One harmful effect of AI is its potential use in surveillance and control.

Some AI ethics-focused groups and thinkers have raised concerns about contemporary AI applications such as facial recognition and predictive policing being used to exert social control, especially targeting marginalized communities. These capabilities are expected to increase in the future, and some civil liberties organizations such as the EFF have been reporting on these uses of AI in both democracies and autocratic regimes.

Security expert Bruce Schneier has argued that AI is enabling a shift from general surveillance (e.g., pervasive use of CCTV) to personalized surveillance of any citizen. For instance, AI can quickly and cheaply search all phone calls and CCTV footage in a city to form a detailed profile on one individual, which was previously only possible through laborious human effort.[1] Furthermore, traditional spying could only gather information *after* the target was identified as a person of interest, whereas the combination of AI and mass recording allows for the inspection of a target's behavior in the past.

In the future, more powerful AI surveillance, along with other AI-enabled technologies like autonomous weapons, might allow authoritarian or totalitarian states to make dissent virtually impossible, potentially enabling the rise of a stable global totalitarian state.

As of early 2024, access to the most advanced models is moderated through API access by Western corporations, which allows these corporations to restrict uses of their models that they do not condone. These corporations are incentivized not to collaborate with totalitarian governments, or to authorize use of their models by projects perceived as authoritarian, lest they face public backlash. As capabilities increase and powerful models become more accessible to smaller actors[2], this state of affairs might change.

A number of prominent researchers who mainly focus on risks from misalignment (rather than misuse) nevertheless view AI-enabled surveillance as one of the most salient risks that could arise from near-term AI:

- Daniel Kokotajlo has speculated that LLMs could be used as powerful persuasion tools to disproportionately aid authoritarian regimes.
- Nick Bostrom has discussed the potential incentives for widespread surveillance systems augmented by AI, based on state responses to concerns about living in an extremely risky and "vulnerable" world.
- Buck Shlegeris claims that risks of AI-enabled totalitarianism are "at least 10% as important as the risks [he] works on as an AI alignment researcher".
- Richard Ngo has claimed that outsourcing the task of maintaining control (e.g. through surveillance) to AI makes it easier to consolidate power, which in the limit leads to authoritarianism.

While it is important to undertake measures to mitigate these kinds of risks of AI misuse, it's not

[1] Schneier calls this a shift from "mass surveillance" to "mass spying", although other authors do not separate these two categories.
[2] This could happen for instance when powerful models are open-sourced.

sufficient; even well-intentioned actors have the potential to accidentally pose an existential risk when deploying AGI because of misalignment.

Alternative phrasings

- Isn't the real concern AI-enabled authoritarianism?
- Isn't the real concern AI-enabled totalitarianism?

Related

- Isn't the real concern X?
- Isn't the real concern AI being misused by terrorists or other bad actors?
- Isn't the real concern misuse?
- Wouldn't slowing down or stopping AI require an invasive global surveillance …
- What can we do to prevent bad actors from using AI for bad?
- How will we stop dangerous rogue AI if we don't develop our own safe superin…
- Is large-scale automated AI persuasion and propaganda a serious concern?

Scratchpad

Generated by AI, probably by stampy

> The concern about totalitarianism intersects with AI alignment because the development and deployment of AI systems can have significant societal implications. If AI systems are designed with values that prioritize control and suppression, they can exacerbate the risk of totalitarianism. Additionally, the increasing power and influence of AI systems may lead to a shift in decision-making authority from humans to AI algorithms, potentially undermining democratic processes and diminishing human control over governance.
>
> It is important to note that the risks associated with totalitarianism and AI alignment are not mutually exclusive. Addressing AI alignment is crucial to ensure that AI systems are aligned with human values and do not contribute to the concentration of power or the suppression of individual freedoms. At the same time, efforts to prevent totalitarianism and protect democratic principles should consider the potential impact of AI technologies and the need for safeguards against their misuse.
>
> In summary, while totalitarianism facilitated by AI is a significant concern, it is closely intertwined with the issue of AI alignment. Both aspects need to be addressed to mitigate the risks associated with the development and deployment of AI systems.
>
> Sources:

- Some thoughts on risks from narrow, non-agentic AI - Richard Ngo - 2021-01-19
- National Power After AI - Matthew Daniels, Ben Chang, Igor Mikolic-Torreira, James Baker, Jack Clark, Remco Zwetsloot, Teddy C
- 2012 Robin Hanson comment on "Intelligence Explosion: Evidence and Import" - Rob Bensinger - 2021-04-02
- [AN #59] How arguments for AI risk have changed over time - Rohin Shah - 2019-07-08

Notes: Old version mid 2023

As with the other concerns listed here, AI enabled totalitarianism is a valid concern. This concern is similar in many ways to the use of AI by bad actors such as terrorists, and as such, the reasons why this website does not concentrate on it are similar.

Basically, we currently don't know how to reliably steer AGI systems towards *any goal at all*. In other words, the problem of outer alignment is the problem of specifying a reward signal which captures your intended goals. This is actually pretty hard. We discuss the difficulty of outer alignment a bit in 'Why work on AI safety early?', but, in short, AI systems will be trained by powerful optimization processes, and there's no guarantee that AIs will end up actually being aligned, rather than being deceptively aligned as a way to achieve their long-term goals.

On this view, concerns about AI-enabled totalitarianism run (at best) a distant second to the issue of training AIs to learn *any* sort of goal we might want to specify. If you're concerned about AI-enabled totalitarianism, such regimes will, in effect, have to solve the problem of outer alignment. If outer alignment is as hard as some claim, we're very unlikely to get AI-enabled totalitarianism, but this is little comfort: instead, we'll very likely end up dead.

Perhaps counter-intuitively, *optimism* about outer alignment should move you towards *pessimism* about AI-enabled totalitarianism. If we make steady headway on outer alignment, we can train AIs to achieve specific goals. But that doesn't mean we'll like whatever goals powerful actors use to train advanced AIs.

Unfortunately, work on the risks of AI-enabled totalitarianism is pretty sparse. Daniel Kokotajlo has speculated on the potential use of LLMs as powerful persuasion tools to disproportionately aid authoritarian regimes, and Nick Bostrom has discussed the potential incentives for widespread surveillance systems augmented by AI, based on state responses to concerns about living in an extremely risky and 'vulnerable world'.

We prefer not to label issues in terms of whether or not they are 'the real concern'. AI-enabled totalitarianism could be a very big concern, depending on your views (among other things) about the difficulty of outer alignment.

So, just how big is the concern of AI-enabled totalitarianism? It's hard to say with a great deal of precision. But, for what it's worth, Buck Shlegeris (CTO of Redwood Research, an AI alignment organization) is reasonably worried, and claims that risks of AI-enabled totalitarianism are "at least 10% as important as the risks [he] works on as an AI alignment researcher".

See also:

- Carl Shulman's thoughts on technology-enabled totalitarianism