Arms Races

“AI arms races might happen, but they are sector specific, and should be addressed through sector-specific regulations.”

Evaluating AI

“AI benchmarks are useful for measuring progress in methods; unfortunately, they have often been misunderstood as measuring progress in applications, and this confusion has been a driver of much hype about imminent economic transformation. For example, while GPT-4 reportedly achieved scores in the top 10% of bar exam test takers, this tells us remarkably little about AI’s ability to practice law.25 The bar exam overemphasizes subject-matter knowledge and under-emphasizes real-world skills that are far harder to measure in a standardized, computer-administered format. In other words, it emphasizes precisely what language models are good at—retrieving and applying memorized information.”

“This pattern appears repeatedly: The easier a task is to measure via benchmarks, the less likely it is to represent the kind of complex, contextual work that defines professional practice. By focusing heavily on capability benchmarks to inform our understanding of AI progress, the AI community consistently overestimates the real-world impact of the technology.”

Real-World Impacts

“Intelligence is not the property at stake for analyzing AI’s impacts. Rather, what is at stake is power—the ability to modify one’s

environment. To clearly analyze the impact of technology (and in particular, increasingly general computing technology), we must investigate how technology has affected humanity's power. When we look at things from this perspective, a completely different picture emerges.

Are These Systems "Intelligent?"

"De-emphasizing intelligence is not just a rhetorical move: We do not think there is a useful sense of the term 'intelligence' in which AI is more intelligent than people acting with the help of AI. Human intelligence is special due to our ability to use tools and to subsume other intelligences into our own, and cannot be coherently placed on a spectrum of intelligence.

Forecasting Geopolitical Events and Predicting Persuasion Capabilities

"We predict that AI will not be able to meaningfully outperform trained humans (particularly teams of humans and especially if augmented with simple automated tools) at forecasting geopolitical events (say elections). We make the same prediction for the task of persuading people to act against their own self-interest.

Model Alignment and Human-in-the-Loop

"Discussions of AI control tend to over-focus on a few narrow approaches, including model alignment and keeping humans in the loop. We can roughly think of these as opposite extremes: delegating safety decisions entirely to AI during system operation, and having a human

second-guessing every decision. There is a role for such approaches, but it is very limited. In Part III, we explain our skepticism of model alignment. By human-in-the-loop control, we mean a system in which every AI decision or action requires review and approval by a human. In most scenarios, this approach greatly diminishes the benefits of automation, and therefore either devolves into the human acting as a rubber stamp or is outcompeted by a less safe solution. We emphasize that human-in-the-loop control is not synonymous with human oversight of AI; it is one particular oversight model, and an extreme one."

The Practical Problem of the "Superintelligence" Framing

Technical AI safety research is sometimes judged against the fuzzy and unrealistic goal of guaranteeing that future "superintelligent" AI will be "aligned with human values." From this perspective, it tends to be viewed as an unsolved problem. But from the perspective of making it easier for developers, deployers, and operators of AI systems to decrease the likelihood of accidents, technical AI safety research has produced a great abundance of ideas. We predict that as advanced AI is developed and adopted, there will be increasing innovation to find new models for human control.

The Analogy to Nuclear Weapons

"AI is often analogized to nuclear weapons. But unless we are talking about the risks of military AI (which we agree is an area of concern and do not consider in this paper), this is the wrong analogy. With regard to the concern about accidents due to the deployment of (otherwise benign) AI applications, the right analogy is nuclear power. The difference between nuclear weapons and nuclear power neatly illustrates our

point—while there was a nuclear weapons arms race, there was no equivalent for nuclear power. In fact, since safety impacts were felt locally, the tech engendered a powerful backlash in many countries that is generally thought to have severely hobbled its potential."

U.S.-China Arms Race

"The U.S. versus China arms race rhetoric has been strongly focused on model development (invention). We have not seen a corresponding rush to adopt AI haphazardly. The safety community should keep up the pressure on policymakers to ensure that this does not change. International cooperation must also play an important role."

The Limitations of Alignment

"Model alignment is often seen as the primary defense against the misuse of models. It is currently achieved through post-training interventions, such as reinforcement learning with human and AI feedback.62 Unfortunately, aligning models to refuse attempts at misuse has proved to be extremely brittle. We argue that this limitation is inherent and is unlikely to be fixable; the primary defenses against misuse must thus reside elsewhere."

"Model alignment seems like a natural defense if we think of an AI model as a humanlike system to which we can defer safety decisions. But for this to work well, the model must be given a great deal of information about the user and the context—for example, having extensive access to the user's personal information would make it more feasible to make judgments about the user's intent. But, when viewing AI as normal technology, such an architecture would decrease safety

because it violates basic cybersecurity principles, such as least privilege, and introduces new attack risks such as personal data exfiltration."

"We are not against model alignment. It has been effective for reducing harmful or biased outputs from language models and has been instrumental in their commercial deployment. Alignment can also create friction against casual threat actors.

Yet, given that model-level protections are not enough to prevent misuse, defenses must focus on the downstream attack surfaces where malicious actors actually deploy AI systems.66 These defenses will often look similar to existing protections against non-AI threats, adapted and strengthened for AI-enabled attacks."

What Policy Should Focus On

"Defense against superintelligence requires humanity to unite against a common enemy, so to speak, concentrating power and exercising central control over AI technology. But we are more concerned about risks that arise from people using AI for their own ends, whether terrorism, or cyberwarfare, or undermining democracy, or simply—and most commonly—extractive capitalistic practices that magnify inequalities. Defending against this category of risk requires increasing resilience by preventing the concentration of power and resources (which often means making powerful AI more widely available)."

Researching Risks

"Current AI safety research focuses heavily on harmful capabilities and does not embrace the normal technology view. Insufficient attention has

been paid to questions that are downstream of technical capabilities. For example, there is a striking dearth of knowledge regarding how threat actors actually use AI. Efforts such as the AI Incident Database exist and are valuable, but incidents in the database are sourced from news reports rather than through research, which means that they are filtered through the selective and biased process by which such incidents become news."

Final Thoughts

"AI as normal technology is a worldview that stands in contrast to the worldview of AI as impending superintelligence. Worldviews are constituted by their assumptions, vocabulary, interpretations of evidence, epistemic tools, predictions, and (possibly) values. These factors reinforce each other and form a tight bundle within each worldview.

For example, we assume that, despite the obvious differences between AI and past technologies, they are sufficiently similar that we should expect well-established patterns, such as diffusion theory to apply to AI, in the absence of specific evidence to the contrary.

Vocabulary differences can be pernicious because they may hide underlying assumptions. For example, we reject certain assumptions that are required for the meaningfulness of the concept of superintelligence as it is commonly understood.

Differences about the future of AI are often partly rooted in differing interpretations of evidence about the present. For example, we strongly disagree with the characterization of generative AI adoption as rapid (which reinforces our assumption about the similarity of AI diffusion to past technologies).

In terms of epistemic tools, we deemphasize probability forecasting and emphasize the need for disaggregating what we mean by AI (levels of generality, progress in methods versus application development versus diffusion, etc.) when extrapolating from the past to the future.

We believe that some version of our worldview is widely held. Unfortunately, it has not been articulated explicitly, perhaps because it might seem like the default to someone who holds this view, and articulating it might seem superfluous. Over time, however, the superintelligence view has become dominant in AI discourse, to the extent that someone steeped in it might not recognize that there exists another coherent way to conceptualize the present and future of AI. Thus, it might be hard to recognize the underlying reasons why different people might sincerely have dramatically differing opinions about AI progress, risks, and policy. We hope that this paper can play some small part in enabling greater mutual understanding, even if it does not change any beliefs."