Philosophy of AI
Vol. 1, 2025, 86-92
10.18716/ojs/phai/2025.11844

AUTOMATION AND ITS DISCONTENTS: LEVERAGING AUTOMATION TO SAFEGUARD HUMANITY

Aksel Sterri, Peder Skjelbred

University of Oslo, NO

Abstract

This paper examines strategies for establishing a long-lasting ban on the development of advanced artificial intelligence (AI) to mitigate existential risk from AI. We evaluate Cappelen, Goldstein, and Hawthorne's proposal to leverage AI accidents as warning shots and find that it faces substantial ethical and practical challenges. As an alternative, we propose leveraging social unrest from rapid AI-driven automation. We argue that widespread job displacement could mobilise global labour movements against AI advancement, creating sufficient political pressure for an enforceable international ban on frontier AI research, and that this offers a more viable path to human survival in the face of advanced AI than current alternatives.

1. Introduction

To mitigate existential risk from AI, humanity may need more than incremental reform and technical AI safety work. Indeed, Cappelen, Goldstein, and Hawthorne (2025, hereafter CGH) provide a compelling argument that current AI safety work might paradoxically increase existential risk from AI by contributing to the advancement of the very systems that pose such risks.[1] The authors argue that to ensure humanity's survival, we may need a permanent ban on advanced artificial intelligence AI research. This paper explores strategies for achieving such a goal.

We first explicate CGH's critique of the prevailing AI safety paradigm and evaluate their proposed alternative: allowing and leveraging AI accidents as warning shots to gain support for a ban on advanced AI research. While this strategy has intuitive appeal, we identify substantial ethical and practical challenges that count against its implementation.

Building on CGH's core insight about leveraging negative AI impacts to gain support for a long-lasting ban on advanced AI research, we propose an alternative strategy: harnessing social unrest from rapid AI-driven automation. Rather than waiting for potentially catastrophic AI accidents, which might not come in time, we argue that widespread job displacement could mobilise labour movements worldwide against AI advancement, creating sufficient political pressure for what CGH call a 'cultural plateau' in AI development, where a ban on risky AI development is successfully implemented and enforced.[2]

Our analysis expands the strategic possibilities in AI governance by suggesting that the

[1] References to pages are, if not otherwise specified, to CGH.

[2] Cultural plateaus serve as an enabler for two of the other survival strategies discussed in CGH, namely successfully aligning advanced AI or successfully keeping oversight and control over them, in virtue of buying more time for safety research. 'Bans' should therefore be understood broadly, including temporary moratoria that enable these other strategies. If alignment and control are reliable, they would ultimately be much more desirable than a blanket ban on research into powerful AI.

most effective path to human survival in the face of AI may run through worker movements.

2. Accident Prevention and Accident Leveraging

A 'strategy' will here be understood as a set of actions designed to effect policy change to achieve one's ultimate goal. 'Policies' are the set of collective actions undertaken by either companies or governments; in this case, either ambitious AI safety technologies or a ban on research into powerful AI. 'Actions' are the set of behaviours that a group of agents is to undertake to get those policies into effect. A strategy should be assessed based on its moral desirability and effectiveness.

In our discussion, we adopt a somewhat idealised approach to the question of which agents, or set of agents, should follow the strategies under consideration. The 'AI safety community' should be thought of as an approximation of a unitary agent that perfectly or nearly perfectly complies with the prescribed strategy. This is warranted, we think, because understanding optimal collective actions provides useful guidance for addressing the more complex question of what individual agents should do under conditions of partial compliance (Acemoglu & Restrepo, 2019).

Current AI safety work could be conceived of as a strategy of 'accident prevention,' in CGH's terminology. It includes technical work to align AI systems with human values, using techniques such as reinforcement learning with human feedback (RLHF), as well as legal restrictions on dangerous or risky use, such as the EU AI Act. People might engage in such work for different reasons, but our focus here is whether accident prevention can help prevent an existential catastrophe.

AI accident prevention is, on its face, a promising strategy for reducing existential risk. Much of it is aimed at preventing harm while allowing advances in AI. If successful, this will enable societies to reap the benefits of AI while preventing adverse consequences. It is scalable in that preventing minor accidents provides insight into how to prevent catastrophic accidents as well, and, for that reason, it is likely to be effective.

Still, accident prevention is not without flaws. One concern is that theoretical work on AI safety might accelerate AI development, thereby increasing the risk of catastrophic outcomes. RLHF and other alignment techniques have arguably been crucial for making AI systems reliably useful to users, and therefore for achieving commercial success.[3]

CGH present a related backfiring argument. They argue that accident prevention, by reducing the number and severity of accidents that could serve as warning shots to wake up a complacent public, might lower 'the chance of banning capability-improving AI research' and thus increase the existential risk from AI (p. 20). The argument is intuitively appealing. It is unlikely that people will support a ban on advanced AI research unless they can see the harm from AI. Without significant accidents, people are likely to become complacent. If a ban is necessary to ensure humanity's survival, the accident prevention strategy might, in fact, make things worse.

CGH offer an alternative strategy to reduce existential risk from AI. It might be better, they argue, to allow accidents to happen, and instead, work to leverage a 'given accident (or near accident)' to ban 'research into potentially dangerous AI systems' (p. 20). This strategy is, to our knowledge, novel in the context of AI-related existential risks, but familiar to researchers and practitioners of social movements. Whether one should prevent or leverage harm is perhaps the central question facing social movements from the labour movement and civil rights movements to the animal liberation movement.[4] Incremental reforms may

[3] For a nuanced but critical overview of this argument, see Christiano (2023).

[4] Should socialists work against social democratic reforms to avoid pacifying the working class by patches to the capitalist system (despite these reforms' likely beneficial effects on the lives of the working class)? Similarly, should civil rights activists take an incremental or a revolutionary approach, fearing that incremental reforms may make the movement lose momentum? Finally, should animal abolitionists oppose animal welfare reforms that improve the lives of animals but also risk putting a friendly face on a practice that should be abolished?

improve outcomes in the short run, but risk dampening the movement's revolutionary spirit.

Enforcing a ban on risky AI research requires securing sufficient support to implement it and the willingness to sustain enforcement over time. This demands, at the very least, sustained political urgency or moral consensus that advanced AI is deeply dangerous. CGH argue that accidents can play this role, as the wakeup call that will unite the world around a ban (p. 10). This is more likely to occur the more severe and frequent the accidents are, at least until the point at which the accidents create chaos and panic rather than support. Accident leveraging strategies seem more promising than accident prevention strategies on their face.

Still, there is a significant ethical hurdle that must be overcome. It is doubtful that it is permissible to employ a strategy that will significantly increase the likelihood of serious harm from accidents in order to rally support for a long-term goal that may or may not come about and that may or may not be necessary for preventing catastrophic harm in the future. However, this is precisely what must be the case for the accident leveraging strategy to get off the ground, and CGH does not offer an argument for it.

Beyond the ethical issues, we have reason to question the effectiveness of the accident leveraging strategy. As CGH (p. 10-11) admit, accidents are unpredictable, and their meaning is often subject to debate. If a powerful AI system malfunctions and kills a small number of people, those who profit from AI might frame it as a fixable technical glitch. Conversely, if a near-existential accident were to occur, it might be too late to effectively rein in the technology. The 'warning shot' must therefore strike the perfect balance: alarming enough to trigger collective action, but not so grave as to cause irreparable damage.

The biggest problem with the accident leveraging strategy, however, is that we might not even see any serious warning shots. Indeed, early AI safety theorists, such as Yudkowsky (2017), have argued that what makes AI risk uniquely difficult to tackle is precisely that there may be no 'fire alarm' for dangerous AI. The AI models that actually pose an existential risk may be deceptively aligned up until they are sufficiently powerful, at which point they may take a 'treacherous turn' (Bostrom, 2014, chapter 8). In worlds where there are no warning shots, accident leveraging strategies are futile.

Thus, there are significant obstacles to the accident leveraging strategy. In the next section, we propose an alternative leveraging strategy: harnessing social unrest arising from rapid AI-driven automation. This approach is less vulnerable to the objections above. AI is likely to cause widespread automation or rational expectations of widespread automation. Indeed, it is arguably already underway (Brynjolfsson et al., 2025). Extensive job losses are harder to dismiss as fixable glitches, and the ethical concerns, though present, are less severe, as we discuss in Section 4. This offers a more promising path to a ban on advanced AI research.

3. Automation and Its Discontents

If AI continues to advance, it will replace workers in many tasks in a wide range of sectors. There is disagreement among economists about the long-run effects of AI-driven automation on labour markets and wages.[5] However, if replacement happens *rapidly*, this will likely lead to significant unemployment and welfare losses. Reallocation of labour takes time due to both *frictional* and *structural* unemployment; temporary unemployment during job searches and longer-lasting unemployment resulting from periods of retraining after worker skill sets are obsolete, respectively.[6] These effects might be amplified through several psychological mechanisms, most notably a form of *anticipatory helplessness*, whereby workers reason 'why bother retraining and reapplying for jobs if AI will automate these

[5] See Acemoglu (2024) and Korinek and Suh (2024).

[6] Acemoglu and Restrepro (2019) show that automation has created significant unemployment and welfare losses in the recent period in the United States because firms have been better at using technology to displace rather than reinstate labour.

jobs before I am in a position to perform them?'[7].

This process of rapid deskilling and unemployment could lead to widespread anger and radicalisation, as it has historically with the Luddite rebellion and other movements fighting against labour displacement (Mokyr 1990). Such anger is a potent force for political change. Rather than relying on AI accidents or near-disaster to instil fear, the mounting evidence of automation's social costs could be exploited to mobilise people against the development and deployment of powerful AI systems.

However, cross-border worker solidarity relies on automation functioning as a rapid, exogenous shock; gradual displacement, by contrast, risks forestalling mobilisation. A slower transition allows the reinstatement of labour in new sectors and enables elites to manage dissent through incremental concessions. Crucially, it fractures collective identity: rather than perceiving themselves as victims of a shared external crisis, atomised workers may internalise displacement as a personal failure to upskill.

In the short run, this might, somewhat paradoxically, suggest supporting accelerationist policies and opposing policies that would reduce the speed and impact of automation.[8] Workers need to feel the pain to rise to the occasion, and measures that soften the costs of automation might prevent mobilisation. When automation, or rational expectations of it, has become extensive enough, the moment must be seized to channel public anger into political demands for restricting AI development and deployment.

This strategy has several advantages. By focusing on economic impacts rather than technical failures, it engages broader constituencies than typical AI safety advocacy does. Labour unions and economic justice organisations are obvious candidates. However, one might also expect support from others who benefit from the current order, such as employer organisations and small- and medium-sized business owners, provided they face substantial financial risks from rapid AI displacement. These groups, who might otherwise remain disengaged from AI safety discussions, become natural allies when AI risk is framed in terms of employment and financial stability.

In addition, such movements might be better equipped to coordinate action across states than AI safety organisations consisting of a small elite. In the face of a sufficiently pressing challenge, labour movements would be expected to coordinate across borders to demand that governments sign and enforce global bans. If unemployment spikes worldwide, or if workers in lagging countries anticipate the trend, a significant portion of workers could be mobilised against frontier AI research.

By the time the employment crisis is undeniable, one might find broad support for the claim that the safest course is to freeze AI at a level where it cannot further imperil human employment, thereby incidentally protecting against the existential risk posed by extremely powerful AI systems.

4. The Risks of Permitting Large-Scale Automation

The approach we have just outlined depends on accelerating, or, at a minimum, allowing, large-scale AI-driven automation in order to create a backlash against AI development. This strategy is not without problems.

One concern is that it may be impermissible to cause, or even to allow, serious harm to prevent highly uncertain but catastrophic outcomes, mirroring the criticism of the accident leveraging strategy.

A second concern is that AI-driven automation could be an *existential risk multiplier* through at least three channels. First, it might multiply risks from other AI risk channels, most notably those associated with the misuse of general-purpose AI technology by mali-

[7] Anticipatory helplessness plays on the concept of 'learned helplessness' in psychology. Learned helplessness is a state of stressful inaction due to an unpredictable environment, and is a prominent explanation of depression (Seligman, 1972).

[8] Examples of such policies are labour protections, taxes on AI deployment, and income transfers such as a universal basic income, which dampen social costs.

cious actors. Misuse risk is a function of two factors: the availability of the technology and the number of skilled people who are willing to use it. Rapidly increasing unemployment, especially among the highly educated, is a plausible catalyst for increasing the number of skilled people willing to perform atrocities (Benmelech et al., 2012; Urdal, 2006). Second, rapid increases in unemployment and economic instability might also produce a fascist or authoritarian backlash (Ballard-Rosa et al., 2017; King et al., 2008). If so, this reduces the probability of achieving the cooperation necessary to solve collective action problems; that is, it lowers the likelihood that cultural plateau strategies work, thereby indirectly increasing AI risk. Finally, by causing rising geopolitical tensions and heightened national instability, allowing widespread automation might increase the likelihood of non-AI existential risks, such as nuclear war and engineered pandemics (Caldara & Iacoviello, 2022).

Given these risks, one might be tempted to pursue a preventive strategy instead: rather than allow for rapid automation to generate backlash, seek policies that slow or moderate its pace. This would avoid the existential-risk-multiplying effects just described while leaving open other paths to reducing existential risk, such as accident leveraging or direct advocacy for a ban.

We have two responses to this objection. First, even if such a strategy were desirable, it is unlikely to be feasible. For reasons mentioned above, it is challenging to build a coalition against AI before there is a clear threat to workers. Furthermore, even if a preventive strategy were possible, it would likely be dangerous. A ban on AI *deployment* would, at best, be a temporary victory with weak popular support. Crucially, it is unlikely to prevent further research into powerful and dangerous AI systems. Investors biding their time could fund such research while working to circumvent the ban or turn public opinion against it. As these systems become more powerful, the temptation to deploy them will grow, making any ban increasingly difficult to sustain. At the point of sudden deployment following the lifting of a ban, you would expect a substantial increase in risk from several sources. These capable AI models would themselves pose a significant existential risk, and arguably a greater risk than if they were deployed and tested gradually in the real world. A sudden implementation of highly capable models would also result in a much more dramatic increase in unemployment than in a world where automation occurred more gradually, potentially leading to extreme social unrest.

Let us now consider whether the automation leveraging strategy is permissible. We grant that such strategies are, at the very least, ethically dubious.[9] However, it is less problematic than accident leveraging, at least in this case. Automation is typically considered a permissible means of harming people, in large part because it offers to benefit almost everyone in the long run. Indeed, automation is a crucial reason for the staggering increases in economic welfare achieved since the Industrial Revolution. A standard assumption in business ethics is therefore that firms may replace workers with machines whenever that is profitable and basic procedural requirements are satisfied, such as providing workers with due notice (Heath, 2014). Allowing harm to workers through automation is therefore morally superior to explicit accident leveraging, which is closer on the spectrum to engaging in terrorist activities to create the necessary fear among the population to achieve otherwise laudable objectives.

Still, there is a delicate balance to be struck: one needs to permit enough automation for labour's alarm bells to ring, but not so much that society plunges into chaos or authoritarian rule. Along current margins, this balance may not be that difficult to find in practice. Evidence from recent technological advances suggests that accelerating automation may be the appropriate strategy for a considerable time. Implementation of AI in businesses up until now seems to have been quite slow despite evidence to suggest that current state-of-the-

[9] Reflecting on the ethical problems associated with AI survival stories can lead one to question the importance of humanity continuing to exist, as opposed to AI descendants with moral status carrying on the torch. In this paper we say nothing on the all-things-considered moral case for any strategy but merely take for granted the goal of avoiding an existential catastrophe for humans. For discussion, see Hong (2025).

art AI models are partial substitutes for large parts of knowledge work.[10]

The necessity of maintaining credibility acts as a check on any efforts by the AI safety community to hasten automation. It is difficult to convincingly leverage an AI accident for safety purposes if you have done nothing to prevent it. Similarly, it will be difficult to organise a worker uprising against capable AI if you have done nothing to prevent automation. For such a policy of accelerationism and leveraging to be possible, one would have to engage in a clear division of labour within the AI safety community, between accelerationists triggering a backlash on the one hand and, on the other, coalition builders who exploit the backlash to unite with labour friendly organisations to achieve a ban on advanced AI research. Whether such coordination is feasible without the strategy becoming publicly known, and thus self-undermining, remains an open question."

5. Objections and conclusion

One objection against the automation leveraging strategy is that it will come too late. AI systems will be dangerous long before they are deployed and integrated widely enough in the economy for a labour backlash to occur. There are many barriers to automation, but fewer barriers against developing highly capable models that might threaten humanity's survival.

In response, we want to stress that the enormous investments necessary to produce highly capable AI models require a valid business case. Unless investors can expect their investments to be profitable, money will dry up. It is thus plausible that AI development requires an iterative approach, in which AI companies launch products to demonstrate to investors that subsequent investments are likely to be profitable. If this iterative model holds, we might expect rapid automation to occur before models threaten humanity's survival through misalignment.

In addition, it is worth considering how a labour-led cultural plateau would interact with other aspects of the CGH taxonomy. If labour enforces a plateau for decades, alignment research might be stalled. Even the science of AI alignment could be stigmatised or banned, out of fear that it would lead inevitably to the kinds of advanced systems that displace workers. On the other hand, labour movements may be able to discriminate between 'alignment research' and 'capability research,' ensuring that alignment progress continues while top-end capabilities remain frozen. That, in turn, might yield the best-case outcome: a gradually improving alignment science, awaiting the day when the ban can be safely lifted. At this point, AI may be sufficiently advanced without posing existential risk or mass unemployment.

In sum, labour mobilisation arising from the threat of large-scale automation offers a plausible, if complex, route to establishing a cultural plateau for AI. It neither guarantees success nor avoids substantial social risks. Yet it speaks directly to the persistent question of how to build the political force to overcome powerful incentives that push AI research ever further. Whether one finds this scenario attractive or unsettling, it hopefully represents a useful addition to the set of strategic possibilities in AI existential risk discourse. Existential safety measures will likely arise not from rational deliberation about risk alone, but when social and economic pressures are channeled into a demand that leaders can no longer afford to ignore.

6. Bibliography

Acemoglu, D. (2024). *The Simple Macroeconomics of AI*. National Bureau of Economic Research. https://doi.org/10.3386/w32487

Acemoglu, D., & Restrepo, P. (2019). Automation and New Tasks: How Technology Displaces and

[10] Google DeepMind's Gemini 3 Pro Preview scores 38 percent on Humanity's Last Exam, a test consisting of extremely difficult

Reinstates Labor. *Journal of Economic Perspectives*, *33*(2), 3–30. https://doi.org/10.1257/jep.33.2.3

Benmelech, E., Berrebi, C., & Klor, E. F. (2012). Economic Conditions and the Quality of Suicide Terrorism. *The Journal of Politics*, 74(1), 113–128. https://doi.org/10.1017/S0022381611001101

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

Brynjolfsson, E., Chandar, B., & Chen, R. (2025). *Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence* [Working paper]. https://digitaleconomy.stanford.edu/publications/canaries-in-the-coal-mine/

Caldara, D., & Iacoviello, M. (2022). Measuring Geopolitical Risk. *American Economic Review*, *112*(4), 1194–1225. https://doi.org/10.1257/aer.20191823

Cappelen, H., Goldstein, S., & Hawthorne, J. (2025). AI Survival Stories: a Taxonomic Analysis of AI Existential Risk. *Philosophy of AI*, *1*, 1–18. https://doi.org/10.18716/ojs/phai/2025.2801

Center for AI Safety, & Scale. (2025). *Humanity's Last Exam*. https://agi.safe.ai/

Christiano, P. (2023). *Thoughts on the impact of RLHF research*. https://www.alignmentforum.org/posts/vwu4kegAEZTBtpT6p/thoughts-on-the-impact-of-rlhf-research

Estlund, D. (2019). *Utopophobia: On the Limits (If Any) of Political Philosophy*. Princeton University Press.

Heath, J. (2014). *Morality, Competition, and the Firm: The Market Failures Approach to Business Ethics*. Oxford University Press.

Hong, F. (2025). Group prioritarianism: why AI should not replace humanity. *Philosophical Studies*, *182*(7), 1705–1723. https://doi.org/10.1007/s11098-024-02189-5

Korinek, A., & Suh, D. (2024). *Scenarios for the Transition to AGI*. National Bureau of Economic Research. https://doi.org/10.3386/w32255

Seligman, M. E. P. (1972). Learned Helplessness. *Annual Review of Medicine*, 23(Volume 23, 1972), 407–412. https://doi.org/https://doi.org/10.1146/annurev.me.23.020172.002203

Urdal, H. (2006). A Clash of Generations? Youth Bulges and Political Violence. *International Studies Quarterly*, *50*(3), 607–629. https://doi.org/10.1111/j.1468-2478.2006.00416.x

Yudkowsky, E. (2017, October 13). *There's No Fire Alarm for Artificial General Intelligence - Machine Intelligence Research Institute*. https://intelligence.org/2017/10/13/fire-alarm/

questions from widely different disciplines (Center for AI Safety & Scale, 2025).