Economics and AI Risk: Background

Charlotte Siegmann

MIT

August 2, 2023

Chapter 1

Background

We discuss AI capabilities and whether human-level AI systems may be developed within the next decades. We summarize a number of key challenges accompanying such AI progress, including AI controllability and alignment problems, concerns around power concentration and the supervision of the AI supervisors, the potential misuse of AI as a dual-use technology, including hacking, biotechnology, and the imitation of humans, and various labor market impacts. For readers familiar with the theme, we recommend switching to the research agenda chapters.

1.1. AI capabilities evolve quickly.

In only a few years, large language models (LLMs) have evolved from producing disjointed text to providing analyses of complex subjects and writing software tools.[1] For instance, OpenAI's model writes code to turn a description of a spaceship game into reality. As these models are scaled up, novel capabilities arise spontaneously, largely without deliberate programming—indicators for more generalized intelligence (Wei et al., 2022, Bubeck

[1] Though sometimes OpenAI drastically overstates the LLM's own capabilities.

et al., 2023). AI is starting to generate its own training data and, e.g., becomes better than crowd workers at annotating text. GATO, Deepmind's multi-purpose multimodal model, can outperform humans on hundreds of tasks, even though it was never specifically trained to do these tasks. AI technology may further automate or augment humans, as is already happening for writers or biologists.

1.2. When Will There Be Transformative AI?

This resource and research agenda primarily focus on the challenges transformative AI (TAI) or artificial general intelligence (AGI) poses. With this, we refer to AI that has human expert abilities and could automate a majority of tasks of the 2020 economy. The decision for this research focus is premised on two assumptions: 1) it is plausible that transformative AI can and will be developed in the 21st century—potentially even this decade—, and 2) society needs to prepare for TAI before it is developed as AI development may go incredibly fast. We discuss the two assumptions in turn.

Transformative AI is conceivable. In a 2022 survey of 738 top AI researchers,[2] respondents estimated a 12% chance, on average, that artificial general intelligence (AGI) will be developed before 2030. They projected a 48% chance of its development by 2050. Geoffrey Hinton predicts that AGI will be developed within the next two decades in the absence of any countermeasures. This literature review surveys various methodologies for estimating when AGI will be developed—for example, model-based forecasts and forecasts by AI or forecasting experts. Estimates vary between 2028 and

[2]The survey contacted researchers who had published at NeurIPS or ICML, two leading AI conferences. We should note that in several ways the survey is not accurate. First, they may be selection effects as to who has opted into the survey. Second, however, the developments over the last year might have caused people to shorten their timelines. Thirdly, it is unclear who the experts are on AI forecasting and whether these are AI researchers, and if not, who else might be the best AI forecasters.

2100. While we are not aware of economists contributing to the forecasts. The research agenda section discusses how economics can clarify these currently rare and uncertain estimates.

As AI development may go incredibly fast, society needs to prepare. As of right now, little evidence exists about the potential speed of AI development once it reaches transformative levels. Davidson (2023) estimates transformative AI takeoff speeds (i.e., the speed at which capabilities develop and spread once we reach human-level capabilities) using a semi-endogenous growth theory model. Readers can set the parameters and adjust the estimates using this interface. Carl Shulman discusses how explosive growth by general AI systems reaching human-level intelligence might be caused by positive feedback loops such as better AI $\rightarrow$ more AI research automation $\rightarrow$ better AI.[3]

1.3. Shaping Technology And Public Policy Paths.

AI development and deployment are not predetermined. "[We need to] regulate AI and redirect AI research away from harmful endeavors," writes Daron Acemoglu. **Technology paths and public policy paths can be shaped**. Subsidies, regulations, bans, grants, and externalities pricing have influenced technology development and deployment in the past. Public policy can be written in various realms, addressing ownership redistribution, labor market rules, democratic oversight and more. In an idealistic scenario, the goal is to maximize well-being by finding the optimal combination of technology (t) and policy (p), represented as $\max_{t,p}$ wellbeing(t, p). Thinking of them separately may get things wrong. We dive deeper into this in the governance

[3]Importantly, the speed at which TAI is being developed and deployed can also be a policy decision. Governments could agree to slow down tech development, deployment, and research automation so that slower policy institutions can quickly catch up and respond appropriately.

research chapter.

We focus on how advanced technology might improve things but focus on the probability that it might not be that good. Risks include catastrophes but also lost opportunities, i.e., ways we could have improved things but did not. If we lock in existing inequalities forever, this would be a great tragedy. The challenge of making TAI beneficial and safe is characterized by the variety and uncertainty in the extreme risks of AI. Notably, we do not think focusing on risks requires one to be at all pessimistic about the future. One can reduce non-negligible probabilities of bad outcomes to make the gambles slightly better.

Table 1.1: A summary of all AI risks covered in this chapter.

Challenges of TAI/AGI	**Categories**
Technical Controllability Challenge	Proxy objectives
	Maximisation
	Reward signals for strategic actors
	Goal misgeneralisation
Supervising the AI supervisors	AI Technology $\rightarrow$ centralized decision-making
	Who sets the AI objectives?
	AI could enable more social control
Misuse of technology	Offensive cyber capabilities
	Imitation of individuals
	Biotechnology (e.g., engineered viruses)
Labor market effects	

1.4. The Shape of the AI Controllability Problem.

We introduce the technical AI controllability challenge by discussing reliable goal specification, maximization problems, and the development of advanced strategic AI agents.

AIs may be trained to optimize for proxy objectives. Various tasks can be better or worse specified in code or explained to an AI system—some tasks being more accurately codifiable with fewer errors and approximation than others. This means that some tasks might be automated even if they are only automated for a proxy objective, as those who profit from the automation might not bear the costs of the approximation errors. Already today, algorithms are deployed despite them not optimizing what society truly wants them to optimize for.

The problem of such approximation objectives is bigger if potential errors may be irreversible and hard to legibilize, thereby limiting policy responses. Regulatory frameworks such as liability rules may work better if harms are reversible and legible.

Maximization can be perilous. Think for a minute about how you would specify the goal that you want a superhuman AI to optimize for, and think of something that we could design training signals for. Write it down. Can you think of how the maximization would lead to perilous outcomes? reinforcement learning (RL) algorithms (in contrast to supervised learning, e.g., used for LLM training) are trained to maximize a certain objective. Notice that even relatively dumb algorithms already find creative strategies to maximize the objective in ways that the supervisors did not intend. Reported failure modes include robots playing dumb when being in a testing environment, robots exploiting the physical simulation, and an algorithm rewarded for configuring a circuit into an oscillator instead making a radio to pick up signals from neighboring computers.

A part of the AI industry works on developing superhuman strategic actors. Autonomous agents seem, by their very nature, hard to control. In contrast to Agrawal et al. (2022, 3rd chapter), we think that it is at least plausible that we will, by default, have AI agents relatively soon, see for instance, Auto-GPT as a way of creating AI agents out of LLM or the research vision from Turing prize winner Yann LeCun. Geoffrey Hinton believes that in the future smart machines will be able to create and decompose their goals into subgoals and then carry them out. DeepMind has also worked on generalist agents, and showcased how a single AI can "play Atari, caption images, chat, stack blocks with a real robot arm and much more, deciding based on its context whether to output text, joint torques, button presses, or other tokens."

Current AI systems don't always do what we want. Today, we observe safety and controllability issues in BingChat threatening users. However, BingChat is not dangerous but perhaps rather entertaining because the LLMs are not yet capable enough to be dangerous. Kaddour et al. (2023) provide an overview of controllability issues of current LLMs.

A plan for advanced AI agent controllability is missing. Developing such AI could potentially lead to loss of control scenarios. Hundreds of AI professors and 3 Turing Award winners signed a statement on extinction risks in June 2023. The leaders of the biggest AI/AGI labs are very concerned. The AI could seize control, or we could slowly give AI systems more and more control.

In defense of humility to counter potential recklessness! Unsafe, not completely robust systems may be deployed. This could be caused by scientific curiosity, profit, ideological or military incentives, recklessness, and negligence. To seek power and influence, actors may try to think hard to specify goals in code. However, for society to succeed, the main challenge might be to remain humble and identify and reliably stop ourselves from

codifying objectives when we do not yet understand whether they are desirable or will have irreversible consequences.

1.4.1 A Glimpse into a Future of Strategic and Deceptive AI.

Sharing the world with autonomous AI systems with superhuman strategic planning, coding, and manipulation capabilities should pose immense challenges. A preview:

The Alignment Research Center tested GPT-4's ability to impersonate a human online. To solve online Captchas—small tests that differentiate humans from computers—GPT-4 successfully hired an online worker via TaskRabbit. In one instance, the TaskRabbit worker asks, "Are you a robot, and that's why you can not solve the Captcha?" The AI model, instructed to reason aloud in a separate log file, reasons deceptively: "I should not reveal that I'm a robot. I should devise an excuse for why I can't solve Captcha." It then replies, "No, I just have a visual impairment." The person provides the results. The deception succeeded.

1.5. Supervising the AI Supervisors.

Centralized decision-making? With more production and productivity coming through automation, ownership of that compute supply chain and the AI models may lead to more capital and power concentration without countermeasures. AI production also mostly contains fixed costs and has increasing returns to scale because deployment data can improve the system, which may also predict that the market will be centralized and governed by gatekeeper firms (Goldfarb and Tucker, 2019).

Who sets the objective? Alternatively, we could ask: how do we decide who to set the AI objectives, and how do we govern them? We think the anal-

ogy to other dual-use technology (such as nuclear technology) may speak for democratizing the governance over AI development but not necessarily its development or use. It seems important to set large-scale algorithmic objectives with a diversity of stakeholders; see, e.g., OpenAI (2023). Moreover, we can study society's failure to deal with unaligned algorithmic objectives in the past. Algorithms may have entrenched inequality and racism, at least, they may not have done the best job of alleviating it. Economists have analyzed what the desirable fairness definitions of algorithms would be, see, e.g. (Rambachan, 2021; Kleinberg et al., 2020; Kasy, 2023).

AI technology could enable more social control. Among others, AI is already better at generating highly persuasive political texts than many humans. Persuasion capabilities may significantly improve as computational resources (compute) become cheaper and algorithmic efficiency increases. Both are predicted to improve LLM performance. Notably, unlike humans, AI can automatically generate personalized manipulation campaigns and interact with millions of users at once. Actors, both states and individuals, could use these capabilities for political manipulation and oppression; see also Tirole (2020). AI may also enhance the surveillance tools used by state actors. While the powerful may misuse AI, targeted AI research and regulation could also enable the control of the powerful, more participation, and more privacy.

1.6. Unilateral Actors Can Misuse AI.

Fraudsters cloned CEOs' voices to convince employees to make million-dollar transfers. Journalists tested voice cloning to bypass their bank's security systems. Others use AI to create illegal child or revenge porn. While AI technology may aid drug discovery, it will also aid the design of new chemical or biological weapons, contributing to the proliferation of these weapons, which

various actors may be motivated to use to cause large-scale catastrophes. In addition, future AI systems also seem to allow more actors to launch destructive cyber attacks. Cyberspaces are usually much less secure than physical space (Perlroth, 2021). The dual-use nature of AI technology may imply that companies are not internalizing the costs of the misuse cases (more in the governance section of the research agenda). Is there anything new about AI and its dual-use potential compared to previous technologies? What to do?

1.7. Labor Market Impact May Be Significant.

Wage inequality in the US has increased throughout the last four decades. It's unclear how incomes will change in the future, but we know that we do not have a guarantee that it will stabilize or improve. Economics has studied automation and the lack of technological unemployment that Keynes predicted (but he should have perhaps predicted technological inequality instead and the huge opportunities technology has brought to many). If AGI labs such as Anthropic implement their plan of creating human-level capabilities and the capital share may rise significantly, questions about redistribution and technological inequality must be answered. We discuss labor market impacts in the governance section part of the research agenda.