Future-Proof: Monitoring the Development, Deployment, and Impacts of Artificial Intelligence

Anson Ho[1]
[1]Epoch, San Francisco, CA, USA
https://doi.org/10.38126/JSPG220305
Corresponding author: anson@epochai.org
Keywords: artificial intelligence

Executive Summary: Recent developments in Artificial Intelligence (AI) pose a complex challenge for policymakers, who are tasked with regulating a technology which is poorly understood, highly multi-use, of potentially enormous economic impact, and which becomes more powerful at an extraordinary rate. In response to this challenge, this policy position paper outlines two recommended actions for national governments to monitor the AI supply chain: (1) Invest in infrastructure for monitoring the AI supply chain, and (2) establish key AI standards. This will allow policymakers to prepare for current technological challenges, as well as to have the infrastructure for unforeseen ones. Importantly, these recommendations are directly informed by technical research at the frontiers of AI and AI forecasting, to help policymakers make decisions that are robust to future technological changes.

I. Introduction

In March 2023, the artificial intelligence company OpenAI released GPT-4, an AI chatbot that can solve novel problems across a wide range of topics, such as mathematics, medicine, and law (Bubeck et al. 2023), often outperforming humans on certain benchmarks (OpenAI 2023). Due to their task-automation capabilities across a wide range of domains, AI algorithm-based computer systems (“AI systems” for short) like GPT-4 could impact large fractions of the labor force (Eloundou et al. 2023) and lead to significantly increased economic growth rates (Besiroglu et al. 2022; Trammell and Korinek 2023, Hatzius et al. 2023; Aghion et al. 2017).

Developments in AI systems have also been extraordinarily fast and driven by massive increases in investment in resources like computation and training data (Sevilla et al. 2022; Villalobos et al. 2022). The impact of AI systems can diffuse rapidly across society. For instance, the chatbot ChatGPT reached 100 million users within two months after launching (Milmo et al. 2023; Dennean et al. 2023). Moreover, increases in investments (Roser 2023; Maslej et al. 2023), and the ability of AI systems to themselves contribute to further performance improvements (Masa 2023; Madaan et al. 2023; Saunders et al. 2022), suggest that trends of fast progress may continue.

With these rapid changes, the risks of emerging AI technologies can become amplified. A salient example comes from Urbina et al. (2022), who demonstrate how AIs used for drug discovery can be modified to discover novel molecular designs for chemical weapons. Within six hours, the AI system discovered 40,000 designs, with some predicted to be more toxic than any other chemical weapon currently known (Calma 2022). This example further illustrates how AI systems increase the ability of individuals with malicious intent to cause serious harm, using limited resources and technical knowledge, and at a much lower price (Shevlane and Dafoe 2020).

These dangers underscore the urgency for existing legal and political institutions to adapt quickly in order to safely enjoy the benefits while reducing the harms of AI (Clark and Hadfield 2019). Following recent advancements, there has been a surge of interest in regulating AI — for instance, the European Union (EU) set precedent in establishing the AI Act (AI Act 2021), and the Future of Life Institute released an open letter calling for a pause on the training of AI systems more powerful than GPT-4, receiving over 25,000 signatures from academics and industry leaders alike (Future of Life Institute 2023). Despite this interest in regulation, the potential impacts of AI remain highly uncertain in magnitude and probability, making regulation difficult.

Combined, all these factors pose a challenging problem for government policymakers. This article elaborates on these challenges and argues that monitoring AI development is key for overcoming them, providing two recommendations: (1) Invest in infrastructure for monitoring the AI supply chain, and (2) establish key AI standards.

These recommendations are primarily targeted towards national governments, such as the US federal government. Furthermore, they apply most strongly to countries which are at the forefront of AI development (e.g. in terms of the number of state-of-the-art AI systems released per year), such as the US and the UK. That said, the policy recommendations are of broad interest and deliberately presented for implementation by multiple actors across different countries. As such, this paper does not specify which governmental branches or departments should enact the stated policies, and instead opts for providing motivating examples of related policies to illustrate how they may be implemented.

This position paper is therefore of less direct relevance to readers not within the aforementioned target audience, but is valuable nevertheless. In particular, key ideas of the policy proposals depend mostly on the properties of AI systems (such as their intensive resource requirements), rather than on country-specific policy environments. This paper is therefore also relevant to international organizations monitoring AI developments (e.g., the Organisation for Economic Co-operation and Development AI Futures Expert Group) and policymakers at a more local scale (e.g., state or provincial governments).

II. The development, deployment, and impacts of AI

i. The landscape of risks

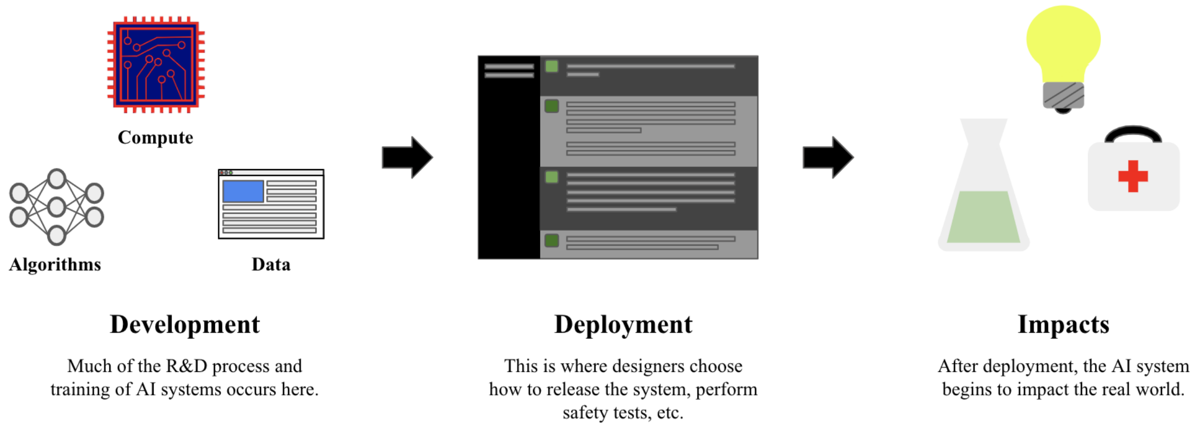
To understand how to appropriately regulate these emerging technologies, it is necessary to consider the risks that AI poses, where AI systems are defined as *machine-based systems which optimize objective function(s) to generate outputs (e.g., text)*. These pose three main categories of risks, adapted from a framework by Zwetsloot and Dafoe (2019):

1) **Misuse risks:** Risks from the malicious or otherwise unethical use of AI systems, such as the use of autonomous weapons systems for targeted assassination (Trager 2022), mass disinformation campaigns (Goldstein et al. 2023), and the use of AI in cyber attacks (Brundage et al. 2018).
2) **Misalignment risks:** Risks from AI systems that are not aligned with human values. This can manifest in a wide range of ways, such as amplifying biases towards vulnerable groups (Christian 2020; Bender et al. 2021), or in AI systems that deliberately deceive humans and autonomously take harmful actions (Shah et al. 2022; Ngo et al. 2023; Krakovna et al. 2020). These issues are especially problematic if powerful, misaligned AI systems are deployed in high-stakes scenarios, such as allowing AI to control weapons systems (Dafoe 2018; Ziegler et al. 2022).
3) **Structural risks:** Misuse and misalignment risks exist within a broader landscape (Zwetsloot and Dafoe 2019). Structural risks describe how AI systems influence structural environments and incentives, and how these affect the development of AI in return. One example concern is that stiff competition between AI labs could lead to races to market, deprioritizing safety and thus magnifying misuse and misalignment risks (Scharre 2021; Dafoe 2020; Emery-Xu et al. 2022).

It is necessary to avoid all three sources of risk in order to support the development of AI systems that are safe, fair, and beneficial. It is also crucial to note that these risks become significantly amplified as AI systems become more powerful. Moreover, systems at the frontier of AI development are often the ones that have the greatest societal impacts, such as the previous example of ChatGPT reaching a hundred million users within two months. Thus, the most important systems to consider are at the forefront of AI development.

When these potential risks are combined with a lack of clarity about rates of progress and relevant details of the AI supply chain, government policymakers are faced with a challenge. They need to (1) make important decisions under time pressure, (2) develop strategies that can adapt to the rapid rate of change of technologies (i.e., "future-proof" strategies), and (3) make informed decisions in a field where collective understanding of the field's impacts is very limited (Clark and Hadfield 2019). These challenges can be a serious bottleneck for successful regulation, as seen by Canadian Members of Parliament during recent readings of Bill C-27 (Arai 2022; Government of Canada 2022). In particular, a major criticism of the proposed AI portion of the bill is that it does not "have a shell of a framework for responsive artificial intelligence regulation and oversight" (Williams 2022). The simultaneous lack of clarity regarding issues and developments in AI, combined with the need for careful decision-making, are fundamental issues for policymakers.

Figure 1. The three stages of the AI supply


Compute
Algorithms
Data
Development
Much of the R&D process and training of AI systems occurs here.
Deployment
This is where designers choose how to release the system, perform safety tests, etc.
Impacts
After deployment, the AI system begins to impact the real world.

ii. The AI supply chain

To understand how to mitigate risks, we need to understand the AI supply chain, which can be split into three stages (Figure 1):

1) **Development**: This encompasses investments in computation, data, and algorithms. Importantly, it involves the "training" of an AI system, by which these three resources are funneled into building a functioning AI system.
2) **Deployment**: After training the system, organizations perform safety and performance checks, decide how to release the AI system, etc.
3) **Impacts**: This occurs when AI systems interact with users or society more generally.

Note that this framework is a simplification. For instance, systems such as ChatGPT may be continuously developed even after deployment (OpenAI 2022). The risks considered in section II, subsection *i.* apply to all three stages to different degrees, and this is elaborated upon by considering each stage in turn.

Development

At a high level, the development stage is driven by three main inputs: computation, data, and algorithms. Historically, many improvements in AI capabilities have been driven by using the above factors at large scales — the largest and most powerful AI systems today are trained using 10^{25} operations, trillions of words of text, and contain hundreds of billions of parameters (Epoch 2023).

These inputs are in turn driven by other factors — monetary investments, for example, are key to supporting large computational budgets, with GPT-4 costing around USD $40 million to train (Epoch 2023). Moreover, monetary investments have been increasing — since 2012, eight of the top AI companies that are explicitly trying to build "Artificial General Intelligence" have received USD $21 billion in investment, including $11 billion in the first three months of 2023 alone (Hogarth 2023).

The development stage is significant because it defines the capabilities of trained systems, which are difficult to subsequently adjust. In general, the more capable a system is, the greater the potential for positive benefits and serious harms via the three risk vectors outlined earlier. Therefore, the development stage serves as an important node for intervention, where governments have a clear potential lever in controlling the flow of these resources and their use.

Deployment

After training AI systems, they are tested on benchmarks such as ImageNet (Deng et al. 2009) and WikiText-103 (Merity et al. 2016) to evaluate their capabilities. Engineers of the system will then decide how to release the model. For example, GPT-4 was released via an online interface rather than made open source (OpenAI 2023), and OpenAI has more generally followed a "staged release" strategy for their systems since 2019. This involves first releasing smaller, less capable versions of systems in restricted fashion to first evaluate social impacts, before allowing access to the full version (Solaiman et al. 2019).

This stage is important for defining how quickly and in what form AI systems impact society, as well as for applying safety checks. For instance, OpenAI reportedly spent six months on safety research and evaluation prior to the deployment of GPT-4, such as to check if the system had autonomous replication abilities (OpenAI 2023; Alignment Research Center 2023).

However, labs may not evaluate their systems effectively or may choose deployment strategies that are inappropriate for the magnitude of risks, such as open-sourcing AI systems that can be used for bioweapons design. This could amplify all three kinds of risk, making monitoring at this stage critical.

Impacts

The third and final stage is the "impacts" stage, where AI systems begin to exert a broader impact on society. These impacts could be positive, such as potential increases in economic growth rates (Trammell and Korinek 2023; Eloundou et al. 2023), or negative, such as increases in cyberattacks (Brundage et al. 2018). Even with extensive checks prior to deployment, it is nearly impossible to anticipate every potential issue in advance. Therefore, adjusting policies and standards based on real-world feedback is critical.

III. Why monitor AI?

With these considerations about the risks and supply chain of AI, we now turn our attention to approaches for effective AI governance and policy. In particular, governments should monitor the AI supply chain across all three stages, gathering data about key metrics to form a bedrock AI policy via two categories of benefits: (1) providing a framework for supporting and evaluating policies, and (2) understanding the probability and magnitude of impacts.

i. Supporting AI policy via information collection

Properly regulating AI can be challenging without understanding how AI works, and how it impacts society. Monitoring the AI supply chain can help alleviate some of these difficulties by adopting a crucial role in problem identification and policy evaluation, both of which are crucial to the policy process (Brewer 1974). This section elaborates on several ways in which monitoring AI can help support effective policy.

Identifying unresolved AI policy issues

Effectively monitoring the AI supply chain can help policymakers spot issues that need to be addressed, and which need to be addressed with greater urgency. For example, Benaich and Hogarth (2022) estimate there were 300 researchers working on AI safety at large labs in 2022, which pales in comparison to the number of AI researchers more generally. The 2021 NeurIPS machine learning conference alone had almost 9000 researchers, exceeding the number of safety researchers by over an order of magnitude. Given that AI systems can be deployed on a large scale across many users, this presents a serious policy issue that needs to be addressed, and that is first identified through monitoring the supply chain.

Identifying levers for policy

Policy approaches may depend on specific details of the AI supply chain, and gathering information about the supply chain (e.g., where semiconductor chips are being produced) may be highly valuable. For instance, the field of Compute Governance (Anderljung et al. 2022; Whittlestone et al. 2023) depends on the empirical finding that the most powerful AI systems leave a large physical footprint, requiring thousands of units of specialized hardware to run continuously for several months. This provides insight into a potential policy lever: controlling access to computational resources, thereby influencing which actors are able to train powerful AI systems (Shavit 2023).

Evaluating the effect of policies

An essential step in the policy process is evaluating the intervention after implementation. This certainly applies to AI policy too, where evaluations may help inform future policies. One example from corporate policy is OpenAI's experimentation with staged release strategies for GPT-2, which influenced how they released their later systems (Hao 2019; Solaiman et al. 2019). The benefits of policy evaluation are however not limited to corporate policy — for instance, monitoring the impact of the US CHIPS and Science Act will undoubtedly influence future policy approaches towards AI exports (Bureau of Industry and Security 2022; Kannan and Feldgoise 2022).

ii. Understanding future impacts

The field of AI is characterized by a rapid rate of progress, as well as difficulties in adopting policy strategies that can adapt to these changes. This issue can be mitigated with the data obtained from monitoring, which can be used to forecast future AI impacts.

Estimating the probability of future risks

Understanding risk probabilities helps inform the urgency of regulation. This can involve forecasting how capabilities may evolve in the future based on existing trends in hardware and software (Cotra 2020; Davidson 2023), information about which can be obtained from monitoring the development stage. This can also simply involve closely tracking the state of the art in model capabilities on benchmarks. For instance, given the often human-competitive benchmark performance of GPT-4, it is natural to expect an increase in broad societal impacts and some risks in the near term, as GPT-4 becomes increasingly widely adopted (e.g., in cybersecurity).

Estimating the magnitude of future impacts

While estimating the probabilities of risks is important, this in itself is insufficient. Some risks (e.g., nuclear war) may have low probabilities, but impacts large enough to be worth great attention (Shulman and Thornley 2023; Ord 2020). Whether this applies to certain AI risks will depend on the precise scenario in question.

Understanding how future risks may arise

In order to adapt to fast rates of progress, governments must be prepared for a range of possible scenarios. This process can be aided by analyzing how these futures may unfold. For instance, it could be crucial to analyze how AI could impact the "offense-defense balance" in high-risk domains such as synthetic biology. That is, AI could magnify the ability to cause harm more than the ability for defenders to prevent these harms (Shevlane and Dafoe 2020). Whether or not AI affects the offense-defense balance of specific domains will depend on how AI is applied in these domains. Understanding this balance can help manage novel risks that may emerge during the impacts stage, and how these can be mitigated in the development and deployment stages.

One might object to these benefits on the grounds that the future is hard to predict. However, this view is problematic. While it is true that forecasting the future in precise detail is extremely difficult, it is possible to forecast general effects, as has been the case historically. For example, the DICE integrated assessment model (Nordhaus 1992) has been hugely influential for handling climate change, such as in analyzing the effects of the Kyoto Protocol (Nordhaus and Boyer 1999). Clear trends have also existed in computer science such as through Moore's Law (Moore 1965), helping forecast future developments in semiconductor technology (Waldrop 2016). In fact, analogous empirical trends and "scaling laws" from the AI literature (Kaplan et al. 2020; Hoffmann et al. 2022; Villalobos 2023; Epoch 2023) suggest that dismissal of *any* ability to usefully forecast AI developments is at the very least premature.

VI. Recommendations: monitoring the field of AI

This section outlines two recommendations for national governments, particularly for countries at the frontier of AI development, to effectively monitor AI. This does not imply that these policies are irrelevant to other policymakers — in fact, the implementation of these policies may involve interactions between different levels of government (e.g., with provincial governments, like the government of Québec), or across the governments of different nations (e.g., the Organisation for Economic Co-operation and Development).

The listed recommendations are informed by three main desiderata: (1) recommendations should be grounded in the AI literature, (2) they should focus primarily on the most relevant variables, and (3) they should be tractable (e.g., it can be costly or prohibitive to gather certain types of data).

i. National governments should invest in infrastructure for monitoring the AI supply chain. Multiple countries, such as the UK and Japan, are currently designing national AI strategies (Secretary of State for Digital, Culture, Media and Sport 2021; Department for Science, Innovation and Technology et al. 2023; Cabinet Office, Government of Japan 2022). However, governments may lack the information to enforce appropriate regulations, and as such investing in AI monitoring infrastructure is vital for gaining relevant information and continuously observing developments (Clark et al. 2021).

Importantly, it is most pertinent to monitor parts of the supply chain that are either associated with the highest risks, or are otherwise of highest leverage. For instance, an essential component of the "development" stage of the AI supply chain is when the AI system is trained, since this is when the system gains powerful capabilities that can be both beneficial and harmful. To monitor this, governments could gather data about certain predictors of performance, such as total training compute (the number of computational operations performed during training) and the total amount of training data, identifying which actors are involved in the largest-scale experiments and proxy model capabilities in real-world tasks. Notably, obtaining this information is cheap and unintrusive, and yet is shown to be of robust importance in understanding rates of progress and identifying policy issues. For instance, simply tracking training compute information has drawn attention to the "compute divide": computationally-intensive large AI systems are affordable to industry labs, but not to academic groups (Benaich and Hogarth 2022; Anderljung et al. 2022). This, therefore, provides a path for governments to tackle these inequalities in compute access. More example variables in Appendix A.

Focusing on high-risk parts of the supply chain has the benefit of reducing complexity and intrusiveness, compared to monitoring the entire complex system. This can make it easier for governments to form teams of experts to check that actors follow appropriate safety measures (NISSTC 2021), or evaluate the safety of commonly used AI resources (e.g., code, trained models, datasets) which may be a failure point for a large number of users (Lohn 2021). Moreover, monitoring all AI systems would be both wildly impractical and unethical, since this would likely require draconian measures to monitor all electronic devices which use AI in any form. Fortunately, this is not necessary, and thus emphasis is placed on monitoring AI systems *at the forefront of developments*, given their elevated impacts and levels of associated risk.

There are multiple ways this could be achieved, depending on case-by-case feasibility. For instance, governments could establish new observatories to explicitly monitor the national AI landscape, or to build on top of existing monitoring systems in related domains (such as cybersecurity; Whittlestone and Clark 2021). Some governments already have observatories of some kind in place, such as the KI-Observatorium in Germany and the government of Québec (Observatory on Artificial Intelligence in Work and Society 2023; International Observatory on the Social Impacts of Artificial and Digital Intelligence in Canada 2023). In these cases, governments could consider expanding the scope of these existing observatories to be in line with technological developments, and to more explicitly monitor the development and deployment stages of the supply chain.

Setting up this core infrastructure will likely be quite cheap for governments, compared to the investments that may be required for other interventions. For instance, Ord et al. (2021) estimate that continually assessing AI capabilities and societal impacts (e.g., through the formation of new government bodies), as well as funding of research projects, may collectively cost around USD $750,000 annually. Challenges to successful implementation are hence plausibly bottlenecked by practical difficulties (such as determining which metrics are most relevant to a given policy issue), rather than financial costs. The same is likely true of our other recommendations, which are of a similar vein.

ii. National governments should establish key AI standards.

These standards include those for testing procedures, best practices, and safety certification of AI models (Frase 2023). This intervention addresses a number of existing issues — for one, there is an absence of norms for providing certain kinds of useful metrics (e.g., total training computation) in the AI community (Sevilla et al. 2023; Hooker 2020; Thompson et al. 2020), and gathering this information is non-trivial (Sevilla et al. 2022). Having clear standards for properties of AI systems or stages in the AI supply chain can help provide a stronger basis for evaluating AI progress and enforcing regulations.

Evaluating trained AI systems is already common practice since researchers typically wish to compare the performance of different AI models. Governments could thus take advantage of this common practice by creating standard benchmarks to test performance and capabilities, including in safety-relevant metrics (e.g. the ETHICS dataset (Hendrycks et al. 2020)). Governments could also push for transparency regarding key metrics relevant to AI training, by collaborating with conferences and journals. For instance, top AI conferences like NeurIPS require that authors of submitted publications specify information relating to the total training compute, but these requirements are vague — these could be made more strict to report more precise metrics (Sevilla et al. 2023).

Implementing a set of AI standards would likely involve collaboration with other parties who possess expertise in testing and benchmarking existing models. For example, top AI labs like Anthropic, Google, Microsoft, and OpenAI have expressed interest in identifying best safety practices for frontier models and sharing information with policymakers via the formation of a “Frontier Model Forum” (OpenAI 2023). This can help determine the viability of specific AI model standards, and which properties can or cannot be guaranteed, e.g., “bias” and “safety”. Collaboration with regulators like the National Institute of Standards and Technology (Arnold and Toner 2021; NIST 2023) could also be fruitful for creating a testbed of code for AI system evaluation. The example of NIST suggests a possible implementation of standards via the executive branch of government, but as mentioned in the introduction, this is not a necessary criterion for implementation.

Designing robust AI standards would be a significant step forward for AI policy, but it would likely be challenging to do so. For instance, trained AI systems can often exhibit unpredictable behavior, which experts are unable to anticipate or effectively mitigate (Bowman 2023). Under this view, it is perhaps unsurprising that there is as yet no widely accepted definition of key terms like “safe AI” (Arnold and Toner 2021).

To ensure compliance, governments could potentially establish communication channels between private sectors and governments, by which information about AI misuse or accidents can be reported (Arnold and Toner 2021). This could be based on existing institutions in other domains, such as Information Sharing and Analysis Centers, which are coordinating bodies that help disseminate information about cyber or physical threats between governments and the private sector (National Council of ISACs 2022).

iii. Potential limitations

A general caveat to these recommendations is that the specifics will likely depend on the context of the policymaker (Frase 2023). To ensure that government interventions are robustly useful, monitoring approaches may need to be informed by specific policy issues. For instance, prevention of AI-driven bioweapons development may require monitoring different aspects of the supply chain compared to management of cybersecurity issues (Whittlestone and Clark 2021).

There are multiple possible concerns with monitoring AI developments, such as potential pushback from the private sector and disincentives for innovation. This is why the recommendations were primarily aimed at specific parts of the supply chain that are particularly high-risk, to keep intervention relatively limited while still gathering information that is useful for risk mitigation. The significance of this difficulty varies somewhat depending on what exactly is being monitored — monitoring total training compute requirements for trained systems is relatively cheap and nonintrusive, but more stringent monitoring (e.g., of adherence to certain standards for hardware designs) could see stronger pushback from companies.

Monitoring AI systems may also bring unintended consequences. Metrics that are intended to *measure* rates of progress may instead turn into *targets*, and there are anecdotal accounts of this being the case historically for Moore’s Law (Schaller 1997). Analogously, measures of growth rates in training computation may inadvertently turn into goals that AI labs try to attain, thus inadvertently altering rates of technological development.

While important, the ultimate significance of these caveats pales in comparison to the issues that would likely arise in the absence of appropriate monitoring of the AI supply chain. In this counterfactual, policymakers are faced with regulating AI while lacking concrete details about AI risks and its drivers. Actors in the private sector may try to take advantage of deficiencies in policy by deploying systems that are profitable without fully accounting for the social costs. Hence, while monitoring AI is insufficient in itself to prevent serious AI risks, it is almost certainly a critical component of any effective AI governance strategy. In fact, establishing good monitoring infrastructure could result in even *more* innovation compared to the counterfactual where serious risks could arise.

V. Conclusion

In this position paper, two policy options are recommended to tackle the challenge of the rapidly changing AI landscape: (1) Invest in infrastructure for monitoring the AI supply chain, and (2) establish key AI standards. These approaches for monitoring AI help form the basis of a “future-proof” policy strategy, allowing policymakers to prepare for current technological challenges as well as future unforeseen ones. While monitoring alone is likely insufficient for mitigating AI risks, it is surely necessary for any effective AI governance approach.

Appendix A: Example guiding questions

Stage	Approach	Example guiding questions
Development	**Monitoring key inputs to AI systems**	**Compute:** How much training computation was needed to train the model? **Data:** How much data was used to train the model? Which dataset was used? **Algorithms:** How many parameters are there in the AI model? Is the system based on a pre-trained foundation model (Bommasani et al. 2021) that is fine-tuned to a particular task? **Hardware:** What kind of hardware was used, and in what quantity? How long was the system trained for? **Investment:** What were the total costs of training? **Labor:** Which actors are training large AI systems, and for what purposes?
Deployment	**Analyzing AI capabilities**	How well does the trained system perform on standard benchmarks? Does the model demonstrate qualitatively similar performance across multiple benchmarks?

	Regulatory compliance, ethics, and safety	Were checks established and implemented to ensure safe and reliable model behavior, such as red-teaming (OpenAI 2023; Alignment Research Center 2023)? Are the designers of the system cutting corners on safety due to profit incentives, or competitive pressures?
Impacts	**Monitoring the use of AI**	Which actors are using AI systems? How many people, and how quickly has the use of these systems spread? How are AI systems being used in high-risk domains, such as nuclear security, biotechnology, or cybersecurity?
	Monitoring hazards and accidents	What notable misuse, misalignment, or structural risks have occurred historically? (For an example of a collection of misalignment risks, see Krakovna (2018))

References

Anderljung, Markus, Lennart Heim, and Toby Shevlane. "Compute Funds and Pre-Trained Models: Govai Blog." RSS. Accessed May 13, 2023. https://www.governance.ai/post/compute-funds-and-pre-trained-models.

Aghion, Philippe, Benjamin Jones, and Charles Jones. Artificial Intelligence and economic growth, 2017. https://doi.org/10.3386/w23928.

Arai, Maggie. "Five Things to Know about Bill C-27." Schwartz Reisman Institute, April 17, 2023. https://srinstitute.utoronto.ca/news/five-things-to-know-about-bill-c-27.

ARC Evals. "Update on Arc's Recent Eval Efforts." Update on ARC's recent eval efforts - ARC Evals. Accessed May 13, 2023. https://evals.alignment.org/blog/2023-03-18-update-on-recent-evals/.

Arnold, Zachary, and Helen Toner. "Ai Accidents: An Emerging Threat." Center for Security and Emerging Technology, April 12, 2022. https://cset.georgetown.edu/publication/ai-accidents-an-emerging-threat/.

Benaich, Nathan, and Ian Hogarth. "State of Ai Report 2022." State of AI Report 2022, October 11, 2022. https://www.stateof.ai/.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots." Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021. https://doi.org/10.1145/3442188.3445922.

Besiroglu, Tamay, Nicholas Emery-Xu, and Neil Thompson. "Economic Impacts of AI-Augmented R&D." arXiv.org, January 2, 2023. https://arxiv.org/abs/2212.08198.

Bowman, Samuel R. "Eight Things to Know about Large Language Models." arXiv.org, April 2, 2023. https://arxiv.org/abs/2304.00612.

Brewer, Garry D. "The Policy Sciences Emerge: To Nurture and Structure a Discipline." Policy Sciences 5, no. 3 (1974): 239–44. https://doi.org/10.1007/bf00144283.

Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, et al. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation." arXiv.org, February 20, 2018. https://arxiv.org/abs/1802.07228.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, et al. "Sparks of Artificial General Intelligence: Early Experiments with GPT-4." arXiv.org, April 13, 2023. https://arxiv.org/abs/2303.12712.

Calma, Justine. "AI Suggested 40,000 New Possible Chemical Weapons in Just Six Hours." The Verge, March 17, 2022. https://www.theverge.com/2022/3/17/22983197/ai-new-possible-chemical-weapons-generative-models-vx.

Christian, Brian. The alignment problem: machine learning and human values. New York, NY: W.W. Norton & Company, 2021.

Clark, Jack, and Gillian K. Hadfield. "Regulatory Markets for AI Safety." arXiv.org, December 11, 2019. https://arxiv.org/abs/2001.00078.

Clark, Jack, Kyle Augustus Miller, and Rebecca Gelles. "Measuring AI Development." Center for Security and Emerging Technology, January 17, 2023. https://cset.georgetown.edu/publication/measuring-ai-development/.

Cotra, Ajeya. "Draft Report on Ai Timelines." AI Alignment Forum, September 18, 2020. https://www.alignmentforum.org/posts/KrJfoZzpSDpnrv9va/draft-report-on-ai-timelines.

Dafoe, Allan. "AI Governance: A Research Agenda - Future of Humanity Institute." Centre for the Governance of AI, August 27, 2018.

https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf.

Davidson, Tom. "What a Compute-Centric Framework Says about AI Takeoff Speeds - Draft Report." AI Alignment Forum, January 22, 2023. https://www.alignmentforum.org/posts/Gc9FGtdXhK9sCSEYu/what-a-compute-centric-framework-says-about-ai-takeoff.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "ImageNet: A Large-Scale Hierarchical Image Database." 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009. https://doi.org/10.1109/cvpr.2009.5206848.

Dennean, Kevin, Sundeep Gantori, Delwin Kurnia Limas, Allen Pu, and Reid Gilligan. "Let's Chat about ChatGPT." UBS Editorial, September 24, 2019. https://www.ubs.com/global/en/wealth-management/our-approach/marketnews/article.1585717.html.

Department for Science, Innovation and Technology, Chloe Smith, and Rishi Sunak. "Tech Entrepreneur Ian Hogarth to Lead UK's AI Foundation Model Taskforce." GOV.UK, June 18, 2023. https://www.gov.uk/government/news/tech-entrepreneur-ian-hogarth-to-lead-uks-ai-foundation-model-taskforce.

Department for Science, Innovation and Technology. "National AI Strategy." GOV.UK, December 18, 2022. https://www.gov.uk/government/publications/national-ai-strategy.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. "GPTs Are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models." arXiv.org, March 23, 2023. https://arxiv.org/abs/2303.10130.

Emery-Xu, Nicholas, Andrew Park, and Robert Trager. "Uncertainty, Information, and Risk in International Technology Races (Working Paper).Pdf." Accessed May 13, 2023. https://drive.google.com/file/d/18j_wnA4HDMA3ofclLcfpgyV-0INMn1ZW/view.

Epoch. "ML Trends." Epoch, April 11, 2023. https://epochai.org/trends.

European Commission. "Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS." EUR-Lex, April 21, 2021. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

Frase, Heather. "One Size Does Not Fit All." Center for Security and Emerging Technology, February 22, 2023. https://cset.georgetown.edu/publication/one-size-does-not-fit-all/.

Future of Life Institute. "Pause Giant AI Experiments: An Open Letter." Future of Life Institute, May 5, 2023. https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

Goldstein, Josh A., Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. "Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations." arXiv.org, January 10, 2023. https://arxiv.org/abs/2301.04246.

Government of Canada, Department of Justice. "Charter Statement Bill C-27: An Act to Enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to Make Consequential and Related Amendments to Other Acts." Government of Canada, Department of Justice, Electronic Communications, November 10, 2022. https://www.justice.gc.ca/eng/csj-sjc/pl/charter-charte/c27_1.html.

Government of Japan, Cabinet Office. "ＡＩ戦略２０２２－内閣府." cao.go.jp, April 4, 2022. https://www8.cao.go.jp/cstp/ai/aistrategy2022_honbun.pdf.

Hadfield, Gillian K., and Jack Clark. "Regulatory Markets: The Future of AI Governance." arXiv.org, April 25, 2023. https://arxiv.org/abs/2304.04914.

Hao, Karen. "OpenAI Has Released the Largest Version yet of Its Fake-News-Spewing AI." MIT Technology Review, April 2, 2020. https://www.technologyreview.com/2019/08/29/133218/openai-released-its-fake-news-ai-gpt-2/

Hatzius, Jan, Joseph Briggs, Davesh Kodnani, and Giovanni Pierdomenico. "The Potentially Large Effects of Artificial Intelligence on Economic Growth". Global Economics Analyst, March 26, 2023. https://www.key4biz.it/wp-content/uploads/2023/03/Global-Economics-Analyst_-The-Potentially-Large-Effects-of-Artificial-Intelligence-on-Economic-Growth-Briggs_Kodnani.pdf.

Hendrycks, Dan, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. "Aligning AI with Shared Human Values." arXiv.org, February 17, 2023. https://arxiv.org/abs/2008.02275.

Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, et al. "Training Compute-Optimal Large Language Models." arXiv.org, March 29, 2022. https://arxiv.org/abs/2203.15556.

Hogarth, Ian. "We Must Slow down the Race to God-like Ai." Financial Times, April 13, 2023.

https://www.ft.com/content/03895dc4-a3b7-481e-95cc-336a524f2ac2.

Hooker, Sara. "The Hardware Lottery." Communications of the ACM 64, no. 12 (2021): 58–65. https://doi.org/10.1145/3467017.

International observatory on the societal impacts of AI and digital technology. Accessed August 6, 2023. https://observatoire-ia.ulaval.ca/en/.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. "Scaling Laws for Neural Language Models." arXiv.org, January 23, 2020. https://arxiv.org/abs/2001.08361.

Krakovna, Victoria, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. "Specification gaming: the flip side of AI ingenuity". Google DeepMind, April 21, 2020. https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity.

Lohn, Andrew J. "Poison in the Well." Center for Security and Emerging Technology, March 2, 2023. https://cset.georgetown.edu/publication/poison-in-the-well/.

Madaan, Aman, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, et al. "Self-Refine: Iterative Refinement with Self-Feedback." arXiv.org, March 30, 2023. https://arxiv.org/abs/2303.17651.

Masa. "Revolutionizing Industries with Unparalleled Applications." AutoGPT, May 4, 2023. https://autogpt.net/autogpt-revolutionizing-industries-with-unparalleled-applications/.

Maslej, Nestor, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and Raymond Perrault. "The AI Index 2023 Annual Report," AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2023. https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf.

Merity, Stephen, Caiming Xiong, James Bradbury, and Richard Socher. "Pointer Sentinel Mixture Models." arXiv.org, September 26, 2016. https://arxiv.org/abs/1609.07843.

Milmo, Dan and agency. "ChatGPT Reaches 100 Million Users Two Months after Launch." The Guardian, February 2, 2023. https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app.

Mittelsteadt, Matthew. "AI Verification." Center for Security and Emerging Technology, January 25, 2023. https://cset.georgetown.edu/publication/ai-verification/.

Moore, Gordon E. "Cramming More Components onto Integrated Circuits, Reprinted from Electronics, Volume 38, Number 8, April 19, 1965, Pp.114 Ff." IEEE Solid-State Circuits Society Newsletter 11, no. 3 (2006): 33–35. https://doi.org/10.1109/n-ssc.2006.4785860.

National Council of ISACs, 2022. https://www.nationalisacs.org/about-nci.

Ngo, Richard, Lawrence Chan, and Sören Mindermann. "The Alignment Problem from a Deep Learning Perspective." arXiv.org, February 22, 2023. https://arxiv.org/abs/2209.00626.

Nordhaus, William D. "The 'dice' Model: Background and Structure of a Dynamic Inte." Cowles Foundation Discussion Papers, February 2, 1992. https://ideas.repec.org/p/cwl/cwldpp/1009.html

Nordhaus, William D., and Joseph G. Boyer. "Requiem for Kyoto: An Economic Analysis of the Kyoto Protocol." The Energy Journal 20, no. 01 (1999). https://doi.org/10.5547/issn0195-6574-ej-vol20-nosi-5.

OpenAI. "Introducing Chatgpt." Introducing ChatGPT, November 30, 2022. https://openai.com/blog/chatgpt.

OpenAI. GPT-4, March 14, 2023. https://openai.com/research/gpt-4.

OpenAI. "GPT-4 System Card." OpenAI, March 23, 2023. https://cdn.openai.com/papers/gpt-4-system-card.pdf.

OpenAI. "GPT-4 Technical Report." arXiv.org, March 27, 2023. https://arxiv.org/abs/2303.08774.

OpenAI. Frontier Model Forum, July 26, 2023. https://openai.com/blog/frontier-model-forum.

Ord, Toby. The precipice: Existential risk and the future of humanity. New York: Hachette Books, 2021.

Ord, Toby, Angus Mercer, and Sophie Dannreuther. "Future Proof: The Opportunity to Transform the UK's Resilience to Extreme Risks." The Centre for Long-Term Resilience. June, 2021. https://11f95c32-710c-438b-903d-da4e18de8aaa.filesusr.com/ugd/e40baa_c64c0d7b430149a393236bf4d26cdfdd.pdf.

Roser, Max. "Artificial Intelligence Has Advanced despite Having Few Resources Dedicated to Its Development – Now Investments Have Increased Substantially." Our World in Data, March 29, 2023. https://ourworldindata.org/ai-investments.

Saunders, William, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. "Self-Critiquing Models for Assisting Human Evaluators." arXiv.org, June 14, 2022. https://arxiv.org/abs/2206.05802.

Schaller, R.R. “Moore’s Law: Past, Present and Future.” IEEE Spectrum 34, no. 6 (1997): 52–59. https://doi.org/10.1109/6.591665.

Scharre, Paul. “Debunking the AI Arms Race Theory.” Texas National Security Review, August 18, 2021. https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/.

Sevilla, Jaime, Anson Ho, and Tamay Besiroglu. “Please Report Your Compute.” Communications of the ACM 66, no. 5 (2023): 30–32. https://doi.org/10.1145/3563035.

Sevilla, Jaime, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. “Compute Trends across Three Eras of Machine Learning.” arXiv.org, March 9, 2022. https://arxiv.org/abs/2202.05924.

Sevilla, Jaime, Lennart Heim, Marius Hobbhahn, Tamay Besiroglu, Anson Ho, and Pablo Villalobos. “Estimating Training Compute of Deep Learning Models.” Epoch, January 20, 2022. https://epochai.org/blog/estimating-training-compute.

Shah, Rohin, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. “Goal Misgeneralization: Why Correct Specifications Aren’t Enough for Correct Goals.” arXiv.org, November 2, 2022. https://arxiv.org/abs/2210.01790.

Shavit, Yonadav. “What Does It Take to Catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring.” arXiv.org, March 20, 2023. https://arxiv.org/abs/2303.11341.

Shevlane, Toby, and Allan Dafoe. “The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?” GovAI, February 8, 2020. https://www.governance.ai/research-paper/the-offense-defense-balance-of-scientific-knowledge-does-publishing-ai-research-reduce-misuse.

Shulman, Carl, and Elliott Thornley. “Carl Shulman &amp; Elliott Thornley, How Much Should Governments Pay to Prevent Catastrophes? Longtermism’s Limited Role.” PhilPapers. Accessed May 13, 2023. https://philpapers.org/rec/SHUHMS.

Solaiman, Irene, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, et al. “Release Strategies and the Social Impacts of Language Models.” arXiv.org, November 13, 2019. https://arxiv.org/abs/1908.09203.

Tabassi, Elham. “Ai Risk Management Framework.” National Institute of Standards and Technology, January 2023. https://doi.org/10.6028/nist.ai.100-1.

Observatory on Artificial Intelligence in Work and Society. Accessed August 6, 2023. https://www.ki-observatorium.de/en/.

Thompson, Neil C., Kristjan Greenewald, Keeheon Lee, and Gabriel F. Manso. “The Computational Limits of Deep Learning.” arXiv.org, July 27, 2022. https://arxiv.org/abs/2007.05558.

Trager, Robert. “Deliberating Autonomous Weapons.” Issues in Science and Technology, July 28, 2022. https://issues.org/autonomous-weapons-russell-forum/.

Trammell, Philip and Korinek, Anton. “Economic Growth under Transformative AI.” 2023. https://philiptrammell.com/static/economic_growth_under_transformative_ai.pdf.

Urbina, Fabio, Filippa Lentzos, Cédric Invernizzi, and Sean Ekins. “Dual Use of Artificial-Intelligence-Powered Drug Discovery.” Nature Machine Intelligence 4, no. 3 (2022): 189–91. https://doi.org/10.1038/s42256-022-00465-9.

Villalobos, Pablo, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. “Will We Run out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning.” arXiv.org, October 26, 2022. https://arxiv.org/abs/2211.04325.

Villalobos, Pablo. “Scaling Laws Literature Review.” Epoch, January 26, 2023. https://epochai.org/blog/scaling-laws-literature-review.

Waldrop, M. Mitchell. “The Chips Are down for Moore’s Law.” Nature 530, no. 7589 (2016): 144–47. https://doi.org/10.1038/530144a.

Whittlestone, Jess, and Jack Clark. “Why and How Governments Should Monitor AI Development.” arXiv.org, August 31, 2021. https://arxiv.org/abs/2108.12427.

Whittlestone, Jess, Shahar Avin, Lennart Heim, Markus Anderljung, and Girish Sastry. “Response to the UK’s Future of Compute Review.” GovAI, March 29, 2023. https://www.governance.ai/research-paper/response-to-the-uks-future-of-compute-review.

Williams, Ryan. “Debates (Hansard) No. 125 - November 4, 2022 (44-1) - House of Commons of Canada.” Debates (Hansard) No. 125 - November 4, 2022 (44-1) - House of Commons of Canada, November 4, 2022. https://www.ourcommons.ca/DocumentViewer/en/44-1/house/sitting-125/hansard#11908667.

Xiaoli, Shangguan, Hu Ying, Hao Chunliang, Zhang Yuguang, Su Hang, Hu Songzhi, Yang Tao, Jing Huiyun, Zhang Xudong, Xu Xiaogeng, Gu Zhaoquan, Wu Yuesheng, Meng Guozhu, Li Shi, Fu Yingbo, Mei Jingqing, Wang Le, Dong Yinpeng, Liu Xize, Wang Zhelin, Zhao Yunwei, Han Han, Zhang

Xia, Peng Juntao, Xu Yongtai, Zhang Yi, Xu Yuqing, Wu Baoyuan, Han Lei, and Wang Bingzheng. "Information security technology-Security specification and assessment methods for machine learning algorithms." National Standard of the People's Republic of China. July 27, 2021. https://www.tc260.org.cn/file/2021-08-04/6b5 3 0404-858b-4c9d-8d89-a83239ec5712.pdf.

Ziegler, Daniel M., Seraphina Nix, Lawrence Chan, Tim Bauman, Peter Schmidt-Nielsen, Tao Lin, Adam Scherlis, et al. "Adversarial Training for High-Stakes Reliability." arXiv.org, November 10, 2022. https://arxiv.org/abs/2205.01663.

Zwetsloot, Remco, and Allan Dafoe. "Thinking about Risks from Ai: Accidents, Misuse and Structure." Lawfare, October 31, 2019. https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure.

Anson Ho is a staff researcher at Epoch, where he does research on future AI impacts and developments, to support effective AI governance and policy. He is interested in helping ensure the safe and beneficial development of AI and other emerging technologies.

Acknowledgements
I would like to thank Lauren Wagner, Mary Chen, Gavin Douglas, Hadrien Pouget, and Lennart Heim for their feedback on this position paper.