AI takeover and human disempowerment

Adam Bales (Global Priorities Institute, University of Oxford)

Global Priorities Institute | April 2024

GPI Working Paper No . 9-2024

Please cite this working paper as: Bales, A . AI takeover and human disempowerment. *Global Priorities Institute Working Paper Series, No. 9-2024*. Available at: https://globalprioritiesinstitute.org/AI-takeover-and-human-disempowerment-a-bales

AI Takeover and Human Disempowerment[1]

Adam Bales

Abstract

Some take seriously the possibility of AI takeover, where AI systems seize power in a way that leads to human disempowerment. Assessing the likelihood of takeover requires answering empirical questions about the future of AI technologies and the context in which AI will operate. In many cases, philosophers are poorly placed to answer these questions. However, some prior questions are more amenable to philosophical techniques. What does it mean to speak of AI empowerment and human disempowerment? And what empirical claims must hold for the former to lead to the latter? In this paper, I address these questions, providing foundations for further evaluation of the likelihood of takeover.

Keywords: Artificial Intelligence · AI Takeover · Human Disempowerment · Power Over · Power To

[1] For feedback, thanks to Tomi Francis, Andreas Mogensen, Oliver Ritchie, Brad Saad, Christian Tarsney, Teru Thomas, Elliot Thornley, Hayden Wilkinson, and Timothy Luke Williamson, as well as reviewers for the journal.

1 *Introduction*

Some worry that artificial intelligence (AI) might disempower humanity. Such disempowerment could take various forms. For example, a group of humans might use AI systems to seize power for themselves, thereby disempowering all other humans (Hendrycks et al. 2023: §2.4). Or the deployment of AI systems in evermore contexts might make the world so complex that humans struggle to understand it and meaningfully shape what happens. This involves a form of disempowerment. These possibilities strike me as worthy of reflection, but in this paper, I'll focus on another potential route to human disempowerment: *AI takeover*, where this involves AI systems actively seizing power and disempowering humanity in doing so (cf. Carlsmith 2021; Cotra 2022).

Much of the discussion of such takeover has focused on establishing that AI systems might, in the near future, seize vast amounts of power. Yet this isn't sufficient to make the case for human disempowerment; it would also need to be shown that AI empowerment would lead in turn to human disempowerment. This further claim is comparatively neglected. In this paper, I'll redress this neglect.

In doing so, I won't aim to decisively resolve the matter. After all, in predicting AI's impacts, many of the issues are empirical in a way that precludes resolution from the philosopher's armchair. Still, in this paper, I'll carry out two tasks that are amenable to philosophy's toolkit. First, drawing on work on the nature of power, I'll distinguish different senses in which AI takeover could bring about humanity's disempowerment. I'll then outline the case for thinking that AI empowerment—of the sort involved in takeover—might bring about each form of human disempowerment, identifying key empirical claims that are relied upon in making this case.

Ultimately, I'll conclude that there are various ways AI empowerment might lead to human disempowerment. However, I'll also show that the most radical forms of human disempowerment require extremely strong forms of AI empowerment. As the case for expecting such empowerment is speculative, more work would be needed to establish that radical human disempowerment was likely.

Note that my focus will be on the case for expecting various forms of human disempowerment. I'll mostly set aside the further questions of to what extent these forms of disempowerment are undesirable and whether some cases of human disempowerment might even be desirable.[2] Consequently, this paper is intended to be just a part of a broader process of reflection on the possibility of AI bringing about human disempowerment.

2 AI Empowerment

When I speak of AI systems acquiring power, I'll follow the existing literature on takeover by using *power* to mean roughly, 'the type of thing that helps a wide variety of agents pursue a wide variety of objectives' (Carlsmith 2021).[3] For example, acquiring money is one way to acquire power in this sense, as money can be used in pursuit of a range of goals. Likewise for developing and controlling technologies that can be used for a range of purposes (like computing technologies and technologies for energy generation).

In this paper, I'll mostly take it as an assumption that AI systems will seek and acquire substantial power and will ask whether human disempowerment would follow.[4] Nevertheless, while I won't provide a thorough argument for AI empowerment, I'll briefly outline why some take this possibility seriously (cf. Bostrom 2014; Carlsmith 2021).

[2] To get a taste of the issues: a case where a cruel, human dictator loses control of society is different to one involving a loss of democratic human control.

[3] I'll clarify this notion in §4.

[4] Many are sceptical about AI empowerment. For example, LeCun—one of the recipients of the 2018 Turing Award who are often called the godfathers of AI—describes existential threat from AI as, 'preposterously ridiculous' (Taylor 2023). However, others take this threat seriously, including Bengio and Hinton, the other two recipients of the 2018 award (Bengio et al. 2023).

My own guess is that AI won't insatiably seek power, partly because I'm sceptical of aspects of the below argument and partly because I expect we'll act to mitigate the risks. Still, I think it's worth reflecting on AI empowerment to gain more certainty about the risks and support better mitigation.

As a starting point, note that there's been rapid recent progress in AI. This progress might continue, and two potential feedback loops could lead it to accelerate.[5] First, in a technical feedback loop, AI-powered tools could help us to develop more sophisticated AI systems, which could help us to develop better AI tools, and so on (Drexler 2019: 18–19).[6] Second, in an economic feedback loop, if AI becomes highly profitable this could lead to more funding for R&D, and so to better systems and hence more profit, and so on.[7] Consequently, it's possible that the coming decades could see radical improvements in AI.

How radical? In some domains, AI systems already dramatically exceed human capabilities, including in playing various games (Silver et al. 2016; Silver et al. 2017; Badia et al. 2020), predicting protein structure (Jumper et al. 2021), and rapidly generating high-quality images and text from written prompts (as seen in language models and text-to-image models).[8] Given continuing progress, AI might come to outperform humans in increasingly many domains.

In addition to becoming more capable, AI systems will plausibly become more numerous. AI systems are already widely deployed, either in practice or in prototypes: they drive cars, trade stocks, operate military drones, act as personal assistants, and so on. Over time, AI is likely to become increasingly pervasive. Eventually, these systems might outnumber humans if they're widely embedded in phones, computers, TVs and cars.[9]

At this point, AI systems would be more capable than humans in many domains, as well as more numerous, and would be embedded in military and economic systems, among other places.

[5] Increased computing power has driven much recent progress. This suggests we might achieve further progress even without fundamental breakthroughs, by utilising additional computing power (though there's debate about how far this can take us). See Sutton 2019; Hoffmann et al. 2022; Lohn & Musser 2022.

[6] An extreme version of this is the *singularity*, which involves recursive improvement leading rapidly from human-level AI to radically superhuman AI (see Chalmers 2010; Bostrom 2014: ch. 4; Thorstad Unpublished). However, I focus on a broader range of possibilities, including mundane ways that AI systems could play a role in AI R&D, perhaps by increasing the efficiency of human coders or helping discover new algorithms. Relatedly, I'm not assuming these feedback cycles must involve AI systems operating independently of human input.

[7] Diminishing returns could undermine these feedback loops, so accelerating progress isn't inevitable. Still, it's a possibility.

[8] AI systems that seem superhuman can fail in ways humans never would (cf. Wang et al. 2023). Still, these systems plausibly exceed human capabilities all-things-considered.

[9] It's unclear how to individuate, and hence count, AI systems, but for my purposes it's unimportant whether many systems might operate simultaneously, or whether it's instead many tokens or threads of systems that might do so.

Given this, AI systems might be able to seize further power if they attempted to, at least if they could coordinate.[10]

The argument for why these systems would make such an attempt appeals to two claims. The first claim: some systems will be goal-driven, selecting actions on the basis that these promote some goal.[11] We might deliberately design systems with this structure—because it's useful to be able to point AI at goals—or it might emerge from a system's training process (Ngo et al. 2023: §3.2). The second claim: given that power-seeking behaviours are instrumentally useful for achieving a variety of goals—by the above definition of power—we should expect these behaviours to emerge in many goal-driven systems (Bostrom 2012; Carlsmith 2021: §4.2). For example, money is plausibly helpful for achieving a wide variety of goals, and so plausibly many goal-driven systems will attempt to acquire money. Further, because more money is typically useful for achieving goals to a greater extent, systems might seek to acquire huge levels

[10] This might involve individual AI systems that are each more capable than humans across many domains (as might happen if labs—like OpenAI and Google DeepMind—achieve their aim of creating generally capable agents). Alternatively, it might involve different systems, each more capable than humans in different domains, cooperating to collectively seize power. Whether we should expect such cooperation depends, among other things, on whether AI systems are capable of the reasoning required for cooperation, whether humans have things to offer AI alliances, and whether AI systems have better mechanisms for cooperation with one another than with humans. I'll consider related issues in §5.2.

Insofar as AI empowerment is most plausible if future systems are generally capable, scepticism might arise. Landgreve and Smith (2023) have argued that such artificial general intelligence is impossible, and even setting this aside, one might doubt that such systems are on the horizon. However, while this scepticism might eventually be vindicated, it's far from settled consensus. Rapaport (unpublished) argues that Landgreve and Smith's results show only what's possible with current techniques and rely on disputable assumptions about what's needed to achieve general intelligence. Meanwhile, many experts take seriously the possibility of generally capable systems being developed within decades (Roser 2023).

My own view echoes that of Colombo and Piccinini (2023: 44), who note Landgreve and Smith's arguments, but conclude that, 'we do not understand general intelligence well enough, or how to reproduce it in machines, to be confident about whether and when artificial general intelligence will be achieved'. Domain general systems remain a live possibility.

[11] One might deny this, arguing that to be goal driven an AI system must 'give a damn', in Haugeland's (1979) sense, which requires being embodied and embedded in a social world. Assuming AI systems cannot be embodied or embedded it follows that they cannot be goal driven. I owe this argument to a referee, but for related discussions, see Adams and Browning 2017 and Chemero 2023.

I think this challenge should shift our views but doubt it's decisive. While current systems might not give a damn, Chemero (2023: 1829) notes that, 'future AI models… might be built so that they do give a damn'. I agree: it's unclear whether future systems will give a damn or whether the limitation is fundamental. Further, it's unclear whether being goal driven in the sense relevant to my discussion requires systems to give a damn in Haugeland's sense. Arguably, other ways that AI can be goal driven suffice for worries about AI empowerment (see Ngo, Chan, & Mindermann 2023: §3).

of wealth rather than being easily satiated. Much the same argument applies to other ways of seeking power, so AI systems might generally seek substantial quantities of power.[12]

For these reasons, some people take seriously the possibility that AI systems will be goal-driven and will seek substantial power. Because of numerosity, capability, and the fact that we'll have handed these systems a great deal of power, these people also take seriously the possibility that this power seeking will be successful.

As a final step, it's argued that ensuring that an AI system has desirable goals requires overcoming unsolved technical problems. Consequently, absent solutions to these problems, we shouldn't expect systems' goals to involve respect for human flourishing and shouldn't expect these goals to preclude behaviour that harms humans (Krakovna et al. 2020; Shah et al. 2022; Ngo et al. 2023).[13] So, some deny that we can simply assume that the power acquired by AI will be used in ways that are desirable from the human perspective. If so then it's worth asking, as this paper does, what the consequences of AI power seeking would be.[14]

Of course, this brief discussion hardly makes an undeniable case for AI empowerment. Indeed, even the more detailed discussions of Bostrom and Carlsmith are very far from decisive. Nevertheless, I take the above to give a sense of why AI empowerment is worth taking seriously enough to ask what would follow if this did, in fact, occur. So in the following, I'll assume that AI systems will acquire substantial power.

[12] Later, when discussing broadly-scoped goals, I'll consider further whether AI systems will insatiably seek power.

[13] One could deny this. Perhaps we'll have enough (imperfect) control of these goals to ensure minimal respect for humans. Or, perhaps patterns of reward during training will lead systems to develop deontological-style restrictions on the means they use in pursuing goals, in ways that preclude human harm. Or perhaps any system capable enough to seize power will be capable of reflecting on reasons, in a way that will lead it to care for human flourishing (Müller & Cannon, 2021). Or perhaps we wouldn't deploy systems that failed to heed human flourishing and so won't see such systems in use (for pushback, see Bostrom 2014: 142–145; Carlsmith 2021: §5; Cotra 2022; Ngo et al. 2023: §4.2). Or perhaps recent advances involving language models make it easier to provide desirable goals to AI systems (Goldstein & Kirk-Giannini forthcoming). While these possibilities call for consideration in the full course of reflection, I'll set them aside here.

[14] As noted above, I won't evaluate whether all forms of disempowerment are undesirable. Still, at various points, I'll focus on forms of disempowerment where there's a *prima facie* case for concern, as these call for particular reflection.

3 Human Disempowerment

In the remainder of the paper, I'll explore whether AI empowerment is likely to lead in turn to human disempowerment. However, it's first necessary to clarify what I mean when I speak of such disempowerment. In particular, rather than discussing a single notion of human disempowerment, I'll discuss three, exploring the case for thinking that AI empowerment will lead to each.

First, one notion of power is *power-over*, where in rough terms, one person has power over another to the extent that they're able to get that other person to undertake actions they wouldn't otherwise undertake (see Dahl 1957). When one person has power over another, that other person isn't necessarily disempowered. Consider Allen's (1998) example of an honourable basketball coach, who exerts power over the team—deciding on training regimes, who will fill which position, and so on—but does so with the consent of the team in order to hone their skills, help them bond, and promote their careers and wellbeing. It would mischaracterise this situation to say that the coach disempowers the team. Still, while power-over doesn't always disempower, when one person has *illegitimate* power over another this does constitute a form of disempowerment for the other person, which I'll call domination.[15] So in the context of AI, one form of AI empowerment might involve AI systems having illegitimate power over humans. Correspondingly, one form of potential human disempowerment is *domination*.

As defined above, power-over, and hence domination, requires the mere ability to control human action, regardless of whether this ability is exercised (see Pettit 1997). In the context of AI, it has been argued that we should want to avoid domination even in this merely-potential

[15] Three clarifications. First, I'll mostly remain neutral on what makes power-over illegitimate, though one factor plausibly relates to whether this power could be exercised contra the interests of those the power is held over or without being responsive to those interests (Pettit 1997; Allen 1998: 34). Second, some might deny that the mere presence of illegitimate power-over *constitutes* a form of disempowerment. Such sceptics could ignore my discussions of domination and focus on the other forms of disempowerment. Third, while I'll largely avoid evaluating the badness of disempowerment, I'll sometimes use thick concepts like domination (in discussing disempowerment, avoiding thick concepts would be needlessly constraining).

sense (Sparrow forthcoming). Still, as Sparrow notes, there's a strong case for expecting human disempowerment in this potential sense if, as I'm assuming in this paper, AI systems become sufficiently capable and numerous, are handed control of crucial military and economic systems, and use this control to acquire wealth, develop advanced technology, and seize political power. Under these circumstances, AI systems could plausibly exert control over humans if they wished to. Further, such systems would plausibly have the capacity to exert this control *illegitimately*, insofar as their capabilities would allow them to avoid any pressure to account for the interests of humans.[16] So these circumstances plausibly involve domination in the potential sense, even if the systems never in fact exert control.[17] So I take the matter to be relatively clear insofar as we focus on domination as mere potential. The case where further reflection seems valuable relates to whether AI empowerment will lead to illegitimate power over humans actively being exercised. Consequently, I'll focus on active domination.[18]

A second notion of power is *power-to*, where a person has power-to to the extent that they have the capacity to undertake a broad range of actions (Pitkin 1972; Morriss 2002). A particularly important form of power-to involves the capacity to construct a flourishing life, and it's this capacity I'll focus on.[19] So in the context of AI, one form of disempowerment might involve AI systems undermining people's power to acquire food and shelter, construct meaningful lives, form social connections, and so on. This would interfere with people's capacity to create desirable day-to-day lives. I'll call this sort of disempowerment *incapacitation.*

[16] On the relationship between interests and legitimacy, see note 15 and Sparrow forthcoming.

[17] While such a case would *plausibly* involve domination in the potential sense, this isn't a certainty. After all, a group of AI systems could develop institutional structures that limit members' capacity to exert power over humans, either to preclude such power or ensure its legitimacy. This might happen if some AI systems respect human flourishing while others don't. The systems concerned with human flourishing might develop safeguards for humanity. Still, given my earlier assumption that AI systems will be unconcerned by human flourishing, domination in the potential sense seems likely.

[18] In doing so, I'm not taking any position on Pettit or Sparrow's views. I'm merely focusing on the case where further work seems most helpful.

[19] See note 15 for an explanation of how this relates to my setting aside of evaluative matters.

Finally, much of the discussion on AI takeover focuses centrally on the possibility that humanity, as a species, will lose whatever influence we would otherwise have over the decisions that shape civilisation (Karnofsky 2022; Cotra 2022). This would mean that AI systems, rather than humans, play the key role in determining how society progresses over time. It would preclude such influence being wielded by democratic consultation of humanity, by the dictatorial whims of a single human, or by any other process in which humans played a central role. I'll call this sort of disempowerment *disenfranchisement*.[20] Humanity is disenfranchised to the extent that humans lack influence over civilisation's shape.[21]

Disenfranchisement is related to concerns about human extinction, which have animated much of the discussion about extreme risks from AI.[22] From one direction, disenfranchisement could increase the risk of human extinction. Disenfranchisement involves civilisation being shaped by AI systems that, per my earlier assumption, lack concern for human flourishing. This removes a safeguard that protects against extinction: the fact that many decisions are currently made by people who care about humanity's continuing existence.[23] From the other direction, human extinction would plausibly result in disenfranchisement: if humanity went extinct then humans would be unlikely to shape society's further evolution (nonexistence typically precludes influence).[24] Still, while disenfranchisement and extinction are linked, my interest here is in human disempowerment, so I'll focus on disenfranchisement and set extinction aside.

[20] Disenfranchisement is related to a loss of *power-with*, our ability to act with others to achieve shared ends (Allen 1998). One might also analyse disenfranchisement in terms of loss of power-over and power-to at the societal level.

[21] One way to exercise power is to influence someone's preferences and beliefs so that these align with your interests (Lukes 2005: §1.4). This suggests a form of AI power whereby AI systems shape human preferences and beliefs so as to indirectly influence civilisation. While this isn't disenfranchisement, it raises similar issues, and much of my discussion will apply to this case.

[22] See for example https://www.safe.ai/statement-on-ai-risk. Works in academic philosophy often focus not on extinction but on existential risk, which includes extinction but also other events that 'permanently and drastically destroy [Earth-originating intelligent life's] potential for future desirable development.' (Bostrom 2014: 115). See also Ord 2020.

[23] An AI system that's unconcerned by human flourishing might care instrumentally about human. I'll consider this possibility later. For now, I note that lack of more direct concern for humans plausibly increases extinction risk.

[24] This is plausible rather than inevitable because humans might make a plan that AI systems enact in our absence. In a sense, this would involve humanity directing civilisation.

So in this paper, I'll explore whether AI empowerment would be likely to lead to human domination, incapacitation, and disenfranchisement.

4 The Dual Nature of Power

A simple argument that AI empowerment would lead to human disempowerment could appeal to the claim that power has a dual nature, in the sense that whenever one party is empowered another is disempowered in turn.[25] If so then AI empowerment would mean disempowerment for others, plausibly including humans.[26] However, as noted above, power can mean various things. So two questions arise: what forms of power, if any, have a dual nature and what forms of power might AI systems acquire?

As to the first question, domination has a dual nature: whenever one dominates, by possessing illegitimate power-over another, that other is dominated and hence disempowered. On the other hand, power-to doesn't have a dual nature: empowerment in this sense needn't mean disempowerment for others. Indeed, sometimes when one person acquires power-to, others also gain power-to. For example, if a researcher develops a cure for cancer then the resulting wealth might give them power-to but improved health outcomes might also give others power-to. The researcher's acquisition of power-to doesn't require disempowering others. So power-over has a dual nature but power-to does not.

To see what sort of power future AI systems might acquire, on the above picture, we need to consider an element of the discussion from §2 in more detail: the claim that power-seeking behaviours are instrumentally useful for achieving a wide variety of goals (and so AI systems are

[25] A more specific claim would hold that power is zero-sum. However, this framing would raise complications. For example, if zero-sum means that the gains to one are lost by another then we have to consider whether the thing lost in disempowerment is the same as the thing gained in empowerment. And if we define zero-sum in quantitative terms, as meaning that gains and losses sum to zero, then we need to quantify gains and losses and show that they reduce to a common measure that can be meaningfully summed. To avoid these complications, I focus on the claim that power has a dual nature.

[26] Even if power has a dual nature, AI empowerment doesn't immediately imply human disempowerment as AI empowerment could come at the cost of animals or aliens. Still, if widespread AI empowerment leads to widespread disempowerment for others then humans will likely be among those disempowered.

likely to seek power). This claim, which is sometimes called the *instrumental convergence thesis*, is most fundamentally a claim about power-to: it's a claim that a wide variety of goals incentivise developing the power-to achieve things in the world. This interpretation is bolstered by paying attention to Carlsmith's (2021) definition of power, which I quoted earlier, which characterises power as, 'the type of thing that helps a wide variety of agents pursue a wide variety of objectives'. This is clearly a notion of the power that an agent has *to* achieve things, rather than a notion of the power that an agent possesses *over* others.

Now we can spell out more carefully the argument that this section began with: power-over has a dual nature; AI systems will acquire power-to; therefore AI empowerment will lead to human disempowerment. Spelled out this way, it's clear that the argument is invalid and that the original version traded on an equivocation between two notions of power.

Still, while the simple form of this argument fails, its failure is informative: it points us towards more sophisticated arguments for human disempowerment. We might develop such arguments by showing either: (a) that AI will plausibly dominate humans; or (b) that AI acquiring power-to would plausibly lead to human disempowerment. Each of these would support an argument for human disempowerment.

5 Domination

Domination has a dual nature. Consequently, if it could be shown that AI systems will dominate humans then we would have a case for human disempowerment, at least in the sense of domination. So the first purpose of this section will be to clarify the circumstances under which we should expect AI to dominate humans. The second purpose will be to clarify what AI power-over would entail for incapacitation and disenfranchisement.

5.1 AI Power-Over & Domination

As a starting point, I'll outline two reasons AI systems might acquire power over humans, setting aside for a moment the question of whether this power-over would be illegitimate and so constitute domination.

First, power over humans could allow an AI system to make humans carry out tasks that promoted its goals. In other words, the control of human intellectual and physical labour involved in power-over can be a form of power-to: control of this labour can make it easier to achieve things. Second, power over humans could help ensure that humans didn't threaten AI power-to (Bostrom 2014: 141). After all, if humans were acting freely then we might be tempted to shut the system down or to build a competing system that could threaten it. In this respect, the AI system doesn't seek power-over as a positive means to power-to but rather to preclude threats to power-to that it has otherwise acquired.

So insofar as AI systems seek power-to they might also seek power over humans. And, in this paper, I'm assuming both that AI systems will seek power-to and that they will acquire power when they seek it. Putting all of this together, we have an argument that AI systems might acquire power over humans, as part of their broader quest to acquire power-to. If so, what would follow in terms of human disempowerment?

Domination. AI power over humans wouldn't necessarily involve domination. After all, as with the honourable basketball coach, this could be a case of legitimate power-over. However, this will be a hard line to defend if we accept a view, which Allen points to, on which power over someone involves domination unless it's 'exercised for their benefit' (Allen 1998: 34). In the case being considered, AI systems are exerting power-over in pursuit of their own interests, rather than for the benefit of humans (remembering my assumption, from §2, that AI goals won't involve respect for human flourishing). So AI power over humans would plausibly mean domination of humans.

That said, Allen also points to a view on which the exercise of power-over doesn't involve domination if it benefits a person, regardless of whether it's exercised *for* this person's benefit.[27] On this view, certain forms of AI power over humans might not involve domination. For example, perhaps there's no domination involved if an AI system exerts power over humans by paying for human labour. If so then the case for human domination would depend on the claim that AI systems will pursue power over in ways that aren't beneficial for humans. This might be the case if the most effective way for AI systems to utilise human labour and preclude human threat is via force. I'll consider this possibility in §5.2 so for now simply note that if force is particularly effective then domination of humans seems likely to result from AI power over humans.

Incapacitation. AI power over humans could plausibly impact human power-to: if AI systems can control human action this could easily undermine human capacity to construct desirable lives. Still, the extent to which human power-to is threatened here depends on how the AI systems exercise power over.

As above, an AI system could direct human labour by becoming wealthy and employing humans. This needn't dramatically curtail humans' power to construct desirable day-to-day lives, and indeed such employment could potentially promote human power of this sort. Further, precluding humanity as a threat might simply require restricting our access to a small number of dangerous technologies, in much the way that governments already restrict individual access to some technologies. So AI power over doesn't immediately imply substantial human incapacitation.

Still, there might be stronger grounds for worry if some AI systems will have *broadly-scoped goals*, where these apply to long timeframes and large scales (Ngo et al. 2023: 6). For example, a

[27] Allen defines domination in terms of constraint on another's choice that 'works to the others' disadvantage' (Allen 1998: 34). Contra a natural reading of this line, my own view is that cases where power-over benefits someone, but where the exercise isn't for their benefit, involve domination (it's simply that the harm of domination must be weighed against other benefits).

broadly-scoped goal for a stock-trading system might be to make money, with more always being better and no constraints being placed on when or how this goal is achieved. In contrast, a narrowly-scoped goal for such a system might be to give advice, within the next minute, on whether a particular stock trade should be made. If AI systems have sufficiently broadly scoped goals, and if they optimise heavily to achieve these to the greatest extent possible, then AI systems have a sort of unbounded ambition. This might push them to extract as much work as possible from humans, which is likely to impinge more severely on our day-to-day lives. It might also push them to see even minor human interference as a threat, insofar as it slightly decreases the extent to which the goal is achieved. The systems might then enact repressive measures to preclude minor interference.[28]

While Ngo et al. discuss some reasons for expecting AI systems to develop broadly-scoped goals, it strikes me as an open empirical question whether they'll ultimately do so. I cannot aspire to settle such questions here. Still, insofar as systems do develop broadly-scoped goals, it becomes more plausible that their gaining power over humans will lead to large-scale human incapacitation.

Disenfranchisement. As to disenfranchisement, the issues here plausibly also turn on whether AI systems will have broadly-scoped goals. After all, if these systems merely have narrowly-scoped goals then they'll have little concern for controlling the broad shape of civilisation. Such control would be unnecessary for achieving relatively constrained goals, and seizing control of the entirety of civilisation is unlikely to be necessary to stop humanity threatening the achievement of some constrained goal.[29] So even if AI systems had power over

[28] If many AI systems with narrowly-scoped goals each used their power over humans in minor ways, this might cumulatively lead to incapacitation. Further, many systems pursuing narrowly-scoped goals in a coordinated fashion might sometimes act equivalently to a single system pursuing a broadly-scoped goal. This raises the possibility of incapacitation resulting from narrowly-scoped goals. However, modelling the emergent dynamics arising from multi-system interactions is a complex task that would require its own paper. Here, I set this possibility aside.

[29] In discussing broadly-scoped goals, I've emphasized spatial and temporal scale. Another factor is whether the system cares about miniscule increases to the probability of goal achievement (Bostrom 2014: 151–153; Ngo et al. 2023: fn. 14). As I use the phrase, systems with narrowly-scoped goals won't be concerned about miniscule

humans, if they had narrowly-scoped goals they'd be unlikely to exercise this in a way that brought about substantial disenfranchisement.[30] On the other hand, if a system has goals that stretch far into the future and are large-scale enough to require vast resources to achieve then it will more plausibly be incentivised to use its power over humans to shape civilisation. After all, it'll want civilisation's resources to be focused on achieving its large-scale goal to the greatest extent possible. So given broadly-scoped goals, and sufficient power over humans, large-scale disenfranchisement becomes more plausible.

In summary, one way human disempowerment from AI might result is if AI systems acquire power over humans in order to direct our labour and to preclude threat from humans. This could potentially lead to domination of humans. Further, if AI systems have broadly-scoped goals then this domination could potentially lead to both incapacitation and disenfranchisement.

5.2 Two Caveats

With respect to the above, two caveats are worth highlighting.

The first caveat: the above case for disempowerment relies on a relatively fine balancing act when it comes to AI capabilities. If those capabilities are sufficiently weak then it's unlikely that AI systems would acquire power over humans and so concerns about human disempowerment are ameliorated. On the other hand, if those capabilities are sufficiently advanced then humans would likely pose little threat to AI systems and these systems would likely have little to gain from human labour (because they could carry out tasks themselves more competently). If so, AI systems would have little incentive to develop power over humans, and

influence, and so won't be incentivised to seize control of civilisation to slightly increase the probability of goal achievement.

[30] There are at least two ways to push back here. First, if it's sufficiently easy for AI systems to seize control of civilisation then doing so might be a good strategy even for achieving narrowly-scoped goals. For example, if a coalition of systems could easily seize control then these systems might sign a binding contract agreeing to seize control and use it to promote the goals of all in the coalition. Signing the contract might be the easiest way to achieve even a narrowly-scoped goal. Still, this relies on strong assumptions about AI empowerment and coordination. Second, the actions of many systems with narrowly-scoped goals might cumulatively lead to human disenfranchisement, without the systems aiming at this end. See note 28.

the case for human disempowerment would be weakened.[31] Consequently, if power-over is to lead to human disempowerment then AI capabilities will plausibly need to be neither too strong nor too weak.[32]

One reason to accept this assumption would be the thought that AI capabilities will progress relatively continuously, so that at some point we'll pass through what we could call the Goldilocks zone, where capabilities are at the needed level. If so then at some point we should expect human disempowerment to result from the above mechanisms.[33]

On the other hand, one might deny that the required zone exists: perhaps any system with the capacity to exert substantial power over humans will be capable enough that it'll have little to gain from our labour and little to fear from our anger. If so, the required assumption might never be satisfied.[34] Whether AI power-over leads to human disempowerment depends, in part, on which of these views is right.

A second caveat: the above case for human disempowerment proceeds most straightforwardly if the most effective way for AI systems to utilise human labour, and preclude human threat, is via force. And this seems far from inevitable.

Consider human nations: one nation sometimes seeks to acquire the fruits of another nation's labour via war, but at other times nations acquire these benefits via trade and economic engagement. At least sometimes, this latter approach is desirable not merely on moral grounds but also on self-interested ones. War can be costly even for the victor, and in many cases the

[31] On the other hand, if AI systems are far more competent than humans then they might dramatically change the world, and such change might pose its own risks. Further, if humans provide no value to AI this removes one incentive for systems to avoid harming humans.

[32] Alternatively, AI systems could be strong in respects that allow them to seize power, but weak in some other respects, such that they benefit from human labour. In this case too, human disempowerment might result.

[33] Another argument: as long as there are sufficiently many AI systems, some will likely fall within the Goldilocks zone, even if others don't. However, once we account for interactions between systems this won't necessarily lead to human disempowerment. For example, perhaps the most capable systems will determine what happens to humans, rather than the middling systems within the Goldilocks zone. Or perhaps AI systems that are within the zone will make use of AI, rather than human, labour. Overall, evaluating multi-agent cases would require a paper of its own. I set these possibilities aside.

[34] Alternatively, human disempowerment might be temporary: as a system passes through the Goldilocks zone, it might disempower us until it has progressed sufficiently that humans no longer pose a threat or offer opportunities. At that point, we might be left to our own devices, particularly if the system's goals are narrow scoped.

aggressor cannot be guaranteed victory. Given the costs and risks, peaceful trade might be the best way to acquire another nation's labour. Further, on a prominent view, trade can be pacifying: if two nations engage in trade, this raises the cost of their going to war, because of the resulting trade disruption.[35] If this is right then one way to preclude a threat from another party is to trade with them, rather than establishing hard power over them. So in the case of AI, there might be circumstances where AI is incentivised to use not physical force but instead economic means to benefit from human labour and to preclude human threat.

The same point can be made more directly, rather than via analogy. If an AI system uses force against humans then humans will plausibly retaliate, perhaps by attempting to shut the system down. Further, in doing so, humans will utilise whatever AI systems we control. Even if these are less sophisticated than the adversarial system, they might help to close the capability gap between humans and AI and so increase the chances of successful retaliation. And if we live in a multipolar world—where no single AI system has a decisive strategic advantage and instead multiple systems can compete for power (Bostrom 2014: ch. 11)—then humans might be able to form alliances with some systems against others. In combination, these factors might mean that an AI system takes on a substantial risk if it attempts to forcefully seize power over humans. So AI systems might be incentivised to seek peaceful trade with humans rather than seeking power-over by force.

Of course, even if AI systems use peaceful means, some human disempowerment might result. In the case of human institutions, critics suggest that trade institutions operate as ways to wield power over others. For example, such concerns have been raised against both the United-States-backed Bretton Woods Institutions (Leech & Robert 2004) and China's Belt and Road Initiative (Chellaney 2017). Similarly, perhaps AI systems could use economic levers and social

[35] According to Martin et al. (2008), bilateral trade decreases the risk of war between two parties, but multilateral trade can increase the risk of war (by giving parties more trading options and so making them less reliant on peace with any given nation). In the context of AI, this might suggest trade is most likely to be pacifying in a world with just one advanced AI system.

manipulation to exert power over humans without using force. Still, the argument for human disempowerment now becomes more complex, with discussion needed of why, and in what ways, we should expect economic relationships between AI and humans to lead to human disempowerment. At the very least, the argument that AI power-over would lead to substantial human disempowerment proceeds most straightforwardly if force, rather than trade, is the best route to utilising human labour and precluding human threat.[36]

6 Power To

Moving onto power-to—where this involves the capacity to undertake a broad range of actions—I'm assuming in this paper that AI systems will acquire substantial power of this sort. Yet power-to doesn't have a dual nature, and so AI acquisition of such power needn't lead to human disempowerment. However, while power-to might not have a dual nature as a general matter, there are two reasons to think that AI systems acquiring such power could lead to human disempowerment.

First, some ways of acquiring power-to might have this consequence, and we might expect AI systems to acquire power-to in these ways. For example, a paradigm way that AI systems might seek power-to is by acquiring resources (Carlsmith 2021: 19). Such resources might include money, natural resources, land, and capital like factories and equipment. Sometimes, a resource can be acquired without depriving others. For example, by developing a new mining technology, the developer potentially increases the stock of mined resources in a way that might increase both their own and others' possession of the resource. On the other hand, sometimes when one actor acquires a resource, this comes at a cost to another. Most straightforwardly, this occurs when a resource is stolen, which might happen when theft is easier than acquiring the

[36] One question is how these factors (force beating trade and the existence of a Goldilocks zone) interact. For example, perhaps the balanced capabilities in the Goldilocks zone favour trade over force. Or perhaps not: there are plausibly cases where one group can provide useful labour for another but where force is the most effective way of acquiring this (likewise for precluding threat).

resource from scratch. However, this can also occur simply because there are limited easily-acquirable stocks of the resource. So resource acquisition as a means to power-to might sometimes deprive others of resources and, in doing so, potentially undermine their power-to.

Further, human disempowerment could result not (or not solely) from how power-to is acquired but also from the use of such power. For example, AI power-to might be used to acquire illegitimate power-over humans, for reasons outlined in §5. Or AI power-to might be used to shape the trajectory of civilisation, potentially leading to disenfranchisement. Or AI power-to might be used to shape the world in ways that make it less hospitable to humans, and hence lead to human incapacitation.

The general point is that in acquiring, and utilising, power-to one changes the world, and in so doing, one can disempower others. As Popitz (2017: 117) notes, 'The person who alters realities… generally alters the conditions of existence not only for him- or herself, but also for others. The person who cultivates land, plants trees, poisons forests, biologically kills waters decides possibilities and burdens for generations to follow.' Just as human power to change the physical world can impact, and disempower, other humans, so too could AI power-to have such consequences.[37]

To unpack this, it will be helpful to again consider each form of disempowerment in turn. In doing so, I'll focus for concreteness on disempowerment resulting from AI acquisition of power-to via acquisition of resources. However, similar concerns might arise from AI use of power-to, once it's acquired.

Domination. Per §5, AI systems might desire illegitimate power over humans. Power-to might be used to achieve this, so AI power-to might lead to domination.

[37] See Popitz's (2017: 15–18; 116–118) discussion of the power of data constitution. More generally, Popitz's taxonomy of power can illuminate AI power (as well as how, in developing AI systems, humans can exert power over other humans).

Incapacitation. A similar story could be told about incapacitation—by appealing to the discussion in §5—but I'll set this possibility aside and consider whether AI power-to provides a distinctive route to human incapacitation. Here, the answer turns again on whether some AI systems will have broadly-scoped goals. If all systems have modest goals then their resource needs are also likely to be modest. AI resource acquisition is then unlikely to preclude human power to construct flourishing lives.[38] On the other hand, if an AI system has broadly-scoped goals that push it to insatiably pursue resources to achieve an unbounded goal then human incapacitation could result. In particular, if AI systems can seize basically all resources then they might be incentivised to do so if they have broadly-scoped goals, in a way that deprives humans of what we need to construct desirable lives.

An extreme form of this concern is captured in Yudkowsky's (2008: 333) line that, 'The AI neither hates you, nor loves you, but you are made out of atoms that it can use for something else.' His claim is that if an AI system is sufficiently insatiable then even our bodies contain resources that the system might be unwilling to forego. A chilling thought, of course, though it's worth remembering that it's far from established that AI systems will have broadly-scoped goals to this extreme extent, not to mention that Yudkowsky's claim relies on other disputable assumptions.

Note that these arguments don't rely on the AI system exercising power over humans in a direct sense: the system needn't force humans to carry out certain actions for incapacitation to result. As a result, the argument here doesn't rely on the existence of a Goldilocks zone, where an AI system can overpower humans but benefits from our labour or is threatened by our actions. In the current case, an AI system may be incentivised to take actions that incapacitate humans even if humans pose no threat and offer no opportunities.

[38] This assumes there are relatively few systems outside human control. Sufficiently many systems, each acquiring small quantities of resources, might in combination preclude human acquisition of necessary resources (just as plagues of locusts can strip areas of vegetation).

Disenfranchisement. There are at least three ways that AI power-to could lead to human disenfranchisement.

First, AI could use its power-to to acquire power over humans and this could lead to disenfranchisement, per §5.

Second, echoing the discussion of incapacitation, if AI systems are sufficiently insatiable in pursuing resources then humanity might be left with extremely sparse resources, such that it's impossible for us to construct a desirable civilisation. A familiar refrain: such disenfranchisement is most likely if AI systems have extremely broadly-scoped goals, as it's such goals that are most likely to lead AI to insatiably pursue resources.

Third, humans might retain sufficient resources to shape a flourishing human society but the vast majority of resources might nevertheless be controlled by AI. One version of this might involve humanity retaining control of most of Earth's resources, while AI systems come to control the remaining resources of the Milky Way and use these to construct a galaxy-scale civilisation. In this scenario, humanity is disenfranchised in that we get no say in the broadest sweep of civilisation—which is happening across the galaxy—even if we retain control of our tiny corner of it.[39]

One way this scenario might result is if AI systems have broadly-scoped goals, and yet contra my earlier assumption, have minimal concern for human flourishing and autonomy. Another way that this scenario could result would be if technological development led to certain strategic situations. For example, if humanity had access to sufficiently destructive weapons (and had second strike capabilities) then the benefits, to AI systems, of gaining Earth's resources might be outweighed by the risk of human retaliation. In this case, even purely self-interested AI systems might be incentivised to allow humanity to maintain a minimally flourishing civilisation.

[39] A similar scenario might involve an intermingled civilisation of humans and AI, but where AI systems exert the majority of control. Once such a society existed, this disenfranchisement of humanity might be just, if the AI systems had moral status and exerted more influence due to numbers. However, it doesn't follow that we should see disenfranchisement as desirable ahead of time, when deciding whether to create such systems.

So either based on a concern for human flourishing, or for self-interested reasons, AI systems might be motivated to make use of the majority of the galaxy's resources while reserving enough for human use to allow us to autonomously maintain a flourishing civilisation. In either case, humanity would retain some power but nevertheless be largely disenfranchised, in that we would be unable to shape the broadest sweep of civilisation.[40]

So if we assume that AI will acquire power-to then a case can be made that this will lead to human disempowerment. It might lead to domination, if power-to is used to acquire power-over. And it might lead to incapacitation and disenfranchisement if some AI systems have broadly-scoped goals.

7 Conclusions

On one reading, there's an equivocation involved in moving from the claims that AI will be empowered and that power has a dual nature to the claim that humans will be disempowered: the former claim is most straightforwardly about power-to and the latter most straightforwardly about power-over. However, there are two more sophisticated arguments for human disempowerment, which avoid equivocation.

First, it could be argued that AI might acquire not just power-to but also power-over, perhaps in order to preclude human threat and take advantage of human labour. AI power-over could plausibly involve human domination. Further, if AI systems have broadly-scoped goals, AI power-over might also lead to incapacitation and disenfranchisement. Second, it could be argued that AI power-to might lead to human disempowerment, where this is especially likely to the

[40] Some might deny this amounts to human disempowerment, as humanity remains able to create a flourishing civilisation. Here, I think it's hard to separate semantic and evaluative questions: insofar as this disenfranchisement is undesirable, it's more reasonable to describe it as involving disempowerment. I won't resolve the evaluative question and hence will leave the semantic question somewhat open.

Still, here's three considerations. First, as in note 39, we might distinguish how we should think of this ahead of time—when we can influence whether the AI systems are created—from how we should think of this once the systems' existence is settled. Second, one consideration is whether the AI systems will, in their sphere of influence, construct a civilisation that is morally desirable. Third, a related consideration is the extent to which morality is itself insatiable, in the sense of requiring, if possible, that vast resources be dedicated to the moral project.

extent that AI has broadly-scoped goals.

However, these arguments rely on a series of claims, where many of these are empirical claims about future AI systems and the world in which these systems will be embedded. Further, these claims become particularly crucial if one wishes to argue that humanity won't merely be mildly disempowered by AI but will be radically disempowered. While I cannot hope to assess these claims here, I'll close by highlighting some of the more important of these.

First, four claims about AI and goals: (1) some AI systems will be goal-driven; (2) some of these will have broadly-scoped goals; (3) humans will lack the control of these goals needed to ensure they're conducive to human flourishing; and (4) these goals will incentivise power-seeking.

Second, three claims relevant to why AI will be well placed to acquire power as a result of power-seeking: (1) some goal-driven AI systems will be substantially more capable than humans at a range of relevant tasks; (2) goal-driven systems will be more numerous than humans; and (3) humans will voluntarily hand substantial power to goal-driven systems.

Finally, three claims about the broader context: (1) there exists a Goldilocks zone, where AI systems can establish power over humans but can benefit from human labour or need fear human threat; (2) for AI, human labour will be more effectively acquired, and human threat more effectively precluded, by use of force than by peaceful cooperation; and (3) AI systems will seek to acquire power-to in contexts where such acquisition leads to the disempowerment of others.

These claims don't all need to be true for radical human disempowerment to result, nor would the truth of them all mean a certainty of disempowerment. Still, clarity about each would leave us better placed to assess whether human disempowerment would result from AI takeover.

Those most concerned about takeover, like Yudkowsky and Bostrom, tend to take these claims seriously: they envision AI systems that are radically more capable than humans, with broadly-scoped goals, engaging in unbounded pursuit of power. So, in one sense, this paper vindicates Bostrom and Yudkowsky: if their assumptions hold then the case for human disempowerment is relatively strong.

However, this paper also emphasizes the extent to which the case for disempowerment is most plausible given these extremely strong claims: if these claims don't hold, the case for radical disempowerment is notably weaker. My own view is that existing arguments for these claims are speculative, and so Bostrom and Yudkowsky are not currently vindicated in full. Instead, the matter remains unsettled; for now, we are uncertain.[41] I hope that future work will explore these claims further, and so clarify the likelihood of human disempowerment.

References

Adams, Z. & Browning, J. (eds) (2017). *Giving a Damn : Essays in Dialogue With John Haugeland*, Cambridge, MA: The MIT Press.

Allen, A. (1998). 'Rethinking Power', *Hypatia*, 13/1, 21–40.

Badia, A. P. et al. (2020). *Agent57: Outperforming the Atari Human Benchmark*. https://arxiv.org/abs/2003.13350

Bengio, Y. et al. (2023). *Managing AI Risks in an Era of Rapid Progress*, https://arxiv.org/abs/2310.17688

Bostrom, N. (2012). 'The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents', *Minds and Machines*, 22/2, 71–85.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Carlsmith, J. (2021). *Is Power-Seeking AI an Existential Risk? (Draft Report)*. https://arxiv.org/abs/2206.13353

Chalmers, D. J. (2010). 'The Singularity: A Philosophical Analysis', *Journal of Consciousness Studies*,

[41] In fact, I think that absent strong evidence we should be sceptical of extreme views of AI progress and capabilities and of apocalyptic claims. So I take the evidence's speculativeness to support scepticism about AI takeover (while nevertheless seeing value in clarificatory work). Still, I don't argue for this epistemic position here, so I'll settle for the weaker conclusion that the matter is unsettled.

17/9–10, 7–65.

Chellaney, B. (2017). 'China's Debt-Trap Diplomacy', *Project Syndicate*. https://www.project-syndicate.org/commentary/china-one-belt-one-road-loans-debt-by-brahma-chellaney-2017-01

Chemero, A. (2023). 'LLMs Differ From Human Cognition Because They Are Not Embodied', *Nature Human Behaviour,* 7, 1828–9.

Colombo, M. and Piccinini, G. (2023). *The Computational Theory of Mind*. Cambridge, UK: Cambridge University Press.

Cotra, A. (2022). 'Without Specific Countermeasures, the Easiest Path to Transformative AI Likely Leads to AI Takeover'. *AI Alignment Forum*. https://www.alignmentforum.org/posts/pRkFkzwKZ2zfa3R6H/

Dahl, R. (1957). 'The Concept of Power', *Behavioral Science*, 2/3, 201–15.

Drexler, K. E. (2019). *Reframing Superintelligence: Comprehensive AI Services as General Intelligence*. Future of Humanity Institute, University of Oxford.

Goldstein, S., & Kirk-Giannini, C. D. (forthcoming). 'Language Agents Reduce the Risk of Existential Catastrophe'. *AI & Society*.

Haugeland, J. (1979). 'Understanding Natural Language'*, The Journal of Philosophy,* 76/11, 619–32.

Hendrycks, D., Mazeika, M., & Woodside, T. (2023). *An Overview of Catastrophic AI Risks*. https://arxiv.org/abs/2306.12001

Hoffmann, J. et al. (2022). *Training Compute-Optimal Large Language Models*. https://arxiv.org/abs/2203.15556

Jumper, J. et al. (2021). 'Highly Accurate Protein Structure Prediction with AlphaFold', *Nature*, 596/7873, 583–9.

Karnofsky, H. (2022). 'Why Would AI 'Aim' to Defeat Humanity'. *Cold Takes*. https://www.cold-takes.com/why-would-ai-aim-to-defeat-humanity/

Krakovna, V. et al. (2020). 'Specification Gaming: The Flip Side of AI Ingenuity'. *Deepmind Blog*.

https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity

Landgreve, J. & Smith, B. (2023). *Why Machines Will Never Rule the World.* New York: Routledge.

Leech, D., & Robert, L. (2004). *Voting Power in the Bretton Woods Institutions* (Warwick Economic Research Papers). University of Warwick, Department of Economics. https://wrap.warwick.ac.uk/1472/

Lohn, A., & Musser, M. (2022). *AI and Compute: How Much Longer Can Computing Power Drive Artificial Intelligence Progress?* Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/ai-and-compute/

Lukes, S. (2005). *Power: A Radical View* (2nd expanded edition). New York: Macmillan.

Martin, P., Mayer, T., & Thoenig, M. (2008). 'Make Trade Not War?', *The Review of Economic Studies*, 75/3, 865–900.

Morriss, P. (2002). *Power: A Philosophical Analysis (2nd edn)*. Manchester: Manchester University Press.

Müller, V. C., & Cannon, M. (2021). 'Existential Risk from AI and Orthogonality: Can we Have it Both Ways?', *Ratio*, 35/1, 25–36.

Ngo, R., Chan, L., & Mindermann, S. (2023). *The Alignment Problem From a Deep Learning Perspective*. https://arxiv.org/pdf/2209.00626.pdf

Ord, T. (2020). *The Precipice.* London: Bloomsbury.

Pettit, P. (Ed.). (1997). *Republicanism: A Theory of Freedom and Government*. Oxford: Oxford University Press.

Pitkin, H. (1972). *Wittgenstein and Justice: On the Significance of Ludwig Wittgenstein for Social and Political Thought*. Berkeley: University of California Press.

Popitz, H. (2017). *Phenomena of Power* (G. Poggi, trans). New York: Columbia University Press.

Rapaport, W.J. (unpublished). *Is Artificial General Intelligence Impossible?*

Roser, M. (2023). *AI Timelines: What do Experts in Artificial Intelligence Expect for the Future?,* https://ourworldindata.org/ai-timelines

Shah, R. et al. (2022). *Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals*. https://arxiv.org/abs/2210.01790

Silver, D. et al. (2016). 'Mastering the Game of Go With Deep Neural Networks and Tree Search'. *Nature*, 529/7587, 484–9.

Silver, D. et al. (2017). *Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm*. https://arxiv.org/abs/1712.01815

Sparrow, R. (forthcoming). 'Friendly AI Will Still be our Master'. *AI & Society*.

Sutton, R. (2019). 'The Bitter Lesson'. *Incomplete Ideas*. http://www.incompleteideas.net/IncIdeas/BitterLesson.html

Taylor C. 'Almost Half of CEOs Fear A.I. Could Destroy Humanity 5 to 10 Years From Now—But One "A.I. Godfather" Says an Existential Threat is "Preposterously Ridiculous."', *Fortune*. https://fortune.com/europe/2023/06/15/yann-lecun-ai-godfather-destroy-humanity-threat/

Thorstad, D. (Unpublished). *Against the Singularity Hypothesis*.

Wang, T. T. et al. (2023). *Adversarial Policies Beat Superhuman Go AIs*. https://arxiv.org/abs/2211.00241

Yudkowsky, E. (2008). 'Artificial Intelligence as a Positive and Negative Factor in Global Risk'. In M. J. Rees, N. Bostrom, & M. M. Cirkovic (Eds.), *Global Catastrophic Risks,* 308–45. Oxford: Oxford University Press.

The University of Oxford, United Kingdom