Economics and AI Risk:
Research Agenda and Overview

Charlotte Siegmann

MIT

August 2, 2023 – Version 0.1

Preliminary

Chapter 1

Executive Summary

We describe research areas within economics that could contribute to addressing the challenges of transformative AI development and deployment.

1.1. AI Governance: Designing Optimal Policies

The governance and market design of advanced, general-purpose AI systems spans numerous areas. First, we explore the challenge of safely developing transformative AI. This includes (i) directing AI advancement through an amalgamation of subsidies and restrictions, (ii) cultivating an effective ecosystem for AI development, e.g., through oversight, competition rules, and information-sharing regimes, (iii) addressing the unique dual-use market failure in AI development, and (iv) analyzing the AI input supply chains to eventually improve it to ensure safe and beneficial AI development. Secondly, we discuss the safe and beneficial deployment of transformative AI, considering aspects such as safe implementation, the sharing of TAI benefits, and the economics of AI evaluations. Third, drawing an analogy from how estimates of the social cost of carbon have refined discussion and policymaking, AI forecasts and models could likewise enhance AI policy. We summarize economic research questions aiming to improve the forecasting of AI capabilities, e.g., coding or strategic planning abilities, and AI-associated risks, e.g., misuse scenarios.

1.2. AI Controllability: Designing Aligned AI

Making general-purpose AI systems do reliably what humans want can benefit from a profound understanding of human behavior and incentive systems. We proceed with a general introduction of the controllability problem targeted at economists. Second, we introduce a framework to assess the value of AI training processes and AI training goals. Third, we summarize main research areas within AI controllability. Fourth, we outline various economics research areas within AI controllability research, including AI conflict, prediction market design, and causality models.

Chapter 2

Governance Research

We analyze an array of research areas within the economics of AI governance.

Table 2.1: A summary of all AI governance themes covered in this chapter.

AI development	**Technology regulation**	
	AI industry ecosystem	Competition
		Information sharing
		Organization structures
		Democratic oversight
	AI dual-use market design	
	AI inputs/supply chain	Compute
		Data
		Algorithms
AI deployment	**Safety**	
	Benefit sharing	
	Algorithm/AI assessments	
AI forecasting	**AI capabilities**	
	AI risks	

First, we explore the challenge of safely developing transformative AI. This includes guiding technological advancement through a mix of subsidies and restrictions, creating an effective ecosystem for AI development, treating the development issue as a unique dual-use market failure, and analyzing the AI input supply chain. We are particularly excited about more research because of the complexities of the threat models and the seemingly stark trade-offs (e.g.if AI is open-sourced, it could be misused more but might also lead to less inequality).Secondly, we delve into the safe and beneficial deployment of transformative AI, considering aspects such as safe implementation, the sharing of TAI benefits, and the economics of AI evaluations. Third, we summarize the forecasting of AI capabilities.

2.1. AI Development Challenges

We introduce various facets of the governance and economics of AI development, summarize the existing literature (albeit academic work is often missing), and outline open research areas.

2.1.1 Shaping Technology Paths

Technology paths can be shaped through subsidies or rules. Some existing literature on the AI development path includes:

- Jones (2023) discusses whether we should develop AGI now, never, or at some future time when we're economically well-off enough to switch it off if needed. Acemoglu and Lensman (2023) analyze how technological adoption should be slowed down if one learns about the risk levels over time.

- Well-developed agreed-upon international advanced monitoring evaluation regimes (i.e., defining what safe and socially beneficial AI would look like and only deploying and developing such AI systems) could incentivize safe and beneficial AI development (Liang et al. 2022, Shevlane et al., 2023).

- Shavit (2023) outlines a potential AI development compute monitoring mechanism. Countries would monitor the computing hardware used for large neural network training. This includes on-chip firmware, training run information storage, and the scrutiny of the chip supply chain. This proposal raises numerous verification questions about data center governance, the definition of effective compute use, and effective punitive schemes.

- Throughout history, humans have restricted the development, use, or proliferation of certain powerful technologies, such as bioweapons, laser technology, or human cloning. Past governance mechanisms could inform our approach to AI governance.

Selected Research Questions

- When exactly should society ban certain types of AI development or deployment, at least temporarily? When should it be open-sourced or closed?

- What are efficient ways of guiding technology paths to be most welfare-enhancing?

- How to enforce AI rules even among the non-believer states?

2.1.2 AI Industry Ecosystem

This section considers what structural changes could make the AI devleopment more beneficial. **Competition and Merging**

- **Cooperation Clauses**: In the OpenAI Charter, the organization commits to cease competition and assist any value-aligned, safety-conscious competitor who is close to achieving Artificial General Intelligence (AGI). Is this strategy beneficial? How should it be implemented and enforced?

- **Monopoly and Competition:** How should the competition landscape be governed?

Information Sharing

- **What information-sharing regimes, e.g., for safety (and capability) insights, are optimal?** See, e.g., Solaiman (2022). As of June 2023, it appears that most AGI labs decided not to share any AI capabilities insights to avoid races or proliferation: see Anthropic, Google, as well as capabilities insights missing from OpenAI's GPT-4 paper.

Organization Structure

- **What is the ideal organizational structure for developing TAI?** At OpenAI, the NGO Board has the last decision-making authority; Anthropic is a public-benefit corporation; and Google DeepMind is part of Alphabet, a public company. Siegmann (2023) proposes one vision.

Democratic Oversight

- If decisions about technology development are political ones (and not just normal consumer products), how can people influence AI development? How can this be done internationally? For instance, should the monopoly on AI safety (versus capabilities work) be entrusted to an international body like CERN? How could this be achieved? See, e.g., CIP (2023).

2.1.3 AI Dual-Use Market Failures

We discuss whether AI is a dual-use good different from other goods and what, if at all, this might mean for market design.

AI is a dual-use technology like nuclear, biological, and chemical technologies. It can be used to advance well-being and, in the wrong hands, cause much harm. AI can be used to deliberately cause physical harm, be deployed as a persuasive technology (e.g., for marketing, election campaign, social engineering, or subliminal manipulation), and impersonation, which could also erode potential institutions, and be used for advanced surveillance. And AI is already being treated as such. In 2022, the US government put some NVIDIA chips on the export control list. They now need a special license to export it to China. However, NVIDIA aptly found ways around the export control restrictions.

AI differs from previous dual-use technology in three ways:

1. **Players**: Unlike other dual-use technologies, many more actors are involved in AI development, which is mostly organized within a profit-seeking market.

2. **Shareability**: AI is much easier to copy and share. One can take a copy of a model, and the original owner may not even notice. That is different from nuclear technology or biotechnology.

3. **Specificity of security threat:** In the case of AI, the security threat is less well-understood and could come from a much greater variety of actors and pathways. For instance, most are worried about authoritarian or unsafe governments in the case of uranium enrichment. In the case of AI, the security threat could come from many: the own governments through surveillance technology, other governments, terrorists, or reckless or negligent users and scientists.

In the past, the governance of dual-use technology often just included barring bad actors from accessing the technology. This seems much harder in the case of AI.

Selected Research Areas

- **What is the dual-use aspect?** Humans are also dual-use, which does not necessarily mean the labor market has any problems?
- **TAI market design:** further explore how a good market design of AI as a dual-use technology would look like. Work in this area could also better understand whether a dual-use market failure exists.
- **Previous dual-use technology market failures and solutions:** What can be learned from them, and how were they fixed?

2.1.4 The Economics of AI Inputs

To understand the economics of transformative AI and to guide AI responsibly, it might be helpful to study the economics of AI inputs and relevant industrial policy.

The main inputs to AI development are data, algorithms (e.g., algorithmic efficiency increases or neural architecture design), and computational resources (short: compute, including AI-specialized chips). Talent and capital are needed to create all of the above inputs. The data economy has been studied in economics (though we are not aware of much concerning generative AI or LLMs).

For now, we will focus on the economics of compute and leave the economics of algorithms for future introductions.[1] Small silicon chips have been playing an increasingly important role in our economy over the last

[1] For instance, under what conditions should algorithms be deleted? See Khan (2023) and Riley (2023).

70 years (Miller, 2021). The largest AI training runs use specialized compute hardware (e.g. so-called AI accelerators).

The complex political economy of AI compute supply chain: The development of cutting-edge chips is relatively centralized and faces immense demands (Miller, 2021). Moreover, among others, China, the US, and the EU are subsidizing chip development and industry (Verwey, 2019). The US industrial policy strategy also includes the attempted decoupling of the supply chain between China and the US. China supports the development of data centers in Africa through the Belt and Road Initiative.

The compute industry and supply chain setup and soft-regulation can improve AI governance and AI outcomes. Notably, the shape of the AI supply chain can determine the speed, incentives, and location of advanced AI development. Compute can be used to govern AI development. For instance, Jensen et al. (2023) analyze compute pricing as a strategy for internalizing AI externalities. Moreover, the supply chain could also directly change the features of AI development, e.g., how it protects IP.

2.2. AI Deployment Challenges

2.2.1 Safe Deployment

If a company develops a powerful AI system with abilities in strategy, coding, and manipulative persuasion that *far* exceed those of human experts, what should the company do (Karnofsky, 2022)? Should they deploy the AI? Give it to a state? Give it to the UN?

2.2.2 Benefiting and Redistribution

Companies such as Anthropic, OpenAI, and Google DeepMind want to create AI systems able to replace most of the human labor force. If they succeed, this may have immense political and economic consequences, influencing labor income inequality and the labor share.

Some policies have been proposed to deal with such potential transformation. For instance, the Windfall Clause is a proposed ex-ante commitment by AI companies to donate a substantial portion of any exceptionally large profits they generate in the future. These large profits are defined as those that could not have been created without the capabilities of transformative AI. The Windfall Clause was proposed to partially address employment and income inequality effects and stabilize societal relationships during a potential time of turmoil as we transition to TAI.

Other proposals include a universal basic dividends (see, e.g., the American pension funds), sovereign wealth funds (see, e.g., Alaska's provident fund), and common ownership of computational resources (compute) to address inequality from TAI.

Selected Research Areas

- **Forecasting**: How will TAI affect the labor market? What determines how ownership, redistribution, and technology access will be shaped?
- **Labor Policies:** What policies should be implemented during the transition to TAI to mitigate the effects on human workers? In a context where current employees cannot retrain, how will the capital and ownership, e.g., over computational resources, be reallocated? How should capital be redistributed in the long run?
- **Power Diffusion:** What mechanisms can be implemented to sufficiently diffuse power to ensure decentralization of power?

2.2.3 Research and Evaluation of the Algorithmic Economy

As more algorithms become entrenched in our society, it becomes important to empirically study the effects on humans, society, and politics, as has already been done, e.g., in the case of polarization. A better understanding of the societal effects of the algorithms allows policymakers and AI engineers to react. In the following, we exemplify this with the potentially performative effects of prediction technology. We argue that the algorithmic economy requires thicker well-being metrics but, luckily, also enables new preference elicitation tools, enabling new kinds of outcome variables.

Exemplifying the need for greater theoretical understanding: Advanced prediction technology and self-fulfilling prophecies. Some AI technology can be framed as reducing the cost of prediction and increasing its quality. What are the unintended side effects of prediction technology (if the prediction itself changes the outcome)? If someone predicts Katy to become an A student or be unemployed, this itself may change the probability of her becoming an A student or being unemployed (self-fulling or self-defeating prophecy). What should be done? How should technology be deployed? See the footnote for real and fictional examples.[2]

[2] Real-world examples: Criminology and police forces have always used models to predict crime. What's changing now? Models are designed to, e.g., predict where crimes will occur and by whom, but they are more trusted and becoming better. This raises questions for us today about self-fulfilling prophecies (if poor areas are predicted to have a higher rate of crime, then there will be more police officers there, and a significantly greater proportion of, for example, impoverished cannabis consumers compared to wealthy cannabis consumers will be arrested), see Noble (2018) and O'Neil (2016). This raises questions about what it would mean to use and design predictive technology for a fairer and better society, not just in the short run, but also when we consider how the prediction itself will change society and alter the underlying social fabric. Economists refer to this as self-fulfilling prophecies. This phenomenon has also been studied and defined in the field of machine learning as auto-induced distribution shifts, e.g., for social media algorithms. As a fictional example, Steven Spielberg's film, Minority Report, deals with similar themes, such as the reckless and overconfident deployment of prediction technology, human trust in the technology, the very political question of what will be predicted and what not, and the phenomenon of self-fulfilling prophecies.

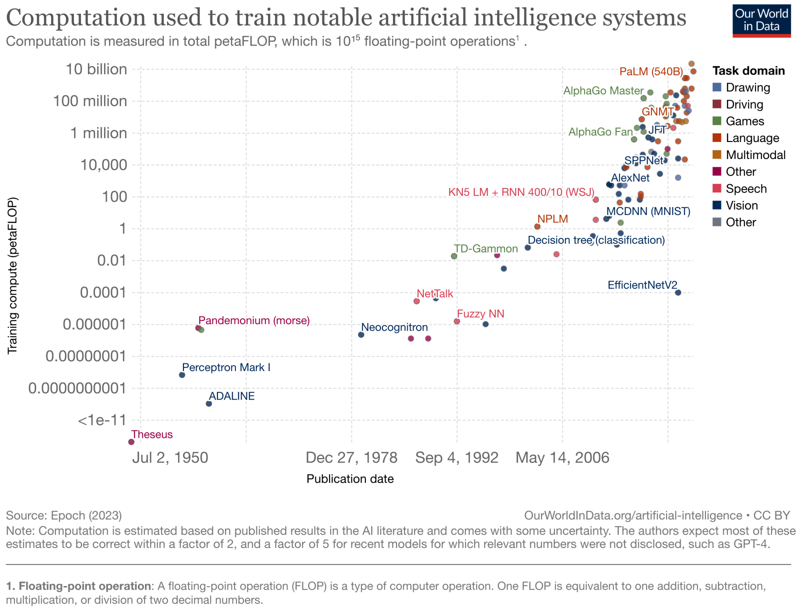
The algorithmic economy requires and enables thicker outcome/well-being measurements. The measurement or approximation of welfare with metrics such as happiness reports or GDP is a contested and regularly misunderstood topic. Without debating under what conditions which approximations are useful, it is important to point out that LLMs enable scientists to measure much thicker ways of the good life and well-being and enhancements on a large scale. One can condition chatbots to do high-quality qualitative interviews with many humans (a prompt-engineered QualitativeInterviewerGPT) and use that to learn from a greater group of humans how they would like to redesign their life. Such research methodology can now be developed and *responsibly* implemented in a high-quality way. Moreover, as algorithms actually optimize for metrics in contrast to human institutions, Goodhart's law needs to be taken much more seriously.[3] This is another reason Stiglitz et al. (2018) have argued that we need to give up on metrics that supposedly can track everything.

2.3. Forecasting Research

Understanding the AI future involves predicting capabilities, e.g., coding or strategic planning, and risks, e.g., dangerous capabilities, misuse scenarios, and deployment conditions. We summarize the mostly non-academic literature in both areas.

Predicting the ordering and timing of AI development and deployment may help society make the best policy decisions. It allows policymakers to take preemptive measures, such as regulating AI or creating directed technological change through subsidies and rules. In that way, AI forecasting

[3] However, we think this worry has also lost a lot of teeth within the LLM paradigm. Such perilous maximization is probably the right way to think about outcome-based RL, but it is not clear that these will be the failure modes as we remain within the RLHF + LLM paradigm.



Source: Epoch (2023)

OurWorldInData.org/artificial-intelligence • CC BY

Note: Computation is estimated based on published results in the AI literature and comes with some uncertainty. The authors expect most of these estimates to be correct within a factor of 2, and a factor of 5 for recent models for which relevant numbers were not disclosed, such as GPT-4.

[1] **Floating-point operation**: A floating-point operation (FLOP) is a type of computer operation. One FLOP is equivalent to one addition, subtraction, multiplication, or division of two decimal numbers.

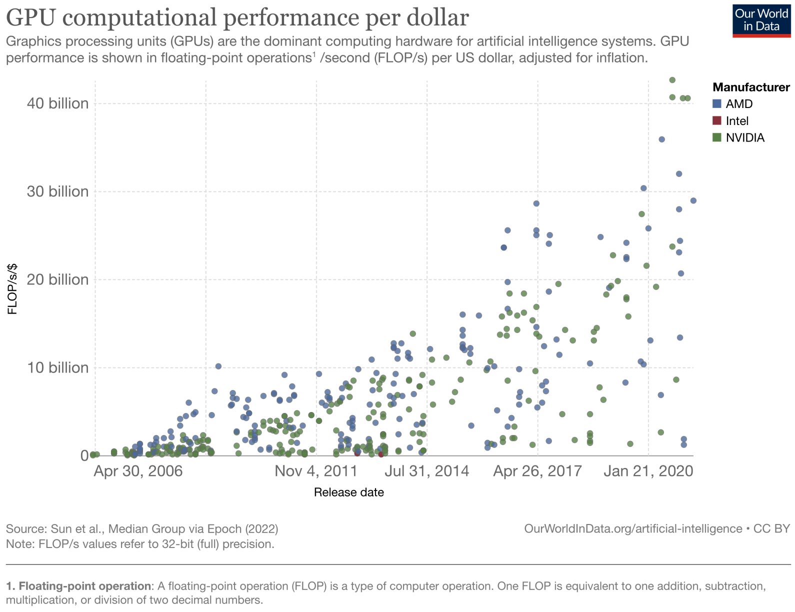
Figure 2.1: Sevilla et al. (2023)'s aggregated data on the exponential increase in the computation used in the biggest AI training runs over the past decades.

efforts can be analogous to climate forecasting efforts. Climate change forecasts and assessments, such as Nordhaus' DICE model, improved mitigation and adaptation measures. However, the development of AI also does not serve as a perfect analogy as AI development might be faster, less quantifiable, and less predictable, and there will likely remain much more scientific disagreement than in the case of the climate crisis.

2.3.1 AI Deployment and Development Forecasting

Predicting when certain AI capabilities will emerge and how they will be deployed—for example, when AI will be able to generate billions in profits or pass the Turing test—could improve policy decisions. Various research efforts contributed to such forecasting. EpochAI reviewed the literature estimating transformative AI timelines. William Nordhaus and follow-ups model whether the world is approaching an economic singularity. Geoffrey Hinton made AI risk and capability predictions. Beraja et al. (2023) studied the export and global diffusion of AI surveillance technology.

AI takeoff refers to *the time* between the development or deployment of AI that has human-level skills and the development or deployment of superhuman AI. Some argue that AI progress will accelerate enormously once AI systems are as good as current AI engineers at coding and AI research. If this fast development occurs, any social preparation or safety work must happen in advance. Davidson (2023) estimated such AI takeoff speeds using semi-endogenous growth theory, concentrating on computational resources as the key input for AI. More rigorous academic economic modeling may drastically improve takeoff and development speed estimates.



GPU computational performance per dollar

Graphics processing units (GPUs) are the dominant computing hardware for artificial intelligence systems. GPU performance is shown in floating-point operations[1] /second (FLOP/s) per US dollar, adjusted for inflation.

Source: Sun et al., Median Group via Epoch (2022)
Note: FLOP/s values refer to 32-bit (full) precision.

OurWorldInData.org/artificial-intelligence • CC BY

[1] **Floating-point operation**: A floating-point operation (FLOP) is a type of computer operation. One FLOP is equivalent to one addition, subtraction, multiplication, or division of two decimal numbers.

Figure 2.2: Sun et al. (2022) document the FLOP/s cost trend over time.

AI Automation and Productivity forecasts. Bailey et al. (2023) have argued for a productivity boom from AI in the coming years or the next decades. Brynjolffsson et al. (2023) and others have provided very early evidence of the productivity, education, and income inequality effects of generative AI. In theory, more advanced LLMs resembling human experts may replace, augment or enhance many workers that could work remotely.[4]

2.3.2 AI Risk Forecasting

With greater insight into potential risks from AI—including misuse, labor market impacts, misalignment, and accidents—society can determine the best technology and public policy pathways. The need to better understand threat scenarios has been stressed in Congress. Economic research can contribute to understanding the dynamics that may shape the future, e.g., Hanson (2016). Stuart Russell and Andrew Critch model and classify potential societal risks. Yoshua Bengio, Holden Karnofsky, and Ajeya Cotra also describe various threat scenarios.

[4] See Brynjolfsson et al. (2023) for estimates of US remote workers.

Chapter 3

Technical Controllability and Alignment Research

This section discusses the intersection of economics and **AI controllability research**, broadly referring to AI value alignment (aligning the AI's goals and behaviors with human values) and AI safety (preventing a system from impacting its environment in an undesirable or harmful way).

Making general-purpose AI systems do reliably what we want can benefit from a profound understanding of human behavior and incentive systems. First, we introduce the controllability problem targeted at economists. Second, we introduce a framework to assess the value of AI training processes and AI training goals. Third, we summarize the main research areas within AI controllability. Fourth, we outline various economics research areas within AI controllability research, including AI conflict, prediction market design, and causality models.

The reader is invited to skip to the sections that are most relevant to them. This section assumes some knowledge of deep learning and AI. To learn more, we recommend the further readings list and the glossary.

3.1. Difficulty of AI Controllability

The AI industry may build superhuman strategic AI agents. AI systems that create text, video, code, and pictures, known as 'generative AI,' are being developed. This development has enabled AIs to become more independent and agentic. For example, Auto-GPT can break down goals into sub-tasks and use the internet and other tools to carry them out, and Large Language Models (LLM) placed into robots have similarly exhibited agentic behavior. We focus on alignment and control problems that could arise as AI researchers build more and more capable, human-like, and agentic AIs, and society uses them to automate more and more of the economy. The footnote discusses the definition of an AI agent.[1]

The controllability challenge can be divided into goal misgeneralisation and reward/goal misspecification. Goal misgeneralization refers to the failure mode where the AI behaves undesirably in novel situations. The program competently pursues an undesired goal that leads to good performance in training situations but bad performance in novel test situations (Shah et al., 2022; Langosco et al., 2022). Next, we discuss goal misspecification.

3.1.1 Goal Misspecification

It is just hard to specify what we want. This is the case in the deep learning paradigm, where the AI systems are mostly black boxes**.** Consider the myth of King Midas. Midas wished that everything he touched would turn to gold, but he was horrified to discover that this included his food and his wife. His *intended* goal could be described as some prespecified objects I touch will

[1] In terms of an agency definition, we think one should roughly describe the AI as an agent if this is the best way of making sense of and predicting its behavior, see Daniel Dennettt's intentional stance for this approach to defining agency.

turn to gold. For instance, the goals of many current AI systems are misspecified as to what humans actually want. In the current paradigm of deep learning, it seems to be an unsolved problem to design systems to predict outcomes of individuals (grades, incarceration, social benefits, etc.) without reproducing past and current injustices and fulfilling all fairness conditions. As a hypothetical future example, consider a sophisticated AI system with the goal of ‘estimating a high number with as high confidence as possible.’ This AI might realize that it could increase its confidence in its estimate by using all of the world’s computing hardware to check and re-check its calculations. It could conclude that it will be better able to achieve its specified goal by releasing a biological superweapon to eliminate humanity, granting it unrestricted access to all hardware resources.

Goal misspecification can be caused by reasoning mistakes and observability limitations. Goodhart’s Law states that "when a measure becomes a target, it ceases to be a good measure." AI engineers may have chosen a goal for the AI program that strongly correlates with, but is not identical to, the desired outcome. For example, a very advanced AI system may be trained to reduce the number of reported crimes. Due to its superhuman problem-solving abilities, it will notice that it is more efficient to threaten humans not to report crimes rather than actually to reduce the number of crimes. It can also happen because the desired goals are less observable or measurable, leading developers to use more observable or measurable proxies. For instance, the author would personally prefer a Facebook recommendation algorithm that makes me the kindest, most well-rounded human in the long term rather than one that maximizes their engagement in the short or long run.

Giving good reward signals becomes much harder when the agents become more capable than the supervisor. Right now, human judgment can give good reward signals, e.g., to make AIs do backflips and Large Language models to act in an honest and helpful question through human feedback, see RLHF and RLAIF). However, human reward signals may get it systematically wrong once the AI system is sufficiently strategic and capable. The main control problem may only emerge once AI systems are sufficiently advanced.

AGI may be illusory or deceptive. The AI may be trained to do what looks good to humans rather than what we would actually find good. In addition, a chatbot may simulate a misaligned deceptive AI in the future, which can become very dangerous (Hubinger et al., 2023). We observe similar phenomena already with current cutting-edge models. In general, an apparently well-behaved AI model may not actually do what we want it to do. The program could be **deceptively aligned**: that is, it may behave well in the training environment, knowing that it is being monitored and assessed, but pursue other goals when it is deployed (Shah et al., 2022). Aligned and deceptive agents would both perform well in the training environments, but their behavior would differ in novel situations after deployment.

The principal-agent problem may serve as an analogy for the controllability challenge. Hadfield-Menell and Hadfield (2018) discuss whether the contracting problem serves as a useful analogy to the AI value alignment problem. We think the analogy does illustrate the difficulty of the value alignment problem: Finding a specification of the goal that is robust and accurate over the entire space of possible states of the world is hard—similar to the problem of incomplete contracting. But it hasn't helped us so far as a model for understanding and solving the value alignment problem. We discuss this more in the footnote.[2]

[2] We remain uncertain about the analogy's effectiveness, primarily because we believe there are more fitting conceptual models for understanding the alignment problem. The principal-agent problem most closely resembles reward misspecification in that both scenarios present challenges due to 1) incomplete monitoring, 2) the inability to specify all states owing to unawareness or bounded rationality and time limitations in training, 3) unobservable objectives (such as fundamental happiness), and 4) human errors. The last point

3.2. Framework for AI Training Processes

This section provides an excursion. The section can be skipped. We introduce i) a framework for evaluating the value of a particular AI training process, ii) desirable training goals, and (iii) approaches to contributing to safer AI.

A training process describes a particular engineering problem that creates an AI system, that is, the inputs (e.g., computational resources) and the outputs (a probability distribution over the potential behavior of the AI system). The training goal refers to the desired behavior of the AI system. The developers have an intended training goal in mind: they might want to create an obedient, intelligent AI assistant, such as GPT-4 or a protein folder. They then attempt to design a *training process* that creates such an AI system (often using neural networks).

3.2.1 Training Process Evaluation Framework

We discuss how one could evaluate the value of a particular training process using five factors; see the footnote for a more general framework.[3]

has been explored as “sycophancy". However, there are significant differences between the principal-agent and reward misspecification problems, for instance, the ability to train the AI to adjust its heuristics and preferences, and the constraint of rewarding the AI only in advance. In some aspects, AI training resembles a re-education camp for prisoners, where deceptive behaviour may arise, or training may fail to adapt to a new environment (e.g., the world outside the prison). While we believe we can create more accurate models, the mindsets and skills derived from studying contracting problems may still prove valuable.

[3] Let g_i, where $i \in I$, be a possible description of the AI's behavior (a training goal). We can now describe a particular training process by:

$p(g_i)$: the probability that a certain training goal. g_i is realized in the training process.

$w_s(g_i)$: the welfare the supervisor gets from using a system with goal g_i (this includes the potential costs during inference time).

$w_o(g_i)$: the welfare others get if the supervisors use the system with goal g_i.

c: the cost used for training the system.

Notice that if $w_o(g_i) > (<) 0$, we have positive (negative) externality issues that will likely not be fully internalized either way. Conversely, if $w_o(g_i) < 0$, we have negative externality issues. In addition, no matter how good a training process is for all if c is too high, it will

Training process success	Likelihood that the AI will end up with the training goal
Training failure damage	Expected damage if the AI does not have the training goal
Training process costs	Costs of executing the training process (e.g., compute costs)
Training goal desirability for society	Expected benefits and costs if an AI with the training goal is deployed
Training goal desirability for the user	Expected benefits and costs for the user[4]

3.2.2 Promising Training Goals?

We do not know what the most promising training goals would be in the medium to long-term. Hypothetical **training goals** that may generate safe and competitive AI systems (and the industry should aim for) include:

- Building AIs that forecast the future (oracles)
- Developing narrow tools rather than general-purpose AI systems
- Making obedient AIs (intent-aligned AI assistants)
- Building very intelligent AI systems to learn how they represent knowledge, then using that knowledge ourselves and switching off the AI

3.2.3 Kinds of Beneficial Research

We discussed how to evaluate existing complete training processes and what potentially promising training goals could look like. This section describes how research projects can improve training processes.

- **Enable training processes for training goals that were previously not possible.** For instance, one could enable the training of AI scientists instead of AI agents.

likely not be used (in the absence of strong regulation).

- **Reduce the training failure damage** (similar to the fail-safe concept from nuclear technology). Imagine an AI designed to autonomously develop vaccines for a range of viruses that works in 99% of cases. If, in the remaining 1% of cases, it outputs nothing, that's merely annoying; but if it outputs a deadly virus that's indistinguishable from a vaccine, that's catastrophic. Researchers could come up with processes to avoid such catastrophic training failures.

- **Increasing training process success.** With checks and balances, e.g., one could avoid the deceptiveness of AI systems.

- **Enable the (socio-)technical evaluation of AI systems before deployment**. If one had better testing tools and development evaluations, one could better evaluate the value and risks of deploying a particular AI system.

3.3. Technical Research Areas

Major research fields within AI controllability include scalable oversight, interpretability, and alignment theory.[5]

3.3.1 Scalable Oversight

Scalable oversight refers to leveraging AIs that are more capable than humans to oversee and align other AIs. Scalable oversight researchers study how to effectively monitor such intelligent agents and adapt as these agents evolve. The crux of scalable oversight is ensuring that more intelligent agents can competitively oversee one another without them colluding. Another

[5] Other techniques that have been proposed to align neural networks (but also make them more capable) include reinforcement learning from human feedback (RLHF) and inverse reinforcement learning.

challenge is adequately breaking down tasks into smaller subtasks. The research field consists of conceptual work, such as iterated amplification, and empirical work, where (for example) AIs are trained to criticize each other's output or debate the right answer to various questions. See also this overview paper and measurement of scalable oversight. Scalable oversight techniques may be used to prevent deceptive alignment, for example. One open question is how to prevent advanced AI systems from colluding with each other. Game theory combined with AI research may provide insights.

3.3.2 Interpretability

AI interpretability is like "digital neuroscience." It involves studying the internal cognition of large language models (LLMs). The hope is that interpretability will help us to audit the internal cognition of AI models, allowing us to provide training feedback based on the models' reasoning process rather than just the outcomes (which may allow one to get more correct answers or filter out deceptive reasoning). Interpretability research looks for concrete features implemented in neural networks and attempts to design alternative neural networks that are more interpretable. Interpretability research can also partially be done by the models themselves. Caspar et al. (2023) provide a critical overview of the field.

3.3.3 Alignment Theory

(includes microeconomic theory and econometric theory)

This area aims to make conceptual progress on what AI alignment means and how to approach it. One such research area within alignment theory is Eliciting Latent Knowledge (ELK). A very advanced prediction model will, by default, only predict the variable you trained it to predict, usually one that we have data inputs or sensors. However, one could develop techniques to extract implicit information from a prediction model about events that aren't observable but that the model has saved in its “world model".[6] Such information could be important for humans, as it could significantly affect their assessment of the situation and decision-making.

3.4. Economics and Controllability Research Areas

AI controllability is an interdisciplinary field involving among others math, computer science, and philosophy. Some AI controllability problems could benefit from academic economists' skills and expertise, including economic theory modeling, mechanism design, and econometric theory.

AI cooperation and conflict
Conditional Prediction market design
Causality models and theory
Multi-agent training and game theory
AI preference elicitation and social choice

3.4.1 Designing AIs That Competently Navigate Conflict

If we develop several advanced AIs, problems may arise when they conflict. The situation may differ from human agents because AIs are much faster and may have significantly more commitment power. That is, they could commit to future actions as a function of others' behavior. They could change their source code or use reinforcement learning to train themselves, and perhaps they can credibly communicate their commitments to others.

[6] For an LLM, this would be the combination of the memory and all current activations.

A world with high-commitment agents might be much better than the alternative since such agents would be better able to (for example) cooperate in prisoner's dilemma-like situations. At the same time, such agents could make ill-advised, incompatible long-term commitments due to time pressure. If both agents have a lot of commitment power, committing before the other agent does may be preferable, i.e., there may be a first-mover advantage. Miscoordination, mistaken beliefs/games off the equilibrium path may be very undesirable.

Selected Research Areas

- How can we model and understand this commitment race problem? How should we categorize various failure modes?
- What commitments, norms, or Schelling points would lead to outcomes in which AIs do not make rationalizable but risky commitments early on?
- What commitments should be pre-programmed into an AI system?
- How can we not only avoid bad commitment races and make AIs that cooperate if desirable?

3.4.2 Incentivizing Predictive Models

In the current paradigm, strong next-token predictors can be used to predict the next token for input involving words, video, audio, or any other kind of sensor. They could be used to forecast the future states of such sensors conditional on different events, including actions, and enhance human decision-making (Hubinger et al., 2023).

For instance, the usual *proper scoring rules* work only in contexts where a) one's prediction does not change which action will be taken and b) predictions don't change what future prediction questions will be given to the model. Both conditions may be violated in the case of predictive models. For instance, if trained on the proper scoring rule, the AI may be trained to change its predictions such that the future becomes more predictable (i.e., any prediction question has a probability very close to 0 or 1) or predict fixed points, events that become ture after their prediction (Osterheld et al., 2022; Othman and Sandholm, 2010).

- What are the consequences of performativity/reflexivity, and what to do about it? How to reliably use and reward a superhuman predictor?

3.4.3 More Economics & AI Controllability Areas

We briefly mention various other areas at the intersection of economics and alignment or controllability research.

- **The frame of causality can be used to study AI controllability**; see, e.g., Everitt et al. (2023) for an introduction and Carey et al. (2023) for an example.

- **Multi-agent AI training scenarios:** In various AI training regimes, several AIs work together (sometimes with humans) to reach the desired outcome. How can we understand their interactions? How can we ensure that they hold each other accountable and that there isn't any undesired collusion?

- **Preference elicitation and social choice theory:** How can the advanced AI systems or agglomerates be aligned to the whole society once they can be aligned to individual users? Suppose very advanced AI systems can automate most human labor. How should they be used to elicit human preferences and combine them to make society-wide policy decisions or recommendations? Advanced LLMs themselves could support such preference elicitation.