AI & SOCIETY (2025) 40:6277–6289
https://doi.org/10.1007/s00146-025-02371-1

OPEN FORUM

A polycrisis threat model for AI

Adam Bales[1]

Received: 3 March 2025 / Accepted: 14 April 2025 / Published online: 7 May 2025

Abstract
A catastrophic AI threat model is a rigorous exploration of some particular mechanisms by which AI could potentially lead to catastrophic outcomes. In this article, I explore a polycrisis threat model. According to this model, AI will lead to a series of harms like disinformation and increased concentration of wealth and power. Interactions between these different harms will make things worse than they would have been had each harm operated in isolation. And the interacting harms will ultimately cause or constitute a catastrophe. My aim in this paper is *not* to defend the inevitability of such a polycrisis occurring. Instead, I aspire merely to establish that polycrisis-driven catastrophe is sufficiently plausible that it calls for further exploration. In doing so, I hope to emphasise that alongside worries about AI takeover, those concerned about catastrophic risk from AI should also take seriously worries about extreme power concentration and systemic disempowerment of humanity.

Keywords Existential risk · Polycrisis · Threat modelling · AI takeover · Systemic disempowerment · Power concentration

1 Introduction

Some people worry that artificial intelligence (AI) could lead to catastrophe, perhaps even causing human extinction (cf. Bostrom 2014, ch. 8). Given the potential stakes involved, it's important to assess how seriously we should take these worries so that we can work to mitigate the risks if necessary.

Along these lines, a *threat model* is a systematic assessment of particular mechanisms by which AI might lead to catastrophe. For example, a power-seeking threat model explores whether future, highly-capable AIs will seek and acquire power, before using this to cause either human extinction or the disempowerment of humanity (Carlsmith 2021; Ngo et al. 2023; Carlsmith, Forthcoming).

While this particular threat model is relatively well explored, other potential paths to catastrophe are comparatively neglected in the academic literature. In this paper, I'll partially address this neglect by exploring a largely distinct model: a polycrisis threat model. According to this model, AI will lead to a cluster of harms, including disinformation, broader access to dangerous weapons, and increased wealth inequality, among others. Interactions between these harms will make the whole worse than the sum of its parts. And these interacting harms will result in catastrophe.[1]

My purpose in exploring this threat model isn't to outline the most probable scenario for the future of AI. Instead, I aim to clarify one way that things could go severely awry. Consequently, I won't argue that polycrisis-driven catastrophe is inevitable but merely that it's plausible enough to call for further reflection.

A comparison: when NASA considers how a space mission could go severely wrong they don't provide decisive arguments that some failure will result. Instead, when engaged in a process of risk modelling, they focus not on what's inevitable but on what failures remain sufficiently plausible to call for exploration or mitigation. Here, I consider AI in the same spirit.

2 Assumptions about AI

In this paper, when I speak of catastrophe, I'll have in mind *existential catastrophe*, where this involves, "the destruction of humanity's longterm potential" (Ord 2020, p. 37). A paradigm existential catastrophe would involve human extinction. As further examples of existential catastrophe,

✉ Adam Bales
adam.bales@philosophy.ox.ac.uk

[1] University of Oxford, Oxford, United Kingdom

[1] A related threat model is explored in Kasirzadeh, forthcoming. My discussion and Kasirzadeh's are complementary: we adopt different framings, and draw on different evidence, in a way that bolsters the robustness of the case for taking a family of threat models seriously.

Ord points to the unrecoverable collapse of civilisation and the imposition of a permanent and dystopian dictatorship.

Most people doubt that AI is currently capable of causing catastrophe on this sort of scale. However, as AI becomes more sophisticated and pervasive, catastrophic harm becomes more likely. Consequently, concerns about catastrophe typically relate to AI systems that will be developed in the coming years or decades, rather than to AI as it is now.

This raises a challenge: we don't know what the future will look like, and any threat model relating to the future will need to address this uncertainty. There are two ways to do so: a threat model can either *assess* the plausibility of certain claims about the future or it can *assume* their truth. In the latter case, the threat model is conditional: it assesses the risk of catastrophe via some mechanism conditional on the future unfolding in the specified manner.

I'll take this latter approach and so will assume, without argument, certain claims. Broadly, I'll assume that the future sociotechnical landscape surrounding AI will look similar to the landscape now in important respects, though with technological advances along the way and with AI diffusing out into wider use. More concretely, I'll assume:

1. *A Multipolar, Human-Driven World*—While AI might cause some concentration of power, no single actor will decisively dominate the world in the coming decades. Instead, power will be held by a diversity of humans and human institutions, like governments and corporations. This assumption is primarily intended to rule out AI allowing a single actor to behave with confident impunity, even in the face of united opposition from the rest of the world (see Bostrom 2014, ch. 5). In the intended sense, human civilisation has always been multipolar, including during America's so-called unipolar moment after the Cold War. While the US government possessed a great deal of power during this period, other nations (and corporations) remained sufficiently powerful as to act as checks on US government power. So here, I'm simply ruling out an extreme, unprecedented form of unipolarity.
2. *A Human-Comprehensible Pace of Change*—In the coming decades, AI will accelerate the pace of social and technological change, and harm will result from human failure to adapt. Nevertheless, change will occur sufficiently slowly that humans will remain capable of engaging with it, at least by using AI tools and developing more agile institutions. This can be contrasted with the possibility of AI leading to radical changes over weeks, days, or even hours, such that humans can't fruitfully engage with the changing world (Bostrom 2014, ch. 4).
3. *No Unbounded and Insatiable Agentic AI*—In the coming decades, AI agents will proliferate. For example, military drones might act independently and flexibly in pursuit of a target. Likewise, an AI CEO might pursue a range of strategies to promote shareholder value. Nevertheless, we won't see agents of a particularly insatiable and unbounded sort. An example of such an agent is the hypothetical paperclip maximiser that conquers the world to acquire the resources needed to construct evermore paperclips (Bostrom 2003). I'll assume that AI agents will not be so insatiable nor will they use such radically unexpected strategies in pursuit of seemingly innocuous goals.[2]

This list emphasises similarities between the present and future, but I'll also assume change. In particular, I'll assume AI will rapidly become *increasingly sophisticated*: AI will become able to carry out an increasing range of tasks, many to a level that equals or exceeds human performance.[3] I'll also assume AI will become *increasingly pervasive*: it'll play a central role in the economy, become more important in military and geopolitical contexts, and become integrated with R&D processes, leading to the development of a raft of new technologies.

While I won't argue for these assumptions, I take them to be sufficiently plausible that they represent one scenario that we should reflect on and prepare for.[4] Taking these assumptions as a starting point, this paper will explore whether AI could cause catastrophe.

3 Polycrisis

In this paper, I'll explore a polycrisis threat model, according to which AI will cause a set of interacting harms that lead to catastrophe. This model can be broken down into three claims.

HARMS—First, according to HARMS, AI will involve multiple *harm factors*, each of which captures one way AI

[2] This could be because constrained agency arises naturally from current techniques or is more useful than unbounded agency (as unbounded agents act unpredictably). Or perhaps tools, with humans in the loop, outperform independent agents (because agents are unreliable). Or perhaps regulation constrains the development of agents.

[3] See Grace et al. 2024 for a survey suggesting that many AI researchers take seriously the possibility of AI rapidly increasing in sophistication in this way.

[4] Even if these assumptions hold for only a short period, before further changes follow, it might be important that we navigate this period well. If so, it's worth reflecting on challenges for doing so.

will cause harm, where it's natural and useful to consider these factors independently. For example, one harm factor could be AI-enabled cyberattacks; such attacks are one way AI could cause harm. Another harmful factor could be AI-generated disinformation.

HARMS has a pragmatic component, relating to how it's natural and useful to think. After all, any individual harm factor can be carved up in a finer-grained way to re-characterise it as multiple factors. For example, we could carve up harm from disinformation into harm caused by disinformation distributed on Monday, that caused by disinformation distributed on Tuesday, and so on. Still, it's neither natural nor useful to frame things this way. Establishing HARMS requires showing that AI will involve multiple harm factors in a more natural and useful way of carving things up.[5]

INTERACTION—According to INTERACTION, the interaction between harm factors will make things much worse than it would have been had the factors operated in isolation. If we assume the degree of harm resulting from some factor can be represented numerically then the claim is that the harm resulting from the factors operating in conjunction will be much greater than the sum of harms that would result from each factor operating in isolation.[6]

CATASTROPHE—Finally, according to CATASTROPHE, the harm resulting from the set of harm factors, after accounting for interactions, will be sufficiently large to cause or constitute a catastrophe.[7] In particular, as above, I'll explore the possibility that the interacting harms from AI will constitute, or lead to, an *existential* catastrophe, where such a catastrophe destroys humanity's long-term potential.

In the remainder of this paper, I'll evaluate the polycrisis threat model by evaluating each claim in turn. But before I do so, two caveats.

First, as my aim is to show that polycrisis is worth taking seriously—and not to show that it's inevitable—I won't provide decisive arguments for these claims. Instead, I'll merely show that each claim is sufficiently plausible that, from a precautionary perspective and given what's at stake, it's worth exploring what follows from them.

Second, in discussing each claim, I'll focus on how things might go awry rather than how things might go well. For example, in discussing AI-enabled cyber attacks, I won't discuss the possibility that AI bolsters cyber defence more than cyber offence, leading to more secure digital systems (see Bonfanti 2022; Newman 2024). This is because I'm interested in the scenarios where AI goes most severely wrong, and in these scenarios, I expect harmful impacts to be the dominant consideration. So I focus on these.[8]

With those caveats in mind, on to the arguments.

3.1 HARMS

I'll start by considering HARMS, according to which there will be multiple harm factors, each representing one way AI will cause harm. I'll make the case for this by exploring a series of potential harm factors (my discussion of each will be brief, so I'll mostly flag potential factors and point to discussions elsewhere).[9]

(This paper was finalised on February 15th, 2025. Given the rapid pace of developments in AI and the slower pace of academic publishing, there will undoubtedly be further developments by the time you read this. So this section should be read as a snapshot taken at the time of writing.)

3.2 Misuse

Drawing on a taxonomy from Zwetsloot and Dafoe (2019), one way AI could cause harm would involve its misuse by malicious human actors. I'll consider four types of misuse harm.

Cyberattacks—AI might assist hostile actors in carrying out cyberattacks (Bengio et al. 2025, §2.1.3). It might discover and exploit vulnerabilities in digital systems, develop malware, and help individuals develop dangerous cyber expertise. Indeed, large language models (LLMs) already display cyber capabilities (Xu et al. 2024; Happe et al. 2024; Fang et al. 2024a; Fang et al. 2024b; Shao et al. 2024; Phuong et al. 2024, §4; UK AISI 2024; US AISI & UK AISI 2024a, part II; US AISI & UK AISI, 2024b, part I; Anthropic 2024; OpenAI 2024, pp. 13–16). While these capabilities are currently rudimentary, as LLMs grow

[5] This is a matter of degree. Still, as we'll see in Sect. 4, carving into multiple harms is particularly natural and useful for the polycrisis threat model.

[6] We need to be careful about how we attribute harm to factors, to avoid double counting, but the thought behind INTERACTION is clear enough without unpacking this. Note, the risk being considered is that interactions *actually* make things worse, but as we lack a crystal ball, in assessing the model, we'll consider *expected* harm.

[7] This claim could be strengthened (eg. by requiring that the interactions make a counterfactual difference to catastrophe's occurrence). However, I'll stick with the weaker claim, while nevertheless focusing on cases where multiple harm factors, and interactions between them, make the catastrophe much worse or more likely.

[8] It follows that evaluating the polycrisis model requires evaluating the claim that AI's benefits won't preclude the harms discussed. Space rules out discussion, but I think this too is plausible enough that it's worth exploring what would follow.

[9] This list isn't exhaustive. For example, I don't discuss climate impacts of AI.

more sophisticated they might enable cyber attacks to be conducted at a greater scale, by a wider range of actors.

AI might also assist with social engineering attacks, which involve manipulating people to gain desired outcomes like access to digital systems. For example, LLMs can already assist with email phishing campaigns, where an attacker impersonates a legitimate source to deceive a target into downloading malware or revealing sensitive information (Hazell 2023; NCSC 2024; Heiding et al. 2024).[10]

AI-enabled cyberattacks could erode privacy and enable blackmail by providing access to personal information. They could also diffuse dangerous information and threaten critical infrastructure such as hospitals or power grids. To get a sense of the potential costs of such attacks, consider the WannaCry attack targeting the United Kingdom's National Health Service, which ultimately cost £92 million (Department of Health and Social Care 2018, §6). An increase in the frequency and severity of such attacks could constitute severe harm.

Inference & Surveillance—Beyond cyberattacks, there are two further ways AI could undermine our ability to secure information.

First, using available information, AI can *infer* further information. For example, AI can infer facts about a person's mental health based on their social media posts (Guntuku et al. 2017; Ahmed et al. 2022; Zhang et al. 2022). Indeed even LLMs, which aren't specifically designed for such inferential tasks, are able to make informative inferences about people (Staab et al. 2024).

Second, AI can incentivise and enable *surveillance*. AI's inferential capabilities incentivise surveillance of our online activities because the gathered data can be used both in training inferential models and as the basis for making valuable inferences about us (Véliz 2020; Solow-Niederman 2022; Benn & Lazar 2022, pp. 132–134). Further, AI enables surveillance in the physical world (Feldstein 2019). For example, facial recognition algorithms allow video surveillance to be fruitfully used on a previously impossible scale (Smith & Miller 2022; Hill 2023).

As with cyberattacks, the knowledge generated in these ways could be used for blackmail. It could also be used by states to constrain freedoms, for example by enforcing repressive laws (Strzyżyńska 2022) or identifying protesters (Ryan-Mosley & Richards 2022).

Disinformation—LLMs can cheaply generate textual disinformation at scale (Goldstein & Sastry 2023; Bengio et al. 2025, §2.1.2). Over time, as models have grown more sophisticated, the text generated in this way has become more persuasive, and it's now roughly as persuasive as human-generated text (Jakesch et al. 2023; Spitale et al. 2023; Costello et al. 2024; Durmus et al. 2024; Rogiers et al. 2024; Williams et al. 2024; Bai et al., Unpublished).[11] Further, disinformation generated by LLMs can be targeted at individuals or demographic groups, either by having the LLM personalise the text or by using recommender systems to determine who receives what information. This targeting might make the disinformation more efficacious.[12]

Disinformation could cause various harms. Democratic states could be undermined if disinformation influences election results, increases polarisation, or decreases trust in democratic processes.[13] Harm could also result if AI bolsters autocracies by allowing the creation of personalised, persuasive propaganda. And, in more general terms, AI-generated disinformation could harm societal epistemics, either by leading people to false beliefs or by causing them to distrust reliable information (van Doorn 2023, §1).

Weapons—AI could enable the development and proliferation of chemical and biological weapons, military drones, other autonomous weaponry, and novel weapons. Space precludes discussing all of these, so I'll focus on biological weapons, and in particular, on the possibility that LLMs will enable proliferation of such weapons.

Here, the worry is that LLMs might be able to act as advisors to help groups develop biological weapons when they would otherwise have been incapable of doing so (Bengio et al. 2025, §2.1.4; Hendrycks, 2024, §1.2.1). LLMs already display biological capabilities, with these allowing expert-level performance on some metrics (UK AISI 2024; US AISI & UK AISI, 2024a, part I; US AISI & UK AISI 2024b, part II; Anthropic 2024, p. 25[14]; OpenAI 2024, §5.5). While these capabilities are sufficiently rudimentary that current LLMs probably don't substantially increase the risk of biological weapons proliferation, this could change as LLMs grow more sophisticated.[15] LLMs might then lead to

[10] From the other direction, AI might allow better detection of phishing emails. However, as noted, I'll focus on the possibilities where AI causes harm, on balance, in some domain.

[11] LLMs might also help us detect, and combat, disinformation (Lucas et al. 2023; Wang et al. 2024; Ernst 2024). However, as noted, I'll focus on the case where the harms outweigh the benefits.

[12] More work is needed to reach conclusions about the size of any effect here (Tappin et al. 2023; Simchon et al. 2024; Hackenburg and Margetts 2024; Salvi et al. 2024).

[13] Disinformation campaigns have plausibly had little impact on election outcomes thus far (Stockwell et al. 2024a, b; Stockwell 2024), but this could change as AI becomes more sophisticated.

[14] See also the two addenda to this report.

[15] My assessment is based on the cited evaluations of LLM capabilities. See also Mouton et al. 2024.

the proliferation of biological weapons in more states or to groups like terrorist organisations.

If future LLMs lead to a proliferation of biological weapons, this could cause substantial harm given how dangerous diseases and other biological agents can be. Such harm could result from the deliberate use of biological weapons. It could also result from accidental leaks from labs or storage facilities, which is especially plausible given a history of leaks even from highly secure facilities (Ord 2020, pp. 130–131).

So there are various ways misuse of AI could cause harm.

3.3 Accidents and systemic harms

Not all harms from AI require malicious misuse by humans. Harm can also result from accidents or from AI's systemic impacts (Zwetsloot and Dafoe 2019). In the latter case, harm doesn't result from the actions of any single agent nor from any single, acute event. Instead, it emerges from the actions of many agents and involves diffuse, systemic changes in the world. Here, I'll comment first on accidents and then on two systemic harms.

Accidents[16]—If a self-driving car hits a pedestrian, this needn't involve human misuse of the autonomous driving system. Instead, the harm could be accidental; resulting from a flawed system. As AI becomes more pervasive and is increasingly deployed in critical infrastructure, these sorts of unanticipated failures of AI systems could cause more harm, more often. For example, such failures could lead to blackouts and hospital shutdowns (on accident risk, see Arnold and Toner 2021; Bengio et al. 2025, §2.2.1; Hendrycks, 2024, §1.4).

Of course, accidents already occur even in AI's absence. So if AI is to make things worse then it must cause greater accidental harm than the systems it replaces. This could happen if AI is more error prone than humans in some domains while also being cheaper, such that there's an incentive to deploy it.[17] Alternatively, it could occur if AI is less error prone than humans but leads to larger-scale accidents when it goes awry. This might happen if AI underpins complex, interconnected systems where failure can bring down the entire system rather than an isolated component. So in some scenarios, AI would lead to an increase in accidental harm as compared with the status quo.

Lower Wages and Higher Unemployment—AI will lead to increased automation of intellectual labour. We're already seeing automation as a result of LLMs and other AI models, and we'll see further automation with time (Eloundou et al. 2023; OECD 2023, ch. 3; Cazzaniga et al. 2024; Colombo et al. 2024; Handa et al. 2025). In addition, AI could enable advances in robotics that lead to the automation of physical labour (see Peel et al. 2024).

Most straightforwardly, the displacement of human labour by automation might lead to decreased demand for labour and so to lower wages or higher unemployment (Susskind 2022; Trammell and Korinek 2023). This outcome is particularly plausible, and could be particularly dramatic if AI and robots can cost effectively replace all human labour (Korinek and Juelfs 2023, pp. 750–753; Korinek and Suh 2024[18]; Barnett 2025), a possibility that many experts take seriously (Roser 2023; Henshall 2024).[19]

On the other hand, the displacement of human labour might be counterbalanced by other impacts of AI (here, I draw from Aghion et al. 2019; Acemoglu and Restrepo 2019). AI might increase productivity, leading to a greater demand for humans to carry out tasks that haven't been automated. AI might also lead to the emergence of new, economically valuable tasks that are best carried out by humans, just as the invention of the computer created the need for IT support tasks. If these impacts outweigh displacement then AI won't lead to a persistent decline in demand for labour. Still, even under these conditions, there could be a period during which unemployment rises or wages decline (Korinek and Stiglitz 2017, §5.2; Korinek and Suh 2024[20]). This could happen for various reasons including a skill mismatch, where workers take time to develop the skills needed to carry out new tasks (Susskind 2022, pp. 648–649).

So AI could lead to either persisting or temporary increases in unemployment or declines in wages. This could be harmful if employment continues to play a central role in financial security, social status, and personal meaning.

Concentration of Wealth and Power—AI could cause concentration of wealth and power.[21]

With respect to individuals, AI could lead to increased income and wealth inequality. This could occur if AI increases wage inequality (Acemoglu and Restrepo

[16] Here, I include harms resulting from AI systems' competent pursuit of undesirable goals. Such harm can be accidental when no human intended it.

[17] This incentive might be particularly strong if some accident costs are externalised, such that they aren't borne by the agent who utilises the AI system.

[18] See the discussion of the two AGI scenarios.

[19] This is a claim about increased plausibility, not inevitability. See Trammell 2025.

[20] See the discussion of the bout-of-automation scenario.

[21] Here, I'm interested in AI's structural impacts, so I won't consider misuse of AI-powered technologies, like autonomous weapons and surveillance technologies. And I'll focus on individuals and nations rather than companies (see Bengio et al. 2025, §2.3.3).

2022; Cazzaniga et al. 2024, p. 17; Bengio et al. 2025, pp. 114–115). Such an increase could result either due to automation of tasks that were previously completed by lower paid workers, leading to the declining value of their labour, or due to AI differentially complementing higher paid workers, leading to an increased value of *their* labour. Further, AI could also increase income inequality by decreasing the returns to labour relative to capital (Autor 2022, p. 23; Moll et al. 2022; Cazzaniga et al. 2024, pp. 17–18; Korinek & Juelfs 2023; Bengio et al. 2025, p. 115). Given the relatively large inequalities in capital ownership, this would increase income inequality.

We could also see increasing concentrations of wealth and power with respect to nations (Bengio et al. 2025, pp. 116–122). After all, at the national level, the resources needed to take advantage of AI are highly concentrated. For example, most advanced AI models are developed in a handful of countries, primarily the United States and China (Rahman et al. 2024; Maslej et al. 2024, p. 61). We see a similar pattern when it comes to investment in AI (Center for Security and Emerging Technology 2024; Maslej et al. 2024, pp. 247–253). And similar patterns hold for other relevant metrics.[22] As a result, AI could lead to power and wealth concentration as its benefits accrue to a handful of nations.[23]

So, this section as a whole suggests various harms could be caused by AI. Further, for HARMS to be true, these wouldn't all need to eventuate, as long as sufficiently many did. While we lack a decisive case for this, the above discussion shows that we should take this possibility seriously. HARMS is plausible enough that we should ask what would follow.

3.4 Interaction

I turn now to INTERACTION, according to which various harm factors interact in such a way that the resulting harm is substantially worse than it would be if each factor operated in isolation.[24]

One way to argue for this claim would be to consider interactions between pairs of harm factors. For example, AI-enabled cyber attacks might lead to the proliferation of AI-developed weapons, by making it harder to control the knowledge required to create these weapons. Or alienation created by wealth inequality might make people more susceptible to disinformation.

However, while there's value in such reflections, I'll explore a more general mechanism by which harm factors can interact: some factors might undermine society's capacity to respond to challenges.[25] If so, this would plausibly leave society less able to address other factors and hence might intensify the harm caused by these other factors. We would have a case for INTERACTION.[26]

To make this case, I'll discuss four ways AI harm factors could undermine society's capacity to respond to challenges.

Undermining of the Social Contract—If AI increases unemployment and inequality this might undermine the social contract, leading to unrest and decreased state capacity. While their accounts differ in many details, Turchin (2023) and Acemoglu & Johnson (2023) each argue for such a dynamic.[27] In broad terms, they note that rising inequality pushes people to seek elite membership, both to escape relative impoverishment and to acquire a share of elite power. Typically, this creates a group who see themselves as unable to acquire elite status in the current system and so seek to overthrow the existing elite. This group can draw on popular support, due to the alienation created by inequality, and the resulting conflict can lead to chronic social instability, as well as acute events like coup attempts.[28] This can undermine state capacity. So, if AI

[22] On patents and research, see Center for Security and Emerging Technology 2024. On control of an important form of compute, see Lehdonvirta et al. 2024. On the ability to attract skilled immigrants, see Zwetsloot et al. 2021.

[23] In addition, if AI-driven automation decreases the value of labour, this might impact some developing nations especially severely, as many developing nations are particularly reliant on demand for their labour (Korinek and Stiglitz 2021; Nii-Aponsah et al. 2023).

[24] From the other direction, interaction between harm factors could decrease the resulting harm. For example, if AI causes dramatic accidents, this might encourage careful regulation of AI, which could mitigate other harms from AI. To fully map out the possible impacts of AI, we'd need to consider such interactions. Still, here I'm focused on understanding a scenario where things go dramatically awry, and in such a scenario, I expect the negative interaction effects to dominate. So I'll focus on these.

[25] Such interactions might give rise to feedback loops, where harm factors mutually reinforce one another so that harm is intensified across multiple cycles. For example, if disinformation undermines societal functioning this might lead to increased unemployment, which might increase alienation, which might increase susceptibility to disinformation, and so on. Feedback loops are particularly concerning when the degree of intensification diminishes slowly, or not at all, with each iteration of the loop.

[26] For a complementary discussion, see Kasirzadeh, forthcoming. Kasirzadeh draws on systems analysis to discuss related issues in relatively abstract terms. She then concretely illustrates these issues with a (knowingly) speculative exploration of how the world might look in 2040. My own discussion occupies a middle ground between these approaches. It explores interactions in more concrete terms than Kasirzadeh's appeal to systems analysis, while focusing on more general mechanisms than her illustrative speculations. In addition, Kasirzadeh and I discuss different mechanisms and evidence. So, overall our discussions complement one another.

[27] In the latter case, see chapters 3, 11 and 12.

[28] For further discussion of inequality and instability, see Alesina and Perotti 1996, Posner 1997, Agnello et al. 2017, and Kent A. Clarke Center 2019.

increases unemployment and inequality, this could make it harder for society to address challenges, including those posed by AI harm factors.

Exhaustion of Societal Capacity—If AI causes substantial harm then addressing this might require substantial work. Companies will need to invest in designing safe AI systems, governments will need to regulate AI and address downstream harms, and civil society and international institutions will need to play central roles. For example, if AI enables cyberattacks or leads to a proliferation of biological weapons, this could empower small groups to carry out disruptive attacks. Responding to, and precluding, such attacks could consume substantial resources.[29]

This would come at an opportunity cost; the resources consumed in addressing AI harms could otherwise have been used elsewhere. Further, if AI harms are substantial then addressing them could overload institutions. Covid is an instructive case study for such dynamics. In the UK, for example, it consumed substantial governmental, medical, and scientific resources and at peak times left little spare capacity in the health systems (Fong et al. 2024). Frequent, large-scale AI harms could also induce disaster fatigue, where a series of disasters undermines societal resilience (Ingham et al. 2022, 2023). Under such circumstances, defeatism and increasingly dysfunctional institutions can leave society unable to draw fruitfully upon the resources at its disposal.

As a result, AI harm factors could lead to decreased societal capacity to address challenges, including those relating to other AI harm factors.[30]

Elite Inaction—If AI increases wealth inequality then elites might be insulated from many of AI's harmful impacts. After all, money can often be used protectively. For example, accidents that shut down a public hospital matter less if you can afford treatment in a private hospital. Violence arising from social conflict matters less if you're in a gated community.

However, this insulation means that elites will plausibly be less aware of the harms AI is causing and have less incentive to act to resolve them. Insofar as elites hold much of the power needed for society to act, this might decrease society's tendency to robustly address challenges, including those posed by AI (see Diamond 2005, pp. 430–431 for an argument that elite insulation has contributed to past societal collapses).

Degraded Epistemic Environment—If AI leads to a proliferation of disinformation then this could lead to both false beliefs and greater distrust in sources of information. Further, this could occur against a backdrop where AI-driven innovations are changing society at an unprecedented pace, such that it's difficult to remain well-informed. As a result, society's epistemic environment might be degraded. People might become more ignorant and might occupy different epistemic bubbles in a way that makes collaborative truth-seeking difficult. This might undermine society's capacity to reach consensus on the challenges raised by AI and to address these challenges.

So AI harm factors might undermine society's general capacity to respond to challenges. And this might lead other factors to cause more harm than they would otherwise have done. This provides a case for INTERACTION.

3.5 Catastrophe

I turn now to CATASTROPHE, according to which the harm resulting from interacting harm factors will cause or constitute an existential catastrophe, where this involves the destruction of humanity's long-term potential. Clearly, any argument about the possibility of such a severe catastrophe will be speculative. Still, speculative or not, there are three potential pathways to catastrophe.[31]

In discussing these, I'll feel free to violate my assumptions from Sect. 2. For example, I'll feel free to explore the possibility of extreme power concentration (rather than a multipolar world). My assumptions were intended to constrain the initial scope of inquiries, not the conclusions reached. That's to say, this paper explores what could follow if, in the coming years, the sociotechnical landscape around AI initially takes a certain form. What follows could be further changes to the sociotechnical landscape and my assumptions aren't intended to constrain these downstream impacts. With that in mind, I turn to the pathways.

Extreme Power Concentration—Interacting AI harms could lead to extreme power concentration (related: Hendrycks 2024, §1.2.4). This could result from unequal access to AI, allowing AI's benefits to accrue to a small

[29] Such attacks could also undermine societal capacity in other ways. For example, drone attacks on shipping might disrupt trade, cyber-attacks might undermine critical infrastructure, and terrorist attacks might sow suspicion between groups.

[30] From the other direction, if AI were integrated into infrastructure and government institutions, this might increase state capacity. Still, as I'm interested in scenarios where things go severely wrong, I focus on the case where AI's harmful impacts dominate.

[31] Two other possibilities: (a) decreased societal capacity could leave us more vulnerable to other potential catastrophes, like pandemics or nuclear war (related: Ord 2020); and (b) use of novel, AI-developed weaponry could cause catastrophe, especially if these weapons are more deadly than nuclear weapons or easier to develop (related: Bostrom 2009).

number of people, companies and nations. Meanwhile, autonomous militaries and surveillance technologies might allow autocrats to wield unprecedented power. And disinformation and degradation of society's capacity to coordinate might make it harder for existing institutions to intervene to stop power concentration (eg. by implementing anti-monopoly policies). The result could be an initial concentration of power that intensifies over time, as those with the greatest access to AI accrue further power. If elites use this power to maintain control then power concentration might persist for a long time.

This could lead to an existential catastrophe. We can see this by reflecting concretely on what extreme power concentration could look like. It could involve totalitarian states with unprecedented control over citizens, monopolistic corporations with vast economic and political power, or a small number of nations holding almost all geopolitical power.[32] In each case, reflecting on the closest historical precedents suggests that widespread suffering could easily follow. At the very least, a world largely controlled by totalitarian dictatorships, monopolistic corporations, or a small number of superpowers is unlikely to lead to the sort of highly positive future—a future of widespread flourishing—that might otherwise have been within reach. So if extreme power concentration persists long term, this might destroy humanity's long-term potential. Consequently, it represents a potential pathway to existential catastrophe.

Systemic Control Loss—As AI becomes more sophisticated, it will come to outperform humans at increasingly many tasks. For other tasks, AI might not outperform humans but it might be cheaper, faster, or both. So, there'll be strong incentives to have AI carry out more tasks over time. These will be strengthened by the fact that groups that fail to delegate to AI will be outcompeted by groups that do. Eventually, the most powerful groups might have handed over the majority of control to AI. AI would then make the central decisions that shape society and be pervasive in decision making (Christiano 2019; Critch and Russell 2023, pp. 4–8; Assadi, Unpublished; Kulveit et al. 2025; Kasirzadeh 2025)**.**

Unfortunately, even if these AIs are highly capable, their integration into society could deliver undesirable results, for two reasons.

First, many AIs will presumably be directed at promoting the narrow interests of one group of humans. For example, an AI might be tasked with making trades to enrich some company or with optimising the economy for the benefit of a nation. Consequently, the future might be shaped by systems aggressively competing to promote the interests of different groups. The results of this competition might be the enrichment of the few at the cost of the many. Alternatively, intense competition might see resources consumed in negative sum interactions, like wars, such that even the few ultimately end up worse off.

Second, AIs might perform best when they optimise for easily measurable targets (like GDP) where these are imperfect proxies for the harder-to-measure things we actually care about (like human wellbeing). Yet, arguably, when imperfect proxies are pursued single-mindedly this often fails to promote what matters.[33] For example, GDP might be best promoted by a highly polluting industrial policy even if this won't lead to increased human wellbeing. Likewise, citation counts like the h-index might be optimised by publishing papers containing errors, to encourage the publication of rejoinders. Yet this hardly increases the academic quality of a person's work, which the h-index was intended as a proxy for. If these toy examples reflect a general fact about the pursuit of imperfect proxies then AI that's highly capable of pursuing proxies might produce undesirable outcomes.

Of course, our world already looks somewhat like this, with proxies like GDP driving many decisions and substantial resources directed at promoting the interests of narrow groups. However, there are at least four reasons AI could make things worse. First, as noted above, AI could lead to a radical concentration of power, and this could mean fewer people benefit in an AI-driven world. Second, if we replace human drivers of society with AIs, society could become less shaped by human preferences. Plausibly, this will lead to a less human-friendly world (Kulveit et al. 2025). Third, if AIs become more capable than humans at many tasks then their impacts, and hence any harm caused, might tend to be on a larger scale. Finally, while we already have some sense of the outcomes of a human-driven system, an AI-driven system is unprecedented in a way that creates greater uncertainty. One consequence of this is that we should be more open to the possibility that an AI-driven system will cause unprecedented levels of harm.

Unfortunately, even if AIs produce undesirable outcomes, humans might struggle to intervene. For a start, elites might benefit from the situation and so have little incentive to act. Further, if any given group unilaterally steps back from

[32] In contrast with my earlier discussions, here AI increases state capacity. Overall, AI could lead to low state capacity in some states, or in some respects, while also leading to high state capacity in other states, or in other respects. AI could also lead first to low state capacity and then this could allow authoritarian coups that lead to high state capacity.

[33] This is often framed in terms of Goodhart's Law, which in one formulation holds that, "When a measure becomes a target, it ceases to be a good measure" (Strathern 1997, p. 308). For discussion of AI and Goodhart's Law, see Manheim and Garrabrant 2019; Thomas and Uminsky 2022.

using AI then they might be outcompeted by other groups. Finally, humans might struggle to identify the source of problems and coordinate solutions, due to lower societal capacity, human deskilling, and the rapid (and hence hard to keep up with) pace of AI decision-making.

In this scenario, no human hand holds the wheel, and the world is shaped by the emergent dynamics of interacting AI systems. If this situation persists then it could lead to an existential catastrophe in a similar way to the power concentration case: most humans might live barely tolerable lives given that AI control could lead to undesirable outcomes. Even if life isn't so grim, AI control could mean that humanity fails to achieve brighter futures that we could otherwise have aspired to. In either case, the long-term potential of humanity would have been destroyed.[34]

AI Powerseeking & Takeover—The most prominent concern about existential risk from AI relates to takeover, where agentic AIs acquire substantial power, seize control of society, and either drive humanity extinct or to build a world that's undesirable from a human perspective.

I won't revisit the arguments for this concern (see Bostrom 2014; Carlsmith 2021; Carlsmith, Forthcoming; Ngo et al. 2023). What I will do is note that there are various reasons one might be sceptical about the takeover. For a start, traditional arguments for takeover rely on AI being both extremely power seeking (so that it's incentivised to attempt to seize vast amounts of power, despite potential risks) and extremely capable (so that it's able to seize power when it attempts to do so). We might doubt that AI will be so power seeking or so capable. Further, we might expect that humans will avoid creating such insatiable power-seeking AI and might expect that (with the aid of AI tools) we'd be capable of resisting AI power seeking if it did occur.

However plausible these rejoinders are if we consider a takeover occurring against the backdrop of society as it now stands, they're less plausible if a takeover occurs in a society wracked by polycrisis, for two reasons.

First, *extreme power concentration*, might mean that AI systems tasked with working on behalf of powerful actors are granted large amounts of power from the outset and might be able to use this to acquire more power.[35] Indeed, *systemic control loss* might mean that AIs possess almost all of the power in society, even without making concerted efforts at takeover. If so, it might require very little for a confederacy of systems to seize control of society, especially if they can use *advanced bioweapons* and *autonomous weapons* (and *AI-enabled hacking* might make it hard to deny AI access to these weapons). These considerations decrease how capable an AI would need to be to take over human society and consequently decrease how power seeking an AI would need to be to make the attempt (by making the attempt less risky).

Second, *systemic control loss* and a *loss of society's capacity to address challenges* will undermine efforts to stop takeover attempts, by making society less likely to recognise the dangers, reach a consensus on the need to act, identify appropriate strategies, and coordinate in implementing these.

None of this provides an independent argument for takeover. Still, arguably it bolsters existing arguments: takeover against the backdrop of polycrisis is more plausible than takeover considered in isolation.[36] Takeover might succeed in the former case when it would have failed in the latter. This provides another pathway by which polycrisis might lead to existential catastrophe.

So there are three pathways from polycrisis to existential catastrophe, each of which is speculative but each of which deserves reflection and exploration.

4 Conclusions

Polycrisis could lead to catastrophe. That is, AI could lead to various harms (per HARMS); interactions between these could make things much worse than they would otherwise have been (per INTERACTION); and the resulting harm could be of a sufficient scale to constitute an existential catastrophe (per CATASTROPHE).

Of course, I haven't shown that catastrophic polycrisis is inevitable or even that it's more likely than not. Still, I hope to have shown that this threat model is sufficiently plausible to call for further evaluation.[37] Such further evaluation might also reveal that the risk of polycrisis calls for mitigation, though I don't take myself to have established that here.

Still, we can ask what such mitigation might look like if it came to this. Prior to further reflection, two conjectures suggest themselves. First, particular effort might fruitfully be focused on mitigating power concentration, which

[34] Such catastrophe could result without polycrisis. However, the power concentration and diminishment of societal capacity, which are central parts of polycrisis as I envision it, make systemic control loss more likely and concerning.

[35] One mechanism for acquiring further power: power concentration plausibly makes it easier to seize power from weaker, but still powerful, actors, as a single strike can suffice for acquiring much power. This dynamic is sometimes thought to explain why the Spanish found it easier to conquer swathes of South America, where power was concentrated, than North America, where power was diffuse (Hämäläinen 2022).

[36] From the other direction, polycrisis might make people more aware of the risks of AI and so more wary about takeover. However, given that polycrisis undermines society's capacity to address challenges, I doubt this will leave us better off overall.

[37] My discussion also suggests that work on threat modelling should be attentive to interactions between components of a model and between different models.

played a central role in the pathways to catastrophe. Second, we should plausibly work to mitigate impacts that could undermine society's capacity to reflectively and cooperatively engage with challenges.

In fact, in working to mitigate risk, we needn't solely focus on avoiding harms but might have more positive goals too. In particular, we might work to promote pluralistic, reflective approaches to the decisions that shape our society. So the risk of polycrisis need not solely be a source of concern. It can also encourage us to reflect on what sort of society we wish to build and how we might get there.

Acknowledgements Thanks to Owen Cotton-Barratt, Max Dalton, Raymond Douglas, Rose Hadshar, Elliot Thornley, and Philip Trammell. Thanks also to attendees of the Forethought Foundations Governing Explosive Growth Seminar and the AI Work-In-Progress Seminar of the Global Priorities Institute.

Author contribution This is a single authored paper, so all contributions are by A.B.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests I do not consider myself to have any meaningful competing interests. However, for the sake of complete transparency, this work was completed while employed at the Global Priorities Institute, which receives the majority of its funding via Good Ventures and the Open Philanthropy Project. As at time of writing, I also have £634.62 in shares, most of which is invested in technology companies.

Bibliographys

Acemoglu D, Johnson S (2023) Power and progress: our thousand-year struggle over technology and prosperity. Public Affairs

Acemoglu D, Restrepo P (2019) Automation and new tasks: how technology displaces and reinstates labor. J Econ Perspect 33(2):3–30

Acemoglu D, Restrepo P (2022) Tasks, automation, and the rise in U.S. wage inequality. Econometrica 90(5):1973–2016

Aghion P, Jones BF, Jones CI (2019) Artificial intelligence and economic growth. In: Agrawal A, Gans J, Goldfarb A (eds) The economics of artificial intelligence: an agenda. University of Chicago Press, pp 237–290

Agnello L, Castro V, Jalles JT, Sousa RM (2017) Income inequality, fiscal stimuli and political (in)stability. Int Tax Public Financ 24(3):484–511

Ahmed A, Aziz S, Toro CT, Alzubaidi M, Irshaidat S, Serhan HA, Abd-alrazaq AA, Househ M (2022) Machine learning models to detect anxiety and depression through social media: a scoping review. Computer Methods Progr Biomed Update 2:100066. https://doi.org/10.1016/j.cmpbup.2022.100066

Alesina A, Perotti R (1996) Income distribution, political instability, and investment. Eur Econ Rev 40(6):1203–1228

Anthropic (2024) The Claude 3 Model Family: Opus, Sonnet, Haiku. https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf

Arnold Z, Toner H (2021) AI accidents: an emerging threat (CSET Policy Brief). Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/ai-accidents-an-emerging-threat/

Assadi G (Unpublished) Will humanity choose its future? https://philpapers.org/archive/ASSWHC.pdf

Autor D (2022) The labor market impacts of technological change: from unbridled enthusiasm to qualified optimism to vast uncertainty (working paper 30074; working paper series). National Bureau of Economic Research. https://doi.org/10.3386/w30074

Bai H, Voelkel JG, Eichstaedt JC, Willer R (Unpublished) Artificial intelligence can persuade humans on political issues. https://osf.io/preprints/osf/stakv

Barnett M (2025) AGI could drive wages below subsistence level. Epoch AI's gradient updates. https://epoch.ai/gradient-updates/agi-could-drive-wages-below-subsistence-level

Bengio Y, Mindermann S, Privitera D, Besiroglu T, Bommasani R, Casper S, Choi Y, Fox P, Garfinkel B, Goldfarb D, Heidari H, Ho A, Kapoor S, Khalatbari L, Longpre S, Manning S, Mavroudis V, Mazeika M, Michael J, Zeng Y (2025) International AI safety report (DSIT 2025/001). https://www.gov.uk/government/publications/international-ai-safety-report-2025

Benn C, Lazar S (2022) What's wrong with automated influence. Can J Philos 52(1):125–148

Bonfanti ME (2022) Artificial intelligence and the offense-defense balance in cyber security. In: Cavelty MD, Wenger A (eds) Cyber security politics: socio-technological transformations and political fragmentation. Routledge, pp 64–77

Bostrom N (2003) Ethical issues in advanced artificial intelligence. In: Lasker GE, Marreiros G, Smit I, Wallach W (eds) Cognitive, emotive and ethical aspects of decision making in humans and in artificial intelligence. International Institute for Advanced Studies in Systems Research and Cybernetics, pp 12–17

Bostrom N (2009) Pascal's mugging. Analysis 69(3)

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

Carlsmith J (2021) Is power-seeking ai an existential risk? (Draft report). Open Philanthropy. https://arxiv.org/abs/2206.13353

Carlsmith J (Forthcoming) Existential risk from powerseeking AI. In: Thorstad D, Barrett J, Greaves H (eds) Essays on longtermism. Oxford University Press

Cazzaniga M, Jaumotte F, Li L, Melina G, Panton AJ, Pizzinelli C, Rockall E, Tavares MM (2024) Gen-AI: artificial intelligence and the future of work (IMF staff discussion note SDN/2024/001). IMF. https://www.imf.org/en/Publications/Staff-Discussion-Notes/Issues/2024/01/14/Gen-AI-Artificial-Intelligence-and-the-Future-of-Work-542379

Center for Security and Emerging Technology (2024) Emerging technology observatory country activity tracker: artificial intelligence. https://cat.eto.tech/

Christiano P (2019) What failure looks like. Less Wrong. https://www.lesswrong.com/posts/HBxe6wdjxK239zajf/what-failure-looks-like/

Colombo E, Mercorio F, Mezzanzanica M, Serino A (2024) Towards the terminator economy: assessing job exposure to AI through LLMs. https://arxiv.org/abs/2407.19204
Costello TH, Pennycook G, Rand DG (2024) Durably reducing conspiracy beliefs through dialogues with AI. Science 385(6714)
Critch A, Russell S (2023) TASRA: a taxonomy and analysis of societal-scale risks from AI. https://arxiv.org/abs/2306.06924
Department of Health and Social Care (2018) Securing cyber resilience in health and care: October 2018 update. https://www.gov.uk/government/publications/securing-cyber-resilience-in-health-and-care-october-2018-update
Diamond J (2005) Collapse: how societies choose to fail or succeed. Allen Lane
Durmus E, Lovitt L, Tamkin A, Ritchie S, Clark J, Ganguli D (2024) Measuring the persuasiveness of language models. https://www.anthropic.com/news/measuring-model-persuasiveness
Eloundou T, Manning S, Mishkin P, Rock D (2023) GPTs are GPTs: an early look at the labor market impact potential of large language models. https://arxiv.org/abs/2303.10130
Ernst M (2024) Identifying textual disinformation using large language models. In: Proceedings of the 2024 conference on human information interaction and retrieval, 453–456. https://doi.org/10.1145/3627508.3638315
Fang R, Bindu R, Gupta A, Zhan Q, Kang D (2024a) LLM agents can autonomously hack websites. https://arxiv.org/abs/2402.06664
Fang R, Bindu R, Gupta A, Zhan Q, Kang D (2024b) Teams of LLM agents can exploit zero-day vulnerabilities. https://arxiv.org/abs/2406.01637
Feldstein S (2019) The global expansion of AI surveillance. Carnegie Endowment for International Peace. https://carnegieendowment.org/research/2019/09/the-global-expansion-of-ai-surveillance
Fong KJ, Summers C, Cook TM (2024) NHS hospital capacity during covid-19: overstretched staff, space, systems, and stuff. BMJ 385. https://doi.org/10.1136/bmj-2023-075613
Goldstein JA, Sastry G (2023) The coming age of AI-powered propaganda: how to defend against supercharged disinformation. Foreign Affairs. https://www.foreignaffairs.com/united-states/coming-age-ai-powered-propaganda
Grace K, Stewart H, Sandkühler JF, Thomas S, Weinstein-Raun B, Brauner J (2024) Thousands of AI authors on the future of AI. https://arxiv.org/abs/2401.02843
Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC (2017) Detecting depression and mental illness on social media: an integrative review. Big Data Behav Sci 18:43–49
Hackenburg K, Margetts H (2024) Evaluating the persuasive influence of political microtargeting with large language models. Proc Natl Acad Sci 121(24):e2403116121
Hämäläinen P (2022) Indigenous continent: the epic contest for North America. Liveright
Handa K, Tamkin A, McCain M, Huang S, Durmus E, Heck S, Mueller J, Hong J, Ritchie S, Belonax T, Troy KK, Amodei D, Kaplan J, Clark J, Ganguli D (2025) Which economic tasks are performed with AI? Evidence from millions of Claude Conversations Jared, Jerry Hong, Stuart Ritchie, Tim Belonax, Kevin K. Troy Dario Amodei, Jared Kaplan, Jack Clark, Deep Ganguli. https://assets.anthropic.com/m/2e23255f1e84ca97/original/Economic_Tasks_AI_Paper.pdf
Happe A, Kaplan A, Cito J (2024) LLMs as hackers: autonomous linux privilege escalation attacks. https://arxiv.org/abs/2310.11409
Hazell J (2023) Spear phishing with large language models. The Centre for the Governance of AI. https://www.governance.ai/research-paper/llms-used-spear-phishing
Heiding F, Lermen S, Kao A, Schneier B, Vishwanath A (2024) Evaluating large language models' capability to launch fully automated spear phishing campaigns: validated on human subjects. https://arxiv.org/abs/2412.00586
Hendrycks D (2024) Introduction to AI safety, ethics and society. CRC Press
Henshall W (2024) When might AI outsmart us? It Depends who you ask. Time. https://time.com/6556168/when-ai-outsmart-humans/
Hill K (2023) Your face belongs to us: the secretive startup dismantling your privacy. Simon & Schuster UK
Ingham V, Islam MR, Hicks J, Lukasiewicz A, Kim C (2022) Definition and explanation of community disaster fatigue. In: Lukasiewicz A, O'Donnell T (eds) Complex disasters: compounding, cascading, and protracted. Springer Nature Singapore, pp 341–361. https://doi.org/10.1007/978-981-19-2428-6_17
Ingham V, Wuersch L, Islam MR, Hicks J (2023) Indicators of community disaster fatigue: a case study in the New South Wales Blue Mountains. Int J Disaster Risk Reduct 95:103831
Jakesch M, Hancock JT, Naaman M (2023) Human heuristics for AI-generated language are flawed. Proc Natl Acad Sci 120(11):e2208839120. https://doi.org/10.1073/pnas.2208839120
Kasirzadeh A (2025) Two types of AI existential risk: decisive and accumulative. Philosophical Stud. https://doi.org/10.1007/s11098-025-02301-3
Kent A. Clarke Center (2019) Inequality, populism, and redistribution. https://www.kentclarkcenter.org/surveys/inequality-populism-and-redistribution/
Korinek A, Juelfs M (2023) Preparing for the (non-existent?) future of work. In: Bullock JB, Chen Y-C, Himmelreich J, Hudson VM, Korinek A, Young MM, Zhang B (eds) The Oxford handbook of AI governance. Oxford University Press, pp 746–776
Korinek A, Stiglitz JE (2017) Artificial intelligence and its implications for income distribution and unemployment (working paper 24174; working paper series). National Bureau of Economic Research. https://doi.org/10.3386/w24174
Korinek A, Stiglitz JE (2021) Artificial intelligence, globalization, and strategies for economic development (working paper 28453; working paper series). National Bureau of Economic Research. https://doi.org/10.3386/w28453
Korinek A, Suh D (2024) Scenarios for the transition to AGI (working paper 32255; working paper series). National Bureau of Economic Research. https://doi.org/10.3386/w32255
Kulveit J, Douglas R, Ammann N, Turan D, Krueger D, Duvenaud D (2025) Gradual disempowerment: systemic existential risks from incremental AI development. https://arxiv.org/abs/2501.16946
Lehdonvirta V, Wú B, Hawkins Z (2024) Compute North vs. Compute South: the uneven possibilities of compute-based AI governance around the globe. Proc AAAI/ACM Conf AI Ethics Soc 7(1):828–838
Lucas J, Uchendu A, Yamashita M, Lee J, Rohatgi S, Lee D (2023) Fighting fire with fire: the dual role of llms in crafting and detecting elusive disinformation. In: Bouamor H, Pino J, Bali K (eds) Proceedings of the 2023 conference on empirical methods in natural language processing. Association for Computational Linguistics, pp 14279–14305. https://doi.org/10.18653/v1/2023.emnlp-main.883
Manheim D, Garrabrant S (2019) Categorizing variants of goodhart's law. https://arxiv.org/abs/1803.04585
Maslej N, Fattorini L, Perrault R, Parli V, Reuel A, Brynjolfsson E, Etchemendy J, Ligett K, Lyons T, Manyika J, Niebles JC, Shoham Y, Wald R, Clark J (2024) Artificial intelligence index report 2024. Institute for Human-Centered AI, Stanford University. https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI_AI-Index-Report-2024.pdf
Moll B, Rachel L, Restrepo P (2022) Uneven growth: automation's impact on income and wealth inequality. Econometrica 90(6):2645–2683

Mouton CA, Lucas C, Guest E (2024) The operational risks of AI in large-scale biological attacks: results of a red-team study (RRA2977-2). Rand Corporation
NCSC (2024) The near-term impact of AI on the cyber threat. The National Cybersecurity Centre. https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat
Newman S (2024) Cybersecurity and AI: the evolving security landscape. Center for AI Safety Blog. https://www.safe.ai/blog/cybersecurity-and-ai-the-evolving-security-landscape
Ngo R, Chan L, Mindermann S (2023) The alignment problem from a deep learning perspective. https://arxiv.org/pdf/2209.00626.pdf
Nii-Aponsah H, Verspagen B, Mohnen P (2023) Automation-induced reshoring and potential implications for developing economies (MERIT working papers 2023–018). Maastricht Economic and Social Research Institute on Innovation and Technology (MERIT). https://ideas.repec.org/p/unm/unumer/2023018.html
OECD (2023) OECD employment outlook 2023: artificial intelligence and the labour market. OECD Publishing. https://www.oecd.org/en/publications/oecd-employment-outlook-2023_08785bba-en
OpenAI (2024) OpenAI o1 system card. https://cdn.openai.com/o1-system-card-20240917.pdf
Ord T (2020) The precipice: existential risk and the future of humanity. Bloomsbury Publishing.
Peel M, Rodgers L, Arenas IT, Learner S, Williams J, Bott I (2024) Are the robots finally coming? Financial Times. https://ig.ft.com/ai-robots/
Phuong M, Aitchison M, Catt E, Cogan S, Kaskasoli A, Krakovna V, Lindner D, Rahtz M, Assael Y, Hodkinson S, Howard H, Lieberum T, Kumar R, Raad MA, Webson A, Ho L, Lin S, Farquhar S, Hutter M, Shevlane T (2024) Evaluating frontier models for dangerous capabilities. https://arxiv.org/abs/2403.13793
Posner RA (1997) Equality, wealth, and political stability. J Law Econ Organ 13(2):344–365
Rahman R, Owen D, You J (2024) Tracking large-scale AI models. https://epoch.ai/blog/tracking-large-scale-ai-models
Rogiers A, Noels S, Buyl M, Bie TD (2024) Persuasion with large language models: a survey. https://arxiv.org/abs/2411.06837
Roser M (2023) AI timelines: what do experts in artificial intelligence expect for the future? Our World in Data. https://ourworldindata.org/ai-timelines
Ryan-Mosley T, Richards S (2022) The secret police: cops built a shadowy surveillance machine in minnesota after george floyd's murder. MIT Technology Review. https://www.technologyreview.com/2022/03/03/1046676/police-surveillance-minnesota-george-floyd/
Salvi F, Ribeiro MH, Gallotti R, West R (2024) On the conversational persuasiveness of large language models: a randomized controlled trial. https://arxiv.org/abs/2403.14380
Shao M, Chen B, Jancheska S, Dolan-Gavitt B, Garg S, Karri R, Shafique M (2024) An empirical evaluation of LLMs for solving offensive security challenges. https://arxiv.org/abs/2402.11814
Simchon A, Edwards M, Lewandowsky S (2024) The persuasive effects of political microtargeting in the age of generative artificial intelligence. PNAS Nexus 3(2)
Smith M, Miller S (2022) The ethical application of biometric facial recognition technology. AI Soc 37(1):167–175
Solow-Niederman A (2022) Information privacy and the inference economy. Northwest Univ Law Rev 117(2):357–424
Spitale G, Biller-Andorno N, Germani F (2023) AI model GPT-3 (dis)informs us better than humans. Sci Adv 9(26)
Staab R, Vero M, Balunovic M, Vechev M (2024) Beyond memorization: violating privacy via inference with large language models. The twelfth international conference on learning representations. https://openreview.net/forum?id=kmn0BhQk7p
Stockwell S (2024) AI-enabled influence operations: threat analysis of the 2024 UK and European Elections [CETaS briefing paper]. The Alan Turing Institute. https://cetas.turing.ac.uk/publications/ai-enabled-influence-operations-threat-analysis-2024-uk-and-european-elections
Stockwell S, Hughes M, Swatton P, Bishop K (2024) AI-enabled influence operations: the threat to the UK General Election [CETaS briefing paper]. The Alan Turing Institute. https://cetas.turing.ac.uk/publications/ai-enabled-influence-operations-threat-uk-general-election
Stockwell S, Hughes M, Swatton P, Zhang A, Hall J, Kieran (2024) AI-enabled influence operations: safeguarding future elections [CETaS briefing paper]. The Alan Turing Institute. https://cetas.turing.ac.uk/publications/ai-enabled-influence-operations-safeguarding-future-elections
Strathern M (1997) 'Improving ratings': audit in the British University system. European Review 5(3):305–321
Strzyżyńska W (2022) Iranian authorities plan to use facial recognition to enforce New Hijab Law. The Guardian. https://www.theguardian.com/global-development/2022/sep/05/iran-government-facial-recognition-technology-hijab-law-crackdown
Susskind D (2022) Technological unemployment. In: Bullock JB, Chen Y-C, Himmelreich J, Hudson VM, Korinek A, Young MM, Zhang B (eds) The Oxford handbook of AI governance. Oxford University Press, pp 641–659
Tappin BM, Wittenberg C, Hewitt LB, Berinsky AJ, Rand DG (2023) Quantifying the potential persuasive returns to political microtargeting. Proc Nat Acad Sci 120(25)
Thomas RL, Uminsky D (2022) Reliance on metrics is a fundamental challenge for AI. Patterns 3(5)
Trammell P (2025) The ambiguous effect of full automation on wages. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/cKsknByhuW6Hw2wHj/the-ambiguous-effect-of-full-automation-on-wages
Trammell P, Korinek A (2023) Economic growth under transformative AI (working paper 31815; working paper series). National Bureau of Economic Research. https://doi.org/10.3386/w31815
Turchin P (2023) End times: elites, counter-elites and the path of political disintegration. Penguin
UK AISI (2024) Advanced AI evaluations at AISI: may update. The UK AI Safety Institute. https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update
US AISI, & UK AISI. (2024a) US AISI and UK AISI joint pre-deployment test: anthropic's claude 3.5 Sonnet (October 2024 Release). https://cdn.prod.website-files.com/663bd486c5e4c81588db7a1d/673b689ec926d8d32e889a8e_UK-US-Testing-Report-Nov-19.pdf
US AISI, & UK AISI (2024b) US AISI and UK AISI joint pre-deployment test: OpenAI o1 (December 2024). https://cdn.prod.website-files.com/663bd486c5e4c81588db7a1d/6763fac97cd22a9484ac3c37_o1_uk_us_december_publication_final.pdf
van Doorn M (2023) Advancing the debate on the consequences of misinformation: clarifying why it's not (just) about false beliefs. Inquiry
Véliz C (2020) Privacy is power: why and how you should take back control of your data. Transworld Digital
Wang J, Zhu Z, Liu C, Li R, Wu X (2024) LLM-enhanced multimodal detection of fake news. PLOS ONE 19(10)
Williams AR, Burke-Moore L, Chan RS-Y, Enock FE, Nanni F, Sippy T, Chung Y-L, Gabasova E, Hackenburg K, Bright J (2024) Large language models can consistently generate

high-quality content for election disinformation operations. https://arxiv.org/abs/2408.06731

Xu J, Stokes JW, McDonald G, Bai X, Marshall D, Wang S, Swaminathan A, Li Z (2024) AutoAttacker: a large language model guided system to implement automatic cyber-attacks. https://arxiv.org/abs/2403.01038

Zhang T, Schoene AM, Ji S, Ananiadou S (2022) Natural language processing applied to mental illness detection: a narrative review. NPJ Digital Med 5(1)

Zwetsloot R, Dafoe A (2019) Thinking about risks from AI: accidents, misuse and structure. Lawfare. https://www.lawfaremedia.org/article/thinking-about-risks-ai-accidents-misuse-and-structure

Zwetsloot R, Zhang B, Dreksler N, Kahn L, Anderljung M, Dafoe A, Horowitz MC (2021) Skilled and mobile: survey evidence of AI researchers' immigration preferences. In: Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society, 1050–1059