How AI Progress Can Coexist With Safety and Democracy

Yoshua Bengio and Daniel Privitera outline policy goals for AI progress, safety, and democratic participation

7 min. read · View original

Dive into current discussions about how to regulate artificial intelligence, and you'd think we're grappling with impossible choices: Should we choose AI progress or AI safety? Address present-day impacts of AI or potential future risks? And many more perceived dilemmas. But are these trade-offs real? As world leaders prepare to gather at the upcoming AI Safety Summit in Bletchley Park in the U.K. in early November, let's dig a bit deeper and uncover three core values that underpin most policy proposals in the AI regulation discourse.

The first value is progress. The promises of AI are vast: curing diseases, increasing productivity, helping to solve climate change. This seems to call for a "full steam ahead" approach, in which we attempt to accelerate AI progress even beyond the current level. But moving at breakneck speed comes with increased risks—epidemics of automated fake news, AI enhanced bio terrorism, automated cyber warfare, or out-of-control AI threatening the existence of humanity. We are not currently prepared to handle these risks well. We don't know how to reliably control advanced AI systems, and we don't currently have mechanisms for preventing their misuse.

The second core value in AI regulation is therefore safety. Leading experts as well as the general public are increasingly concerned about extreme risks from AI, and policy-makers are rightly beginning to look for ways to increase AI safety. But prioritizing safety at all costs can also have undesirable consequences. For instance, it can make sense from a safety perspective to limit the open-sourcing of AI models if they can be used for potentially dangerous purposes like engineering a highly contagious virus. On the flipside, however, open-source code helps to reduce concentration of power. In a world with increasingly capable AI, leaving this rapidly growing power in the hands of a few profit-driven companies could seriously endanger democratic sovereignty. Who will decide what very powerful AI systems are used for? To whose benefit or to who's detriment? If superhuman capabilities end up in a few private hands without significant democratic governance, the very principle of sharing power that underlies democracy is threatened.

That is why the third core value in AI regulation is democratic participation. There is a real concern that AI might entrench existing power imbalances at the expense of marginalized groups, low income countries, and potentially everyone but a handful of tech giants building the most powerful AI models. This suggests we need to ensure continued participation from everyone in the future of AI. But focusing exclusively on participation would also come at a cost. Democratizing access to potentially highly destructive technology can lead to catastrophic outcomes, which is why access to certain tech in sectors like nuclear energy or pathogen research is also not democratized, but highly restricted and regulated.

How do progress, safety, and participation interact in practice, then? The "Transformative Technology Trilemma," a framework developed by the Collective Intelligence Project, suggests that discourse about these values is particularly susceptible to an either-or kind of thinking. And unfortunately, public discourse around AI can indeed convey a sense that we have to choose between three mutually incompatible AI futures: either AI Progress, or AI Safety, or AI with Democratic Participation. What a tragic trilemma! But this picture is not accurate. And it can lead to a dangerous, well-studied phenomenon called false polarization, in which people perceive disagreements as more profound than they actually are. This can have grave consequences – ironically, False Polarization has been shown to reinforce actual polarization. If I (wrongly) think the other person strongly disagrees with me, I might actually toughen my rhetoric in response. But in reality, most people care about all three core values underlying AI regulation.

And in reality, we do not have to choose. We can have AI progress *and* safety *and* democratic participation. Here are four policy goals that will help us get all three—a "Beneficial AI Roadmap":

Invest in innovative and beneficial uses of existing AI

Many start-ups are building innovative applications on top of existing general-purpose AI models like GPT-4. They unlock productive use cases of existing AI that can benefit millions of people. Governments could ease the regulatory burden on these SMEs to foster progress by ensuring that the general-purpose models that they build upon, which are usually developed by tech giants like Microsoft and Google, are safe, unbiased and reliable. Three-person startups should not have to deal with safety issues that stem from the technology they use as a foundation for their product. Governments could further invest in existing AI use cases that are not sufficiently valued by markets but can advance important social values like inclusion and participation. These include AI-driven scientific and medical advances, AI that advances the social good in low income countries, AI tutors that support high school students through high quality 1:1 tutoring, or tools that make it easier for people to participate in public discourse—for instance, through large language models that help synthesize and map arguments in a public deliberation process.

Boost research on trustworthy AI

The strongest AI models are increasingly capable, but unreliable and potentially harmful. And while billions of dollars are spent each year to make AI more powerful, funding for research to make AI understandable, free from bias, and safe is tiny in comparison. That is why we need a large-scale effort involving the world's best AI scientists to ensure this technology will continue to benefit humanity instead of harming it. Such an effort could aim to map all potential risks from AI (similar to what the Intergovernmental Panel on Climate Change does regarding climate change) and propose a research agenda for mitigating each risk—and then invest in research into technical solutions for each risk. This research would also provide the information required by regulators, for example to identify the more dangerous forms of AI or where to put the threshold for what should be open-source and what shouldn't. And crucially, this research could enable us to develop safe, defensive AI models that would help humanity protect itself in an AI emergency, like the misuse of powerful AIs by terrorists or an out-of-control superhuman rogue AI. Importantly, advances in capabilities that could help in such AI defense could become a weapon in the wrong hands. They should therefore be developed in a secure way that avoids a single point of failure. This requires strong democratic governance that includes individual governments, civil society and the international community. By doing this research, we would ensure that we can continue to reap the benefits of AI safely, embracing participation, while supporting progress.

Democratize AI oversight

Many of the world's leading AI experts now think that human-level (or even more capable) AI could arrive before 2030. Regardless of the exact timelines, it's clear that unelected tech leaders should not decide whether, when, how and for what purpose such transformative AI is built. This means that, while we might not want to democratize direct access to potentially destructive technology for safety reasons, we urgently need to democratize AI oversight (participation). This is beginning to happen. The E.U. and Canada are currently finalizing their first AI legislation. These efforts will not enable regulators to address all risks, but they are a good starting point. In the U.S., various drafts for binding regulation are circulating, including bi-partisan efforts. But with the current speed of AI development, we do not have time to lose; regulation will have to accelerate and rest on an agile principles-based framework. Voluntary commitments by companies can help in the short term but they are not enough. And at a global level, we need a minimal set of binding rules governing AI R&D worldwide. This will not be easy, but it is in every country's interest to avoid global AI catastrophes. A recent statement warning about AI posing an extinction threat to humanity, for instance, was also signed by leading Chinese and Russian researchers.

Set up procedures for monitoring and evaluating AI progress

With increasingly capable AI models, we will want to know whether somebody in North Korea, for example, is currently using 20,000 AI chips for a huge training run to build their GPT-6. For this reason, governments should consider "compute monitoring": tracking globally who is using the chips needed for building AI models. Governments might also want to mandate on-chip devices that signal to the regulator if they are being used for purposes that violate AI R&D rules, while protecting developers' IP rights. Compute monitoring could be paired with a licensing regime. Only certified responsible labs would then be allowed to train the next generation of general-purpose AI models. This would create transparency and accountability, and it could help solve global coordination problems around AI development. That's important because at some point, it might become necessary to temporarily slow down or pause certain types of AI development globally for safety reasons. Knowing who has the amount of AI chips needed for developing potentially dangerous models would make this easier. Finally, to identify undesirable or even dangerous capabilities in new models, governments should mandate external evaluations: expert "red-teamers" should test models for risks such as dangerous capabilities, bias, dispositions for privacy violations and hallucination.

These four policy goals are not exhaustive, but they would make for a great start. And the best part is that they are mutually compatible, jointly supporting AI progress, safety, and participation. At the upcoming AI Safety Summit in the U.K. and in all efforts to regulate AI, governments should focus on these common-sense, pragmatic policies, and not get distracted by false claims that we have to choose between different core values. Yes, people have different opinions about AI regulation and yes, there will be serious disagreements. But we should not forget that mostly, we want similar things. And we can have them all: progress, safety, and democratic participation.

Yoshua Bengio is a professor at the Department of Computer Science and Operations Research at the Université de Montréal, scientific director of the Montreal Institute for Learning Algorithms and 2018 Turing Award winner. Daniel Privitera is the founder and executive director of the KIRA Center for AI Risks & Impacts and a PhD candidate at the University of Oxford.

Contact us at letters@time.com.