Notes on AI Risk

Zachary Goodsell

February 18, 2026

- I, like most people, have very little political power: if I want things to change in a certain way, there is usually not much I can do about it. However, I am protected from domination and oppression by belonging to classes of people who collectively have significant power. For the most part, a government or corporate policy that would be disastrous for me would also be bad for most people. It would therefore, likely, be opposed by a group of people whose collective power is sufficient to prevent the policy from being implemented. Call this the “protection of collective power”

- In democratic societies, this protection is reinforced by voting power, where policies I disfavour will, on average, be opposed by a majority of voters. Of course some policies are controversial, but the much broader range of almost complete agreement is made invisible by its non-controversiality.

- This form of protection is a powerful force even in the most repressive societies. In feudal times, the peasantry was protected from catastrophic policies by their utility to those in power. Even the most brutal lord has an interested in keeping his peasants alive, productive, and happy. Even in cases where the peasantry is sufficiently under control that there is no hope for any sort of uprising, the peasantry is still protected by their utility. Their collective power, if only to generate wealth, gives the lord a significant incentive to keep them alive and productive.

- The protection of collective power seems more fundamental than more recent protections I enjoy, stemming from democracy, the rule of law, and respect for human rights. The reason that it is very difficult to overturn these more recent protections is, in part, that they are supported by the protection of collective power.

- Collective power may be split into two forms: economic power and coercive power. Economic power is the power to move and create resources (e.g., minerals, factories, food, energy), and coercive power is the power to credibly threaten harm in order to disincentivise certain actions. Of course this is not a totally clear distinction.

- Artificial Intelligence threatens to undermine the protection of collective power in both forms. However, I suppose the primary threat is to economic power. In a world where all research, development, and production is automated, it is very hard to see where people like me contribute. Getting a human like me to do something will, by and large, be an inefficient way of going about things.

- For this reason, I regard a primary threat from AI to be the removal of my protection from collective power. If the presence of people like me is not necessary for production, then the powerful incentive to promote the flourishing of people like me vanishes.

- Of course, the AI that replaces my labour might be under benevolent control (e.g., by a benevolent superintelligence, a democratic government, or a benevolent corporation).

- There are, obviously, serious questions about the probability of a benevolent AI overlord. But what we would need is not just a benevolent overlord but a *robustly* benevolent overlord, who would still feel constrained to keep us happy if this became costly, e.g., in the event of an external catastrophe, or in the event of violent human resistance. Without the protection of collective power, it becomes more difficult to ensure any baseline level of benevolence, to even meet the "benevolence" of a feudal lord who wishes to keep his peasants productive.