments, lead to AGI.

In essence, there are reasons to think that increasing the scale of current transformer models, like the most powerful current LLMs (e.g. GPT-4), suffices for AGI. The scale of the model is determined by the number of parameters comprising the model, the size of the dataset used to train the model and the amount of compute used for training. The first reason is that, in LLMs, scaling has led to massive increases in pre-existing capacities and the emergence of qualitatively new capacities (Wei et al. 2022). This development is often captured in terms of so-called scaling laws: It has been shown that language model performance improves smoothly as the scale of the model increases, where performance has a power–law relationship with each individual factor (model size, size of the training dataset or compute) when not bottlenecked by the other two (Kaplan et al. 2020). These scaling laws pertain to the task the model is directly trained to perform, i.e. predicting the next token in a sequence as accurately as possible.

Improvements in other related capacities cannot be predicted. Nevertheless, in the last five years, performance in virtually all interesting benchmark tasks has improved massively with increased scale (OpenAI 2023; Suzgun et al. 2022). This has been accompanied by the emergence of qualitatively new abilities, e.g. few-shot learning (Brown et al. 2020).[3] So, increases in scale have led to improvements in virtually all interesting abilities in LLMs, including the emergence of abilities which some thought were not available to them. Given this, it seems wise to suspect that further increases in scale will lead to further progress.

It could be that there are limits to what scaling can achieve and that progress will eventually stop. It could be that this limit comes before AGI. However, given that no one has identified an important capacity which does not improve with scaling, the currently best supported hypothesis is arguably that further scaling will bring about AGI. Further support for this idea is that, in terms of real-world usefulness and performance on benchmark tests, the difference in performance between GPT-4 and GPT-1 seems to be a lot higher than the difference between GPT-4 and human-level reasoning.[4] In the light of this, it would be quite coincidental if improvements from increasing scale would level of shortly before LLMs reach human-level intelligence.[5]

I hold that, if scaling transformer models suffices for AGI[6], sufficiently powerful models can be built before 2100. This conditional is supported by Cotra's Biological Anchors Framework (Cotra 2020). While I will not explain the entire framework, it is crucial here that it—based on

[3] In few-shot learning, the pre-trained model learns to perform new kinds of tasks after been given a few examples. This happens during the deployment phase, i.e. without further training of the model.

[4] The capacities of GPT-4 are especially impressive when considering that there is a lot of untapped potential for enhancing its performance. Better ways of fine-tuning or prompting the model are only starting to be explored (Dettmers et al. 2023; Yao et al. 2023).

[5] One could argue that we would expect LLM performance to level off around human intelligence levels because LLMs are trained on human text. However, it also seems that LLMs have various advantages to humans, e.g. faster processing and bigger (working) memory capacity. Thus, we would expect that—if the training data are the only constraint on capability improvements in LLMs due to scaling—LLM nevertheless learn to do superhuman reasoning with human training data. Moreover, LLMs are increasingly trained on multimodal data.

[6] This scaling assumption states that no foundational conceptual breakthrough is required for AGI, it does not say that exactly the same algorithms which are currently used suffice for AGI.

current trends—estimates how the availability of compute for training the largest deep neural networks will develop over time. The framework projects two trends out into the future:

(A) progress in both hardware and software that makes computing power cheaper
(B) economic growth, and an increasing importance of AI in the economy, which increases the amount that can be spent on training large AI models

Given these trends, Cotra estimates that the amount of compute (in FLOPs) which can be used to train frontier AI models will, from 2025 to 2100, increase roughly ten orders of magnitude. In other words, if current trends do not massively slow down, then astronomical scaling of models will be possible in the next decades. If scaling suffices for AGI, then AGI can be expected to be feasible by 2100.

It should be noted that this argument is tentative. Predicting technological and scientific advances is inherently very difficult. However, even if scaling up LLMs does not lead to AGI, there could be other breakthroughs which make AGI possible. After all, a long time will pass until 2100. As the beginning of the field of AI is often dated back to the year 1956, the time period from now to 2100 is longer than the entire existence of the field of AI so far. Given the progress that has been made in this shorter time span of 67 years, it seems wise to place a significant probability on the scenario that AGI can be developed until the end of this century. Since, given the previous argument, AGI would likely be capable of disempowering humanity, one should place a significant probability on the truth of P1.

Note further that the case for P1 does not depend on an "intelligence explosion" where progress in AI research enables AI to build more intelligent AI models, which enables them in turn to better design even more intelligent AI and so forth, in an ever-increasing speed (Chalmers 2010; Thorstad 2022). Superintelligence might as well arise via a slower and more continuous transition from less advanced forms of intelligence. More importantly, as I have argued, superintelligence is not necessary for the capacity to permanently disempower humanity. Some forms of PAI and many forms of AGI are plausibly capable of disempowering humanity. Thus, even without superintelligence, AI may be capable of disempowering humanity.

2.3 Objections to premise 1

In this sub-section, I will briefly discuss two objections to P1. Clearly, those objections can be developed further. Moreover, there are other warranted concerns researchers might have about premise 1. While I respond to the objections which strike me as most obvious and important, I aim to stimulate further discussion.

Since it seems clear to me that sufficiently intelligent AI systems (e.g., AGI) with some other feasible properties (e.g., being able to quickly multiply themselves and coordinate) could disempower humanity, I will focus on the assumption that humanity is capable of building such AI systems. First, there might occur large-scale economic disruptions which significantly undermine long-term global technological progress. For instance, some believe that an all-out nuclear war between great powers may destroy civilization as we know it. If technological progress in general would come to a halt or even be massively reversed, then humanity would not be capable of building more powerful AI. I accept the logic underlying this objection. Since estimating the chance of such large-scale economic disruptions is beyond the scope of the paper, I concede that this is a possibility. Thus, readers should downgrade their confidence in P1 depending on how likely they think such an economic disruption is.[7]

Second, some researchers deny that current deep learning methods will scale to AGI. They think that current deep neural networks (e.g. LLMs) lack (at least) one crucial ingredient for intelligence (Lake et al. 2017; Marcus 2018) which cannot be attained with more scale. There are many ideas for what LLMs might lack: for instance, generality (Dentella et al. 2023), compositional representation, causal and symbolic models of its environment (Chomsky 2023; Marcus 2022a, 2022b), perception and embodiment (Bender and Koller 2020) and the ability to learn from limited data (Chen et al. 2020). However, none of those views strikes me as compelling. Typically, versions of these capacities are even possessed by present-day LLMs[8].

Remarkable capacities for compositionality (Press et al. 2022; Yao et al. 2021) and domain-general reasoning (Bubeck et al. 2023; OpenAI 2023) are present in modern LLMs. With respect to learning speed, it is not clear how to fairly compare the amount of training data to the amount

[7] If the reader has no clear view on how likely significant disruptions stopping long-term economic progress are, they may interpret my argument—including the conclusion—as conditional on the absence of such disruptions.

[8] A special version of this objection is that AI systems are only capable of human disempowerment if they have agency and that current AI systems are not agents (AI myths 2024). I think that there is only a difference in degree—not in kind—between current AI systems and full agents, since behaviors like reward hacking that appear in current systems can plausibly lead to disempowerment, if the system is sufficiently powerful (see Dung 2023 and 2024 for some relevant discussion). Thus, a sufficient degree of agency might emerge as a result of increases in other abilities, particularly for long-term planning and reasoning. Alternatively, a proponent of the argument developed here may argue that there are incentives to intentionally create agentic AI systems (see, e.g., the creation of so-called language agents) (Chan et al. 2023).

of data humans have access to during ontogeny (Buckner 2021). Once trained, modern language models learn very quickly to perform new tasks (Brown et al. 2020). Due to the opacity of LLMs, it is uncertain whether they form representations of the world, including causal relations. However, there is preliminary evidence that LLMs do have something like a world model and causal understanding (Burns et al. 2022; Kıcıman et al. 2023; Li et al. 2023; Meng et al. 2023; though see: Thibodeau 2022). Finally, there are some initial successes in connecting LLMs to embodied agents (Dasgupta et al. 2023; Huang et al. 2022a, 2022b). While this brief survey is by no means the last word, I think the burden of proof is currently on the side of researchers arguing that LLMs won't scale to AGI to specify what precise capacities they lack and why those capacities are unlikely to emerge with increases in scale.

After having motivated and defended the view that humanity is capable of building AI capable of permanently disempowering humanity by 2100, I will now turn to premise 2.

3 Premise 2

3.1 Explaining and motivating premise 2

Premise 2 of the argument is as follows:

P2: If, by 2100, humanity is capable of building AI capable of permanently disempowering humanity, then AI capable of permanently disempowering humanity will be built by 2100.

P2 is easy to explain. The antecedent of the conditional is P1, i.e., the view that humanity will, if it tries, build AI by 2100 which, if it tries, permanently disempowers humanity. The consequent is the claim that AI which will, if it tries, succeed at disempowering humanity will actually be built by 2100. P2 follows logically if one assumes that humanity will actually try to build sufficiently powerful AI, i.e., humanity will devote sufficiently many resources to building such AI and insufficiently many resources to stopping the creation of it. Why would one think that humanity will try to build AI which is capable of permanently disempowering humanity?

I think there are two main reasons. First, there are strong incentives to pursue progress in cutting-edge AI models. AI which automates many, or perhaps eventually all, human tasks has immense financial value. The potential is almost limitless. Since automation via AI will be vastly more efficient than relying only on human labor, military competition will provide a further strong incentive to pursue research on cutting-edge AI. Moreover, these incentives are not just hypothetical. There already are big corporations like Google and Microsoft which have poured enormous sums into AI (Ahmed et al. 2023). The same goes for governments. Thus, there are concrete actors which have a strong interest that progress in AI research continues unabated. In other words, there are strong commercial and geopolitical incentives to push forward with more AI research and those incentives are already entrenched in today's economy and politics.

Thus, even if no actor aims at "building AI which is capable of disempowering humanity" (under this description) there are incentives to try to build AI which is as powerful as possible. Given the previous section, I expect that such AI would in fact be able to disempower humanity.

Second, with sufficient time (e.g., by 2100), it looks like sufficiently powerful AI could be built by many different entities. First, there are not only different companies but also different states which might want to build AGI. Some of them, e.g. the US and China, are in an adversarial political relationship. Within these countries, we are currently in a situation where only a few selected companies have the ability to build systems at the level of the current cutting-edge. But this will probably change. Let's look at why.

Progress in AI mainly depends on two factors: First, hardware improvements. As explained earlier, deep neural networks tend to improve with increases in scale. Thus, when the upscaling of models becomes feasible and computing power becomes cheaper, AI becomes better. Other AI paradigms also benefit from availability of more compute and better hardware. It enables AI to store more information and plan ahead further and researchers to perform more experiments. Second, new scientific insights lead to AI progress. In particular, new ideas can lead to algorithmic improvements so that, holding hardware constant, AI becomes more capable.

The important point here is that both algorithmic and hardware improvements could, over time, position many entities so as to be capable of bringing about AGI. If one entity, e.g. a leading AI lab, has the capacity to build AGI, then—according to the typical trajectory of technological development and deployment—not much later many entities, including governments, companies and at some point ordinary citizens, can build AGI. This is because both factors mentioned earlier—new ideas and hardware improvements—tend (with a delay) to diffuse throughout society and the world.

New algorithms are hard to keep secret, even when investing a lot of resources. Moreover, it is not clear whether secrecy will be given much weight. Compute gets cheaper over time, roughly along an exponential curve.[9] Today's supercomputer will be outmatched in computing power by tomorrow's smartphone, so to speak. Thus, over the typical

[9] As evidence of this trend in the decline of computing costs, see: https://ourworldindata.org/grapher/historical-cost-of-computer-memory-and-storage (Accessed 29th of May 2023).

course of technological development, many actors gain access to algorithmic improvements and increasing compute. Hence, after some time and without dedicated countermeasures, many actors will be able to build AGI once someone has the ability. Moreover, if substantial algorithmic improvements turn out to be unnecessary for AGI, then—even if dedicated AI research were to be stopped—hardware improvements would nevertheless make AGI possible at some point.

The more entities can independently build AGI, the higher the chance that some reckless, malevolent, or ignorant actors will do so. If we reach a situation where most countries and many small groups of citizens could independently build AGI, even a tight regime of political coordination and surveillance of citizens cannot plausibly prevent that someone at some point will decide to build AGI. To summarize, the combination of strong incentives and the likely prospect that, if some entities can build AGI by 2100, then many can build AGI by 2100, makes it very likely that humanity will try to build AI which is capable of permanently disempowering humanity.

3.2 Objections to premise 2

I will discuss three objections. Again, I do not provide definite refutations of all conceivable objections, but my counterarguments serve as discussion starters for why obvious objections might not work. The first objection is that international coordination to prohibit the development or deployment of a technology does sometimes work. An example may be the regulation of nuclear weapons. The technology is available and there are incentives favoring their use, but they are—globally—not deployed. This has been achieved through international agreements. However, on second thought, the analogy between nuclear weapons and AI doesn't necessarily inspire confidence. Large-scale nuclear weapons have been used in the past, even if it was a long time ago. Several times they would have almost been used, with potentially much more destructive effects. Furthermore, the number of countries possessing nuclear weapons has increased over the years, despite international agreements prescribing non-proliferation.

Moreover, there are several reasons to think that stopping the development and deployment of powerful AI is harder than with nuclear weapons. Nuclear weapons can only be built by entities, typically states, with a lot of resources, and building them requires big nuclear projects. The constrained number of actors as well as the possibility to monitor the buildup makes it easier to coordinate around and enforce bans on developing and using nuclear weapons than AI. Besides, there are weaker incentives to develop and use nuclear weapons. Nuclear weapons cannot contribute meaningfully to economic growth. Few individual people or companies have a stake in the development of nuclear weapons. Finally, the risk from using nuclear weapons is obvious. Unlike with AI, there is no disagreement on whether nuclear weapons are dangerous. All these factors make agreeing on and enforcing a ban on nuclear weapons and most or all other technologies easier than a ban on advanced AI.

The second objection to P2 is that it is in everyone's interest to abstain from building AGI if building AGI poses an enormous risk of permanent human disempowerment (as I argue in this paper). For virtually no one wants to be killed or otherwise disempowered by AI. Thus, there are no incentives to build AGI. However, this reasoning does not strike me as persuasive. There are many points which could be made in response, but I will focus on two: the risk of human disempowerment through AI in general is neither obvious nor certain. Moreover, there can be many reasonable disagreements on whether particular AI systems pose a threat of human disempowerment, partly because the emergence of new capacities is often unpredictable. Thus, some people may underestimate or completely ignore the risk from AI in general or from specific systems. Given that we can expect many entities to be able to build AGI, it is likely that some of them will gravely underestimate the risks. In this case, they might build more powerful AI even if it is not in their interest.

The third objection to P2 is that an AI which is advanced, but not yet able to disempower humanity may try to disempower humanity. If so, it may cause massive damage without disempowering humanity and thus serve as a warning sign. Due to this warning, humanity may then be motivated to stop the development of even more advanced AI. While this scenario is possible, it does not address the two factors—incentives and multiple actors—which strongly support P2. Moreover, it is not clear whether such a warning will occur. Partially, this depends on the speed and especially on the continuity of AI progress. The more discontinuous AI progress is, the less likely it is that a useful warning will take place. In addition, an AI which is capable of causing massive damage might be capable of reasoning that it should not try to disempower humanity, until it is confident that it will succeed. If so, advanced AI might not give us a warning sign.

In the next section, I will explain, motivate and defend premise 3 of my argument.

4 Premise 3

4.1 Explaining and motivating premise 3

Premise 3 is as follows:

P3: If AI capable of permanently disempowering humanity will be built by 2100, then misaligned AI capable of permanently disempowering humanity will be built by 2100.

Thus, P3 states that, given the antecedent of the conditional, at least some of the AI systems built by 2100, which are capable of disempowering humanity, will be misaligned. What does 'misalignment' mean? As defined here, an AI model is aligned if and only if it tries to do what its designers want it to do (Ngo et al. 2022).[10] In other words, an aligned model pursues goals, has values or optimizes an objective function which correspond to the goals, values or objective function its designers want it to have. An AI model is misaligned if and only if it is not aligned.

Misalignment can be understood in analogy with humans. Suppose I ask Magnus Carlsen to beat someone in chess. Suppose he loses intentionally. In this case, Carlsen was misaligned with my goals. He didn't try to do what I wanted him to do. If AI is misaligned, it will tend to not do what we want it to do. In at least one respect, misaligned AI is not under our control. Given these clarifications, P3 states that, given the antecedent of the conditional and by 2100, some AI systems capable of disempowering humanity will pursue goals which differ from what their designers want them to be.

Why should we believe in P3? The key reason is that ensuring alignment of AI models, particularly advanced models, is difficult. First, it seems that superior capacity to achieve goals places few constraints on what the goals of AI models are. A model can be very intelligent in the sense of being excellent at achieving its goals for almost any set of goals.[11] There is no guarantee that the goals will strike humans as reasonable or ethical. However, intelligence construed as the capacity to achieve goals is what matters to the question whether an AI is capable of disempowering humanity. This shows that it is at least possible that AI capable of disempowering humanity has misaligned goals. Why is misalignment not only possible, but likely?

There are two kinds of challenges in aligning AI. First, one needs to specify the goals the model should pursue. Second, one needs to ensure that the model robustly pursues those goals.[12] The first challenge has been termed the 'king Midas problem' (Russell 2019). In a nutshell, human goals are complex, multi-faceted, diverse, wide-ranging, and potentially inconsistent. This is why it is exceedingly hard, if not impossible, to explicitly specify everything humans tend to care about. An explicit list is bound to forget some important goals which will then be neglected by AI.

Moreover, even if the goals are specified adequately, it is hard to ensure that the model follows them robustly. With ML models, while alignment can be tested during training, there is the worry that the model may fail to pursue these goals when it encounters situations too distant from the training corpus. Since the number of situations general models can encounter is (virtually) unbounded (e.g., there are no limits on the kinds of texts which can be input to LLMs), it is hard to guarantee that the model will generalize its goals appropriately across all situations.

These alignment challenges are not just hypothetical, but manifest in current ML. In reinforcement learning, it is notoriously difficult to specify the reward function the model is trained to optimize such that, by maximizing reward, the model learns to perform the task it is intended to perform (Pan et al. 2022). "Reward hacking" is a common failure mode (Skalse et al. 2022). Even when the model pursues the intended goal in one context, the model sometimes fails to generalize the goal to another context, outside of its training distribution (Langosco et al. 2023). Even in current models, it is hard to predict in advance whether and how they will be misaligned, and to remedy misalignment. Alignment can often only be achieved after a long process of trial-and-error.

With more advanced models, further problems arise. In this case, it is even harder to predict, or even detect, whether models are misaligned (Dung 2023). For instance, an AGI may be instructed to write papers producing novel scientific insights. If AGI is superior to us in scientific reasoning, it may be hard for us to evaluate whether it aims to increase scientific knowledge, as we intended, or whether it comes up with deliberate falsities.

Moreover, there arise two crucial problems with detecting and remedying misalignment which are specific to advanced models (Dung 2023). If PAI is misaligned to such an extent that its goal is to disempower humanity (more on why to expect this next section), then it has reasons to pretend to be aligned so that humans won't take countermeasures. To disempower humanity, it has an incentive to behave in a seemingly aligned manner until it is ready to actually disempower humanity. Thus, with advanced models, we have to expect that they may try to deceive us.

Even more crucially, with AI which is capable of disempowering humanity, a trial-and-error process to ensure alignment is not an option. If misaligned AGI has the goal to disempower humanity (see next section), then there is no second chance to align it. For once humanity is disempowered, we cannot align an adversarial AGI anymore. Thus, we would have to succeed at alignment at the first try. This is contrary to current methods of AI alignment which are based on observing specific alignment failures, and then modifying the system in response.

[10] There is another facet of the alignment problem which consists in the question of which values an AI system should be aligned with, and who gets to decide these values (Gabriel 2020). Since the technical alignment problem is central to the risk of human disempowerment from AI, not this further, more ethically loaded alignment problem, I focus on the former here.

[11] This is an expression of the so-called orthogonality thesis according to which (instrumental) intelligence and goals are orthogonal (Bostrom 2012, 2014).

[12] Some approaches to alignment are based on only partially fulfilling both of these tasks, however (Hubinger 2021).

In fact, the problem is even harder. As I noted last section, it is likely that, by 2100, there are many entities, e.g. states and companies, which are able to build PAI and AGI. For P3 to be true, only one of those AI models, which is capable of disempowering humanity, has to be misaligned. Thus, when developing sufficiently powerful AI, every one of those entities has to succeed at aligning their models on the first try. Furthermore, with sufficiently many actors able to build AGI, we should even expect that some will build and deploy AGI with dangerous goals malevolently. This suggests that an alternative argument for existential risk from AGI, not the focus here, may proceed via the risk of misuse, rather than misalignment. Roughly, the argument may go: If sufficiently capable AI systems will be build and many people can build or control one of them, then some individual or group is going to use it to usurp control of humanity (see Friederich 2023 for concerns of this kind).

To summarize, since potentially many actors can develop AGI and aligning AI models, as well as predicting whether models will be aligned, is hard, and can be expected to be even harder with more advanced models, we should expect that misaligned AGI will be built, if AGI will be built.

4.2 Objections to premise 3

In this section, I will discuss two objections to P3. As always, my discussion is not exhaustive and is intended to stimulate further critical scrutiny. In particular, there is a field of technical AI alignment research (Ngo et al. 2022) which aims to find methods to ensure that AI is aligned, including more advanced models. Someone might argue that their own distinctive approach to alignment will be successful and will be adopted and that, thus, P3 is implausible. Due to constraints of space, I cannot discuss various different approaches to AI alignment. Thus, I will not engage with this type of objection.

My sense, however, is that most researchers in the AI alignment field believe that we are far from possessing a detailed proposal for aligning PAI that we can be sufficiently confident in. Moreover, I note that, given the possibility that many AI models capable of disempowering humanity might be build, a successful alignment approach would need to be adopted and implemented correctly without exception to make P3 false. This is a high bar.

The first objection to P3 rests on the idea that intelligence and goals (values) are connected. According to this view, a being which is superior to humans in intelligence, e.g. an AGI, must have— or strongly tends to have—reasonable or ethical goals (Müller and Cannon 2022; Railton 2020). Proponents of this objection also adduce the observation that, in humans, learning of moral reasoning and learning of other cognitive abilities are interdependent.

However, observations regarding human moral learning are irrelevant. The claim presupposed by P3 is not that AGI might lack the capacity for moral reasoning. For all we know, AGI might necessarily be better than humans in moral reasoning, as manifested in the capacity to, e.g., write ethics papers. The claim is not that AGI would lack any *ability*, but that it might lack ethical *goals*. While being able to reason ethically, AGI would not have any motivation to act ethically if that is not its goal. While open to debate, the claim that goals and abilities to achieve goals are largely independent in AI is supported by current ML approaches. In reinforcement learning, for instance, a model can be rewarded to obtain arbitrary results. Consequently, the model can become very proficient at bringing about these results, fulfilling its goals, even if the goals themselves are "stupid" or parochial (e.g., some arbitrary metric in a videogame) (for more thorough defenses of the thesis that a superintelligence could have bad goals, see Häggström (2021) and Dung (forthcoming)).

The second objection is that the technical challenge of aligning AI with human goals will turn out to be quite trivial. Given the previous discussion, I am not sure how one can be very confident in this view. Nevertheless, something to be said in favor of the view that alignment might not be as difficult as I describe it is that the task of aligning AI capable of disempowering humanity might be continuous with aligning less powerful systems. If so, we may be able to learn how to align PAI from systems which are not capable of disempowering humanity. If lessons learned from studying AI not capable of disempowering humanity generalize well to AI which is capable of disempowering humanity, then we can learn about AI alignment empirically, by trial-and-error. If so, humanity's odds for learning how to align AGI are better.

Moreover, one might hope that this aligned AGI can then be used to prevent other misaligned PAI models from coming into existence or, failing that, prevent them from disempowering humanity. The aligned AGI model might be used to help align other AI models, to monitor the development and deployment of other AI models in order to intervene when a dangerous model might be built, and to defeat a misaligned model which tries to disempower humanity.

I do not deny that this scenario is possible. However, the scenario depends on several speculative and optimistic assumptions. First, aligning PAI would need to be relatively easy. Second, solutions to the problem of aligning less dangerous AI models would need to generalize very well to AI capable of disempowering humanity, such that no experience and experimentation with those more dangerous models is necessary to apply these alignment solutions. Third, the first actor to build AI capable of disempowering humanity is responsible and cares about AI alignment. Fourth, one of the first AI models capable of disempowering humanity is also able to stop other actors from building other AI models capable of disempowering humanity or, failing that,

stop these models from disempowering humanity. While I hope that all these conditions obtain, one cannot take this for granted and it seems relatively unlikely.

To summarize, while there is hope that the alignment problem is less hard than it seems, it is likely that the problem is very difficult. Thus, P3 seems to be true. In the next section, I move to premise 4 of my argument.

5 Premise 4

5.1 Explaining and motivating premise 4

Let me introduce premise 4:

P4: If misaligned AI capable of permanently disempowering humanity will be built by 2100, then misaligned AI which is capable of permanently disempowering humanity and tries to permanently disempower humanity will be built by 2100.

P4 fills a crucial gap in the argument. So far, I have argued that misaligned AI capable of disempowering humanity will be built by 2100. Yet, there are as many ways for an AI to be misaligned as there are goals the AI could have. P4 states that, if misaligned AI capable of disempowering humanity will be built, AI with the same dangerous capability which actually tries to disempower humanity will be built. The underlying rationale is that many, or all, misaligned AI models capable of disempowering humanity will have the goal of disempowering humanity. Why should we believe that this is true?

The canonical answer in the literature is the instrumental convergence thesis (Bostrom 2012, 2014). The instrumental convergence thesis states that there are certain goals which are instrumentally useful for a wide range of final goals and a wide range of situations (Bostrom 2014, p. 109). Among these goals are self-preservation and the accumulation of power and resources. For if you get destroyed, you cannot work towards achieving your final goal anymore, and if you acquire power and resources, you will be more effective at achieving your final goal. The instrumental convergence thesis suggests that, typically, you will increase the probability that your final goal will be satisfied by preserving yourself and by accumulating power. Thus, for a wide range of goals, AGI would have an incentive to acquire power and to resist being shut down.

To consider an example, suppose an AGI model has an arbitrary misaligned goal, e.g. maximizing the number of paperclips in the universe.[13] This is the model's final goal, i.e., what it optimizes for. Given this final goal, it would be beneficial to the model to acquire power and resources which it can use to produce more paperclips, to prevent being shutdown (which would stop it from making more paperclips) etc. Importantly, if power-seeking is a convergent instrumental goal, then disempowering humanity is. By usurping control, AI makes sure that humans cannot shut it down or otherwise interfere with its goals. This reasoning translates to virtually every other misaligned goal.

Up until now, I have argued that disempowering humanity is a convergent instrumental goal. That is, for a wide range of final goals, disempowering humanity would be useful for PAI and AGI. I claim that this provides us with reason to think that AI capable of disempowering humanity will in fact try to disempower humanity. The claim is: the usefulness of disempowering humanity suggests that AGI will try to disempower humanity, given that AGI is a powerful optimizer. If AI optimizes for a particular set of final goals, then it will try to achieve any goal instrumentally useful for it if it believes[14] that the value of the instrumental goal outweighs the expected costs of pursuing the instrumental goal, measured according to the final goal. This follows from the fact that AI *optimizes* for its goal. This condition is fulfilled if and only if (1) the AI believes it is capable of achieving the instrumental goal, (2) the AI believes that achieving the instrumental goal would be useful and (3) the expected costs of pursuing the instrumental goal are relatively low.

Since AI capable of disempowering humanity would be very smart, (1) and (2) should be fulfilled with respect to the instrumental goal of disempowering humanity. Disempowering humanity would be very useful for a wide range of goals, since it rules out a likely source of interference with its goals and provides access to enormous resources. Thus, it has high expected benefits. At the same time, the expected costs are typically low. For instance, disempowering humanity only limits one's ability to, e.g., build paperclips insofar as one could spend the resources spent on disempowering humanity instead on paperclips. However, given that a sufficiently capable AI might not need many resources to disempower humanity and that the benefits of doing so are big, the expected value of trying to disempower humanity is likely positive. Hence, as soon as misaligned AI capable of disempowering humanity is sufficiently certain that it will succeed, it will try to disempower humanity.

For illustration, we can also see the logic of instrumental convergence at work in human relationships to other species. It is not the case that humans have a final goal or desire that gorilla populations be decimated or that many gorillas are locked up in zoos. However, given human aims such as having space to live, entertaining children etc., disempowering

[13] It seems to me that the same reasoning applies to more complex final goals or a set of final goals, if they are misaligned.

[14] If preferred, one can speak here of 'representations', 'belief-like states' or something similar instead of beliefs.

gorillas was in humans' instrumental interest which is why it occurred.

Moreover, I want to note that instrumental convergence is not the only route to AI capable of disempowering humanity which tries to disempower humanity. If sufficiently many actors will be able to build AI capable of disempowering humanity, including, e.g. small groups of ordinary citizens, then some will intentionally unleash AI trying to disempower humanity. This could happen as a form of (omnicidal) terrorist attack, "for fun" because the people involved underestimate the risk, or in the hopes of taking control over all other humans. This is the main reason why my argument provides only a lower bound to the probability of permanent human disempowerment from AI.

To summarize, most misaligned AI models capable of disempowering humanity will have strong instrumental reasons to try to disempower humanity. Thus, we should expect that they will try to disempower humanity. Moreover, some malevolent or careless humans might intentionally build AI capable of and trying to disempower humanity.

5.2 Objections to premise 4

In this sub-section, I will discuss one objection to P4. I hope that future research will thoroughly scrutinize this premise and develop new arguments, objections and counterarguments. The objection begins with the observation that the logic of instrumental convergence also applies to humans. Nevertheless, humans do not typically try to disempower the rest of humanity. Therefore, instrumental convergence does not seem to imply (massive) power-seeking.

But there are differences to the AGI case. First, humans are not capable of disempowering humanity. Plausibly, some humans would indeed try to disempower humanity if they were capable of it (Carlsmith 2022, Sect. 4.2). However, some humans would not. How do we explain this? Recall that P4 only speaks of misaligned AI. Similarly, humans will not try to disempower humanity if one of their final goals is the wellbeing of the rest of humanity, i.e., if they are "aligned" with the rest of humanity, so to speak. Hence, we can explain that some humans capable of it would not try to disempower humanity by pointing out that they are aligned with the interests of humanity. Thus, a straightforward argument against P4 based on the absence of massive power-seeking in humans does not succeed.

However, this objection draws our attention to something deeper. In humans, it might not be possible to point to a clear boundary between final and instrumental goals. It is not clear where the final goals end and the instrumental goals begin. Instead, it often seems more appropriate to conceive of human motivational psychology as *messy*. On this picture, the human mind comprises a mixture of many, partially competing and partially complementing, drives and psychological dispositions. Any number of these dispositions and drives could disincentivize humans from trying to disempower humanity. For instance, some humans just don't like being the center of attention.

If we picture AGI minds analogously, then it is less clear that they will try to disempower humanity. Any number of features of their motivational or decision-making system could make them reluctant to try to disempower humanity (Thorstad 2023). A clean inference that they will exhibit massive power-seeking cannot be made.

However, I have some reservations with this argument for the claim that misaligned AI will not try to disempower humanity. First, if AGI minds are messy, then this makes predicting what a misaligned AGI model's goals would be quite impossible. This is no strong reason to think that those goals would end up being relatively human-friendly. In particular, of all the logically possible goals for AGI, only a small fraction involves human wellbeing. Moreover, since instrumental convergence is something we find in humans, although constrained by other dispositions, maybe instrumental convergence would win out against other psychological dispositions in AGI and thus create the goal to disempower humanity.

Second, even if human psychology is messy, this does not mean that an AGI's psychology would be messy. It seems like current deep learning methodology embodies a distinction between final and instrumental goals. For instance, in standard versions of reinforcement learning, the model learns to optimize an externally specified reward function as best as possible. It seems like this reward function determines the model's final goal. During training, the model learns to seek out things which are instrumentally relevant to this final goal. Hence, there appears to be a strict distinction between the final goal (specified by the reward function) and instrumental goals.

Consequently, there is some reason to think that methods continuous with current deep learning approaches lead to the development of strict optimizers for a set of final goals. If so, misalignment of AGI will likely lead to AGI trying to disempower humanity. That being said, there is also significant uncertainty about the cognitive processing in PAI and AGI. Thus, there is a chance that the model will acquire psychological tendencies which constrain power-seeking. In the next section, I will describe the fifth premise and give an overview of the entire argument.

6 Finalizing the argument: premise 5 and conclusion

Premise 5 of my argument is as follows:

P5: If misaligned AI which is capable of permanently disempowering humanity and tries to permanently disempower

humanity will be built by 2100, then humanity will be permanently disempowered by 2100.

This premise is very close to a tautology. I defined 'being capable of X' as meaning 'If one tries X, then X'. Thus, tautologically, if misaligned AI is capable of permanently disempowering humanity and tries to disempower humanity, then humanity will be disempowered. The only qualification which saves P5 from being a mere tautology is the reference to time. Conceptually, it is possible that AI capable of and trying to disempower humanity is built before 2100 but only succeeds at disempowering humanity after 2100. Due to this possible scenario, it is conceptually possible for P5 to be false.

However, this scenario does not strike me as plausible. The reasons adduced in favor of P1 support the claim that sufficiently powerful AI will likely be built early enough before 2100, and be capable of disempowering humanity quickly enough, that humanity will be disempowered by 2100. If it nevertheless turns out that humanity is disempowered quickly after 2100, then this technically undermines the letter of my argument, but is not relevant to the spirit and the broader significance of the argument.

Now, we have assembled all premises which compose my argument. To get a second look at the argument in its entirety, the reader may consult the introduction again. It is easy to confirm that this argument is logically valid. In the earlier sections, I have argued for the truth of each premise. If all premises are true, then humanity will be permanently disempowered by 2100.

It is hard to overstate the significance of this conclusion. The conclusion entails that most children born today will witness humanity's disempowerment through AI. The form of human disempowerment which seems most likely to me is extinction. As argued earlier, if AI optimizes for some misaligned goal, it will have instrumental reasons to amass as many resources as possible. AI could plausibly use the resources which humans need to survive (e.g., space) for its own goals. Moreover, killing all humans seems like an effective and simple way to prevent humans from interfering with its goals. By contrast, it is hard to come up with a plausible reason why sufficiently advanced AI (which does not rely on human labor) should try to keep disempowered humans alive. Therefore, I think that—if the path to human disempowerment proceeds via misaligned AI achieving its instrumental goals—extinction is the most likely form of permanent human disempowerment. Thus, if the argument of this paper is sound, one of the most likely causes of death for children born today is being killed by misaligned AI. This provides strong prudential reasons to try to reduce the risk of human disempowerment from AI.

Ethically, human disempowerment through AI means that roughly 8 billion people will either die or be harmed in another very significant way. Moreover, it means that future human beings which could otherwise have led flourishing lives will either not be born or be powerless subjects to the goals of AI. If the future would have otherwise contained a very large number of lives and if future lives have significant value, then this destruction of the future potential of humanity may be much worse than even 8 billion deaths (Bostrom 2013; Greaves and MacAskill 2021; MacAskill 2022).

I mentioned earlier that a natural interpretation of the argument is probabilistic. The combined probability of all the premises places a lower bound on the probability of the conclusion. If one follows my argument, the probability of the conclusion, i.e. permanent human disempowerment by 2100, is higher than many might have thought. This provides a reason for relevant entities—individuals, corporations and governments—to aim to reduce the probability of human disempowerment through AI. Many levers to influence this probability have already been touched upon in the argument: technical research on AI alignment, research on how to best regulate AI and attempts to increase international political cooperation on AI assume a central role.

At last, this paper is also an invitation to skeptics of my conclusion to come forward and explain which of my premises they disagree with, and why. This way, I hope to stimulate a fruitful, rigorous, and constructive discussion on the question whether there is a significant risk of human disempowerment through AI, particularly in the not-too-distant future.

Acknowledgements I thank Jakob Lohmar, Markus Over, Marvin Riemer and David Thorstad for helpful comments and advice. Special thanks go to Jakob Lohmar for inspiring me to write this paper in the first place.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was funded by the German ministry for education and research (BMBF) in the context of the K3I-Cycling project (Project number: 033KI216).

Availability of data and materials Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest The author states that there is no potential conflict of interest.

References

Ahmed N, Wahed M, Thompson NC (2023) The growing influence of industry in AI research. Science 379(6635):884–886. https://doi.org/10.1126/science.ade2420

AI Myths (2024) Myth: AI has agency. https://www.aimyths.org/ai-has-agency#Common-arguments-in-AI-Agency-Debates. Accessed 5 Mar 2024

Bender EM, Koller A (2020) Climbing towards NLU: on meaning, form, and understanding in the age of data. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Presented at the ACL 2020. Association for Computational Linguistics, pp 5185–5198. Doi: https://doi.org/10.18653/v1/2020.acl-main.463

Bostrom N (2012) The superintelligent will: motivation and instrumental rationality in advanced artificial agents. Mind Mach 22(2):71–85. https://doi.org/10.1007/s11023-012-9281-3

Bostrom N (2013) Existential risk prevention as global priority: existential risk prevention as global priority. Global Pol 4(1):15–31. https://doi.org/10.1111/1758-5899.12002

Bostrom N (2014) Superintelligence. Paths, dangers, strategies. Oxford University Press, Oxford

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al (2020) Language models are few-shot learners. arXiv: https://doi.org/10.48550/arXiv.2005.14165

Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al (2023) Sparks of artificial general intelligence: early experiments with GPT-4. arXiv: https://doi.org/10.48550/arXiv.2303.12712

Buckner CJ (2021) Black boxes, or unflattering mirrors? Comparative bias in the science of machine behavior. Br J Philos Sci. Doi: https://doi.org/10.1086/714960

Burns C, Ye H, Klein D, Steinhardt J (2022) Discovering latent knowledge in language models without supervision. arXiv: https://doi.org/10.48550/arXiv.2212.03827

Carlsmith J (2022) Is power-seeking AI an existential risk? arXiv: https://doi.org/10.48550/arXiv.2206.13353

Center for AI Safety (2023) Statement on AI risk. https://www.safe.ai/statement-on-ai-risk. Accessed 20 June 2023

Chalmers DJ (2010) The singularity: a philosophical analysis. J Conscious Stud 17(9–10):9–10

Chan A, Salganik R, Markelius A, Pang C, Rajkumar N, Krasheninnikov D et al (2023) Harms from increasingly agentic algorithmic systems. https://doi.org/10.48550/arXiv.2302.10329

Chen Z, Eavani H, Chen W, Liu Y, Wang WY (2020) Few-shot NLG with pre-trained language model. arXiv. https://doi.org/10.48550/arXiv.1904.09521

Chomsky N (2023) The false promise of ChatGPT—The New York Times. New York Times. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html. Accessed 22 Mar 2023

Cotra A (2020) Draft report on AI timelines. https://www.lesswrong.com/posts/KrJfoZzpSDpnrv9va/draft-report-on-ai-timelines. Accessed 25 May 2023

Cotra A (2021) Why AI alignment could be hard with modern deep learning. Cold takes. https://www.cold-takes.com/why-ai-alignment-could-be-hard-with-modern-deep-learning/. Accessed 15 Jan 2023

Cotra A (2022) Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover. Lesswrong. https://www.lesswrong.com/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to. Accessed 3 Feb 2023

Dasgupta I, Kaeser-Chen C, Marino K, Ahuja A, Babayan S, Hill F, Fergus R (2023) Collaborating with language models for embodied reasoning. arXiv: https://doi.org/10.48550/arXiv.2302.00763

Davidson T (2023) Part 1—what a compute-centric framework says about takeoff speeds: section 2–5 + appendices. Google Docs. https://docs.google.com/document/d/1rw1pTbLi2brrEP0DcsZMAVhlKp6TKGKNUSFRkkdP_hs/edit?usp=embed_facebook. Accessed 20 Jun 2023

Dentella V, Murphy E, Marcus G, Leivada E (2023) Testing AI performance on less frequent aspects of language reveals insensitivity to underlying meaning. arXiv: https://doi.org/10.48550/arXiv.2302.12313

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023, May 23). QLoRA: Efficient Finetuning of Quantized LLMs. arXiv. https://doi.org/10.48550/arXiv.2305.14314

Dung L (2023) Current cases of AI misalignment and their implications for future risks. Synthese 202(5):138. https://doi.org/10.1007/s11229-023-04367-0

Dung L (2024) Understanding artificial agency. The Philos Q pqae010. https://doi.org/10.1093/pq/pqae010

Dung L (forthcoming) Is superintelligence necessarily moral? Analysis. https://doi.org/10.1093/analys/anae033

Friederich S (2023) Symbiosis, not alignment, as the goal for liberal democracies in the transition to artificial general intelligence. AI Ethics. https://doi.org/10.1007/s43681-023-00268-7

Gabriel I (2020) Artificial intelligence, values, and alignment. Mind Mach 30(3):411–437. https://doi.org/10.1007/s11023-020-09539-2

Greaves H, MacAskill W (2021) The case for strong longtermism. https://globalprioritiesinstitute.org/wp-content/uploads/The-Case-for-Strong-Longtermism-GPI-Working-Paper-June-2021-2-2.pdf. Accessed 1 Jun 2022

Häggström O (2021) AI, orthogonality and the Muller-Cannon instrumental vs general intelligence distinction. arXiv. https://doi.org/10.48550/arXiv.2109.07911

Huang W, Abbeel P, Pathak D, Mordatch I (2022) Language models as zero-shot planners: extracting actionable knowledge for embodied agents. In: Proceedings of the 39th International Conference on Machine Learning. Presented at the International Conference on Machine Learning, PMLR, pp 9118–9147. https://proceedings.mlr.press/v162/huang22a.html. Accessed 22 Mar 2023

Huang W, Xia F, Xiao T, Chan H, Liang J, Florence P et al (2022) Inner monologue: embodied reasoning through planning with language models. arXiv: https://doi.org/10.48550/arXiv.2207.05608

Hubinger E (2021) How do we become confident in the safety of a machine learning system? https://www.alignmentforum.org/posts/FDJnZt8Ks2djouQTZ/how-do-we-become-confident-in-the-safety-of-a-machine. Accessed 10 Aug 2023

Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, et al (2020) Scaling laws for neural language models. arXiv: https://doi.org/10.48550/arXiv.2001.08361

Kıcıman E, Ness R, Sharma A, Tan C (2023) Causal reasoning and large language models: opening a new frontier for causality. arXiv: https://doi.org/10.48550/arXiv.2305.00050

Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ (2017) Building machines that learn and think like people. Behav Brain Sci 40:e253. https://doi.org/10.1017/S0140525X16001837

Langosco L, Koch J, Sharkey L, Pfau J, Orseau L, Krueger D (2023) Goal misgeneralization in deep reinforcement learning. arXiv: https://doi.org/10.48550/arXiv.2105.14111

Li K, Hopkins AK, Bau D, Viégas F, Pfister H, Wattenberg M (2023) Emergent world representations: exploring a sequence model

trained on a synthetic task. https://doi.org/10.48550/arXiv.2210.13382
Luo L (2018) Why is the human brain so efficient? Nautilus. https://nautil.us/why-is-the-human-brain-so-efficient-237042/. Accessed 29 May 2023
MacAskill W (2022) What we owe the future. Oneworld Publications, London
Marcus G (2018) Deep learning: a critical appraisal. arXiv: https://doi.org/10.48550/arXiv.1801.00631
Marcus G (2022a) Deep Learning Is Hitting a Wall. *Nautilus*. https://nautil.us/deep-learning-is-hitting-a-wall-238440/. Accessed 22 Mar 2023
Marcus G (2022b) What does it mean when an AI fails? A reply to SlateStarCodex's riff on Gary Marcus. The Road to AI We Can Trust. Substack newsletter. https://garymarcus.substack.com/p/what-does-it-mean-when-an-ai-fails. Accessed 22 Mar 2023
Meng K, Bau D, Andonian A, Belinkov Y (2023) Locating and editing factual associations in GPT. arXiv: https://doi.org/10.48550/arXiv.2202.05262
Müller VC, Cannon M (2022) Existential risk from AI and orthogonality: can we have it both ways? Ratio 35(1):25–36. https://doi.org/10.1111/rati.12320
Ngo R, Chan L, Mindermann S (2022) The alignment problem from a deep learning perspective. arXiv: http://arxiv.org/abs/2209.00626. Accessed 14 Jan 2023
OpenAI (2023) GPT-4 technical report. arXiv: https://doi.org/10.48550/arXiv.2303.08774
Pan A, Bhatia K, Steinhardt J (2022) The effects of reward misspecification: mapping and mitigating misaligned models. arXiv. https://doi.org/10.48550/arXiv.2201.03544
Press O, Zhang M, Min S, Schmidt L, Smith NA, Lewis M (2022) Measuring and narrowing the compositionality gap in language models. arXiv: https://doi.org/10.48550/arXiv.2210.03350
Railton P (2020) Ethical learning, natural and artificial. In: Liao SM (ed) Ethics of artificial intelligence, pp 45–78. Oxford University Press, Oxford. Doi: https://doi.org/10.1093/oso/9780190905033.003.0002
Russell S (2019) Human compatible: artificial intelligence and the problem of control. Viking Press, New York
Skalse J, Howe NHR, Krasheninnikov D, Krueger D (2022) Defining and characterizing reward hacking. arXiv: https://doi.org/10.48550/arXiv.2209.13085
Sparrow R (2023) Friendly AI will still be our master. Or, why we should not want to be the pets of super-intelligent computers. AI and Society. Doi: https://doi.org/10.1007/s00146-023-01698-x
Steinhardt J (2023) What will GPT-2030 look like? Lesswrong. https://www.lesswrong.com/posts/WZXqNYbJhtidjRXSi/what-will-gpt-2030-look-like. Accessed 20 Jun 2023
Suzgun M, Scales N, Schärli N, Gehrmann S, Tay Y, Chung HW et al (2022) Challenging BIG-bench tasks and whether chain-of-thought can solve them. arXiv: https://doi.org/10.48550/arXiv.2210.09261
Thibodeau J (2022) But is it really in Rome? An investigation of the ROME model editing technique. Alignment Forum. https://www.alignmentforum.org/posts/QL7J9wmS6W2fWpofd/but-is-it-really-in-rome-an-investigation-of-the-rome-model. Accessed 22 Mar 2023
Thorstad D (2022) Against the singularity hypothesis. Global Priorities Institute. https://globalprioritiesinstitute.org/against-the-singularity-hypothesis-david-thorstad/. Accessed 19 Jan 2023
Thorstad D (2023) Exaggerating the risks (Part 7: Carlsmith on instrumental convergence). Reflective altruism. https://ineffectivealtruismblog.com/2023/05/06/exaggerating-the-risks-part-7-carlsmith-on-instrumental-convergence/. Accessed 29 May 2023
Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al (2022) Emergent abilities of large language models. arXiv. https://doi.org/10.48550/arXiv.2206.07682
Yao H, Chen Y, Ye Q, Jin X, Ren X (2021) Refining language models with compositional explanations. In: Advances in neural information processing systems, vol 34, pp 8954–8967. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2021/hash/4b26dc4663ccf960c8538d595d0a1d3a-Abstract.html. Accessed 22 Mar 2023
Yao S, Yu D, Zhao J, Shafran I, Griffiths TL, Cao Y, Narasimhan K (2023) Tree of thoughts: deliberate problem solving with large language models. arXiv: https://doi.org/10.48550/arXiv.2305.10601