Existing Policy Proposals Targeting Present and Future Harms

June 5, 2023

The Center for AI Safety exists to reduce societal–scale risks associated with AI. We will soon release policy recommendations to mitigate the future impacts of AI systems, especially harms from increasingly capable general–purpose models. At the same time, we believe many existing policy proposals designed to address the present impacts of AI systems would also promote AI safety if implemented. There is considerable overlap between the regulatory objectives of safety researchers and those of researchers concerned with the fairness, accountability, transparency, and ethics of AI systems. Building consensus around regulatory proposals can help to effectively promote their implementation, thereby both mitigating the present harms of AI systems and preventing future societal–scale risks.

The purpose of this one–page document is to describe three existing policy proposals which we believe promote AI safety. This list is nonexhaustive.

1. **Legal liability for AI harms.** In its policy brief on general–purpose AI (GPAI), the AI Now Institute argues for prioritizing improved legal liability frameworks for the harms caused by AI systems: “Any regulatory approach that allows developers of GPAI to relinquish responsibility using a standard legal disclaimer would be misguided. It creates a dangerous loophole that lets original developers of GPAI (often well resourced large companies) off the hook, instead placing sole responsibility with downstream actors that lack the resources, access, and ability to mitigate all risks.” Improved legal liability frameworks could help redress harms related to misinformation and algorithmic bias. They also incentivize the companies designing AI systems to take safety concerns as a priority rather than as an afterthought.

2. **Increased regulatory scrutiny.** The same policy brief argues for increased regulatory scrutiny targeting the development of AI systems: “GPAI must be regulated throughout the product cycle, not just at the application layer... the companies developing these models must be accountable for the data and design choices they make.” Transparency and regulations targeting training data can help to combat algorithmic bias and prevent companies from profiting from copyrighted materials without compensating those who produced them. From a safety perspective, regulatory mechanisms targeting training data may also help to prevent harms arising from goal misspecification. Increased oversight surrounding training data can ensure that data is responsibly sourced, slowing the dangerous pace of AI capabilities research.

3. **Human oversight over automated systems.** The European Union’s proposed AI Act emphasizes the importance of human oversight in the deployment of high–risk AI systems, and in particular “the possibility for a human to intercede in order to override a decision or recommendations that may lead to potential harm.” Human oversight can help to mitigate concerns about algorithmic bias and the propagation of false or misleading information though AI systems. Keeping humans in the loop can also help to detect and disable hazardous AI systems before they have the potential to cause harm.