Position: Governments Need to Increase and Interconnect Post-Deployment Monitoring of AI

Merlin Stein
University of Oxford
Oxford, UK
merlin.stein@bsg.ox.ac.uk

Jamie Bernardi
London, UK
contact@jamiebernardi.com

Connor Dunlop
Ada Lovelace Institute
London, UK
cdunlop@adalovelaceinstitute.org

Abstract

Language-based AI systems are diffusing into society, bringing positive and negative impacts. Mitigating negative impacts depends on accurate impact assessments, drawn from an empirical evidence base that makes causal connections between AI usage and impacts. Interconnected post-deployment monitoring combines information about model integration and use, applications, and real-world incidents and impacts. For example, chain-of-thought and inference data can be combined with monitoring social media for AI generated text, or monitoring societal indicators of disinformation. Drawing on information sharing mechanisms in other industries, we highlight example data sources and specific data points that governments and their AI Safety Institutes could collect to inform AI risk management.

1 Interconnected Post-Deployment Monitoring of AI as a Government Priority

People are increasingly exposed to AI systems in all areas of life. Language-based AI systems are general-purpose technologies [1], meaning they may be deployed across contexts. Systems like GPT-4, Claude, and Gemini are increasingly being integrated into workflows at Fortune 500 companies [2], public services [3], and in critical sectors like courts [4, 5] and health services [6].

Governments and the public have limited visibility into AI systems use and impacts. While many applications are beneficial, adopting language-based AI systems also carries societal risks [7, 8, 9]. Applicants may be discriminated against based on their names, as recruiters screen CVs with AI systems [10]; certain people's jobs may be displaced [11, 12], and citizens' data can be more readily stolen through AI-assisted cyber attacks [13]. Despite these risks, very little information about how AI is used and its impacts on society is available to governments or the general public [14], which could allow harms to propagate unaddressed.

Pre-deployment information is insufficient to thoroughly assess AI risks. To understand AI risks, governments and civil society have primarily developed mechanisms for gathering pre-deployment information, such as model evaluations [15]. However, pre-deployment information can not fully predict the downstream impacts of AI systems [16]. Risks ultimately arise from real-world usage, and depend on complex interactions of AI systems with people and society. For instance, combining systems with other tools can expand AI systems' capabilities in unpredictable ways [17].

Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS 2024.

Interconnected post-deployment monitoring can improve AI risk management through the use of empirical evidence. By monitoring AI's actual usage and impact, researchers can derive risk taxonomies [18, 19] and acceptable risk tolerances [20]. *Interconnected post-deployment monitoring* means 1) assessing risk by causally connecting different types of post-deployment information; for example, connecting real-world AI impacts with information about where a model is integrated and how it is used, and 2) connecting this causal risk assessment to specific mitigation strategies. For example, studies investigating AI-boosted misinformation and persuasion could benefit from increased visibility into AI usage in writing social media posts, to help inform causal connections about whether and how AI increases the prevalence of disinformation.

Post-deployment monitoring has been at least partly effective in other industries, and more effective when integrated into follow-up processes. The US Food and Drug Administration monitors population-level impacts of drugs linked to individual doctor observations [21]; this helps it apply new warning labels or, in the extreme case, remove a product from the market. Incident reporting in healthcare works best when connected with corrective action procedures [22, 23]. Accident monitoring and investigations by transport safety boards have sharply reduced fatalities across modes of transport, but only in high-income countries [24]. The EU's Digital Services Act 2022 monitors content moderation decisions and aims to link them to structural levels of misinformation [25, 26]. More detailed industry comparisons are required [27, 28, 29].

Current, public post-deployment monitoring of AI systems is driven by civil society, with limited capacity. Civil society organisations and researchers have revealed incidents, misuse and adverse impacts of AI systems [30, 31]. While civil society plays an important role in conducting post-deployment monitoring, restricted access to industrial information usually poses limits on its ability to audit industry [32]. AI companies partly screen usage data and customers [33, 34], but lack incentives to publicly share post-deployment information and monitoring tools [35, 36].

Given these restrictions and limitations on public access to post-deployment information, this position paper argues that **governments need to take an active role in conducting and incentivising post-deployment monitoring**. We contribute an overview of post-deployment monitoring and its impacts (Section 2), a description of its challenges (Section 3) and recommendations for governments, including identifying specific data points to request based on successes in other industries (Section 4).[1] Current practices in post-deployment monitoring are based on large fields of research in social science and computer science. We limit this overview to approaches directly related to general-purpose AI models, acknowledging its inherent high-level nature.

2 What is Post-Deployment Monitoring of AI Systems?

Post-deployment monitoring increases visibility into AI models' integration into applications, usage of AI applications, and AI applications' impacts on people and society. In figure 1, we categorise post-deployment information by the stage in the supply chain at which it is available.

2.1 Types of Post-Deployment Information

Model Integration and Usage Information relates to how AI models are integrated into digital applications. It includes information on how AI models are made available on the market, which application providers use them, and which industries most readily adopt AI models and downstream applications. An example is the US Census Bureau's survey of companies' AI use [12].

Governments and the public have little visibility into how different sectors deploy AI systems. This hinders the ability to monitor cross-cutting risks like over-reliance on AI in certain sectors or geographies, or market concentration and unequal access. Integration information can indicate when and where these cross-cutting risks might emerge.

Application Usage Information relates to how an application is used in context. It is generated when users interact with applications, ideally in the real world. It includes, for example, analysing AI system logs [40], monitoring feedback about AI applications (e.g. model vulnerabilities [41, 42, 43]), or conducting explicit sociotechnical field tests [44, 16]. Application usage data could also be collected

[1] These practices can be implemented in every jurisdiction that regulates AI systems. However, we draw on examples in the EU, the US and the UK throughout this article.

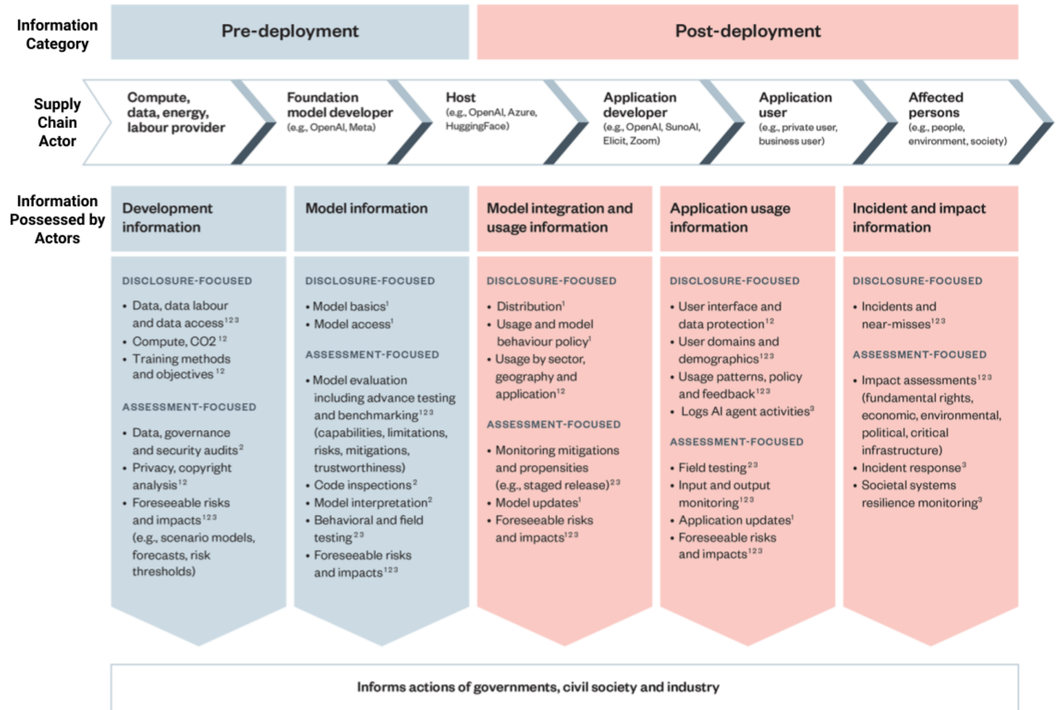

Figure 1: Information types for AI governance, categorised by supply chain stage. Some information sharing involves structured documentation (disclosure-focused), some requires additional analysis (assessment-focused). From Stein and Dunlop [37]. Information subcategories are superscript, and are drawn from 1) the Foundation Model Transparency Index [38], 2) the International Scientific report on the Safety of Advanced AI [39] and 3) the Sociotechnical Safety Evaluation Repository [16].

by monitoring online content for the appearance of AI outputs, which would be aided by requiring AI watermarks [45], content provenance [46] and AI agent activity logs [47, 48] (Section 4.4).

Usage information is especially useful in industries requiring high levels of reliability, safety and assured benefits. Understanding real-world AI usage is essential for assessing risk causally, thus informing effective mitigation strategies [16]. Usage monitoring could find, for example, that a few AI systems are used extensively in CV screening across many companies, which might correlate discrimination risks [10]. It may also show an over-reliance on AI systems for specific tasks, e.g., in critical infrastructure, which could then be reduced to prevent incidents [49].

Impact and Incident Information relates to tracking AI applications' societal effects, and adverse events and near-misses. It can be obtained through incident monitoring and reporting (see Section 4.1), survey of affected populations [50, 51], observing socioeconomic indicators such as income disparity or employment rates [1], or monitoring societal systems and infrastructure[2]. Entire research fields investigate AI impact information.

2.2 Deployment Configurations: Different Supply Chain Actors' Possession of Information

A single entity can fulfil one or many roles in the supply chain. For instance, OpenAI is the foundation model developer, a host *and* application provider for ChatGPT. Commercial relationships between entities affect information availability, due to customers' expectation of confidentiality with their vendor.

[2]Relevant societal functions to monitor could be prioritised based on ratings of their criticality [52, 53] and marginal influence of AI

Some deployment configurations, like open-source models and open centralised hosting, allow for more publicly available post-deployment information and feedback. If open-weight models are hosted privately, less information is available with centralised actors (e.g. Microsoft collects post-deployment information centrally of the open Mistral Large model on Microsoft's Azure servers [54]).

2.3 How Post-Deployment Monitoring of AI Could Inform Action

Empirical risk assessments could help inform institutional decision making. We briefly list non-exhaustive mechanisms that could incorporate post-deployment information. As more governments pursue AI regulation, risk levels could be measured through post-deployment monitoring, which could support government decisions to require corrective action from AI developers on deployed models [55]. Post-deployment information could also help set regulatory designations; designating by training compute [56] has been criticised for overlooking post-deployment capability improvements [57]. More generally, governments disseminating post-deployment information through, for example, international reports [39], could help decentralised actors identify and prioritise interventions that mitigate AI risk [58, 59]. Other, specific utilities of post-deployment information are listed in Table 1.

3 Challenges for Governments Monitoring AI Post-Deployment

Implementing policies for post-deployment monitoring of AI systems poses challenges, some seen in other industries, and others specific to AI technologies:

- **User privacy.** Users expect their AI system usage to be private, thus potentially input personal data in prompts. To monitor usage data directly, it's necessary to employ consent-based data donation [40] or privacy-preserving anonymisation and data analysis techniques [60].
- **Costs and independence.** Who pays the cost of compliance with post-deployment monitoring? Industry-funded monitoring, without appropriate incentive structures, can be low quality [61]. Independent third-parties require appropriate access and funding [32].
- **Information misuse.** Collecting information about incidents and misuse could strategically inform malign actors, requiring coordinated sharing mechanisms [55, 62].
- **Commercial sensitivity.** Information detailing the rate and distribution of AI integration may reveal opportunities for competitors. Whilst current market players keep this information private by default, limited public availability may promote wealth-creating competition [63]. Where governments have offered full confidentiality for post-market monitoring, conflicts of interest can emerge between commercial activity and public safety [64, 65].

4 Recommendations for Governments on Post-Deployment Monitoring

Post-deployment monitoring and follow-up do not happen by default. We outline four recommendations for governments and AI Safety Institutes developing post-deployment monitoring processes.

4.1 Prioritise Incident Monitoring and Reporting with Causal Connection to AI System Use

Incident reporting and monitoring are commonly practiced in many regulated industries [66, 67], and have proven at least partially effective in managing risks [22, 24]. These practices have inspired efforts to evaluate how incident reporting could support AI risk management [68, 21, 69, 70, 71]. Several AI incident databases have already emerged from civil society [30, 72, 73, 74], collecting their data from public channels. These have already informed analyses and taxonomies [75, 76, 77], and have proven to help their users quantify AI harms [78].

To be effective, AI incident reporting and monitoring processes should be designed with clear policy goals, typically one of learning or accountability [79]. These goals drive post-reporting actions, such as sharing learning with relevant stakeholders [62, 80] or implementing safety measures [55]. Governments are often well-suited to facilitate these processes: they have the authority to mandate reporting, act as neutral parties to encourage voluntary reporting, and can provide the resources and authority for follow-up actions.

Since it's difficult to evaluate the most effective incident reporting processes in advance, governments could adopt an iterative approach to their implementation. This would allow them to build expertise and gain insight into reporting gaps over time. As a low cost starting point, government functions that catalogue AI risks - like the UK's Central AI Risk Function [81] - could monitor publicly available data like news or incident databases to collect empirical evidence of AI harm, thus quantifying their risk assessments. From there, governments could explore more involved proposals, such as developing an ombudsman for citizens to report AI harms [82], mandating reporting for major AI incidents [68],[3] and collating AI-related incidents from sector-specific regulators [83]. By conducting further root cause analysis, policy teams could generate risk assessments that connect impact information with usage and integration information, thus informing effective mitigation strategies.

4.2 Establish Mechanisms to Gather Post-Deployment Information

In this recommendation we outline several non-exhaustive strategies that governments and AI Safety Institutes can employ to gather post-deployment information on AI systems and models. Their respective utilities depend on the regulatory and industry context, and the nature of the monitored AI system.

Voluntary Information Provision and Cooperation. Governments can gather information from AI companies through both informal and formal channels for voluntary cooperation. This can involve requests for specific statistics (examples given in Table 1), but could also involve companies providing regular aggregated data streams: the UK's Office for National Statistics receives aggregated data from payment service providers [84], which could be a useful model for governments monitoring AI integration and usage statistics. The UK and US AI Safety Institutes have already established voluntary agreements with leading AI model developers to test their models before deployment [85, 86], and this framework could be expanded to include post-deployment data. Voluntary cooperation strategies are lighter-touch and more flexible than making mandatory requests, but their success is dependent on goodwill relationships, which may incur a selection bias in which companies provide the most information to government [87].

Mandatory reporting through legislation. Mandatory reporting requirements ensures broad compliance, which may be essential for obtaining safety critical information. Mandatory requests often require legislative backing. A useful framework to consider for AI-related information requests is the UK's Digital Economy Act 2017, which empowers its Office for National Statistics to mandate businesses to submit specific data through binding surveys [88]. The EU AI Act already mandates certain post-market reporting, including metric reporting (Article 72) and documentation of serious incidents (Article 73) [89]. An effective approach depends on governments having enough knowledge to request targeted information [87].

Third-Party Research and Independent Monitoring. Academics and other third party institutions play an important role in collecting and analysing post-deployment data, however their data access is often limited to public sources [32]. Third parties have utilised alternative sources like SimilarWeb [90] and building independent datasets for AI usage [40]. Governments can support third party efforts through funding [91], providing researcher access to non-public data [92], and otherwise protecting and supporting third party investigations [32].

4.3 Request Initial Data Points and Build Analysis Capacity

By examining post-market information that has been useful in existing regulated industries, we provide in Table 1 a preliminary, non-comprehensive list of data points that governments could start requesting from companies in the AI supply chain. It includes: integration and model usage information like usage by location and sector; application usage information like intended use case and degree of tool use; and impact information like incident monitoring.

A full effort to understand AI risks would use these data points in combination with other data sources, such as macroeconomic indicators and surveying affected populations. Together, causal connections could be inferred between observed societal impacts and the data outlined in this table. For example, the environmental impacts of AI could be inferred from inference volumes [93]. Economic disparities across genders can be predicted using differing usage amounts [94].

[3] Definitions of *major* incidents are underway [68, 83]

Gathering and learning from information as a government is likely to be an iterative process of identifying an informative data point, requesting it from industry, analysing the provided data, then evaluating its usefulness to generate new lines of inquiry. Requesting and analysing information requires staff time, which governments could hire-in directly [28], fund [91], or facilitate by incentivising a third-party ecosystem [32, 16]. Despite access limitations, third party organisations should not be overlooked; in the past, they have advocated for monitoring functions and the enforcement of the Digital Services Act through analysing public data [95].

4.4 Support Technical Governance Methods that Increase Visibility

As AI outputs become more prevalent, governments should continue to encourage adoption of visibility-building technologies like content provenance [46] and watermarking [45]. As language-based AI agents are developed and become more prevalent, governments should proactively support corresponding visibility standards [48]. This includes AI agents outputting *identifiers*, informing companies and individuals about when they are interacting with agents, indicating *where possible* which developer is accountable, and otherwise creating visibility that third-party researchers could analyse.

Visibility into AI agent behaviour may also involve analysing logs [47]. Researchers have preserved privacy by conducting test tasks, however technical solutions may enable monitoring of real agents [60]. In any case, government agencies should work with agent developers to understand agent behaviour and human-agent interaction early in this technology's development to identify risks, inform technical processes that mitigate them, and surface ways that companies and individuals should adapt to the diffusion of AI agents [59].

5 Conclusion and Future Work

In this paper, we have argued for the critical importance of interconnected post-deployment monitoring of AI systems by governments and their AI Safety Institutes. We suggest causally connecting three kinds of post-deployment information: model integration and usage, application usage, and impact and incident data. We recommend that governments begin building this information ecosystem by:

- Prioritising incident monitoring and reporting, with causal links to AI system use.
- Implementing mechanisms to gather post-deployment information.
- Requesting specific data points from AI companies and build analysis capacity.
- Supporting technical governance methods that increase visibility of AI systems.

We call on the technical and AI governance research communities and AI companies to support these measures, which requires future work on assessing the effectiveness of different post-deployment monitoring approaches and using privacy-preserving techniques to build more post-deployment datasets like WildChat [40] across different sectors and applications.

Table 1: Examples data points for post-deployment monitoring.

Data Point	Utility	Downsides	Analogies
Integration and model usage information (usually provided by model hosts)			
Size of user-base, including total inference volume.	Allocate research by measuring prevalence and growth in AI applications.	Data is coarse, and survey may suffice.	EU Digital Service Act regulation only covers platforms with > 45M active EU users [26].
Usage by sector, e.g. inference volumes by Standard Industrial Classification code.	Identify potential structural risks like over-reliance and market concentration in critical sectors.	Revealing market gaps across industries may be commercial sensitive.	The US Census Bureau collects usage information by survey to understand AI's impact on employment [12].
Usage by location, e.g. inference volume per region.	Monitor adoption effects, e.g. comparing economic outcomes with regional AI use.	Revealing market gaps across geographies may be commercially sensitive.	Regional differences are commonly measured to inform digital inclusion strategies [96].
Model host downtime, e.g. minutes/month of unavailability.	Minimise economic and other harms from downtime as AI reliance grows.	Competitive markets already incentivise minimal downtime (see 'service level agreements').	The UK's Financial Conduct Authority monitors payment service providers' up-time (e.g. Visa [97]).
Application usage information (usually provided by application developers)			
Intended use case of an AI request, e.g. CV screening, therapy, medical.	Prioritise regulatory response based on prevalence of use cases.	Revealing market gaps in use-cases may be commercially sensitive.	The US Food and Drug Administration monitors drug usage as part of broader evaluations [98].
Degree of tool use in AI applications (e.g., web browser access).	Assess AI's potential to operate autonomously.	Revealing specific tool usage may be commercially sensitive.	AI-specific data point, discussed in [48].
Anonymised chat logs, with user consent [40].	Support research on AI impacts like sycophancy, over-reliance and safeguard failures.	Important privacy concerns. User awareness of sharing causes sampling bias.	The UK's Office for National Statistics receives anonymised payment data from providers [84].
Incident and impact information (usually better informed by observation)			
Misuse statistics, e.g. declined requests and account closures.	Measure scale of misuse and safeguard efficacy.	Reporting misuse may incentivise under-detection, and could inform attackers.	EU Digital Service Act requires transparency on moderation decisions and incidents [99].
Incident monitoring and reporting to identify or quantify harm.	Prevent repeated AI failures by informing legislation or safeguards [30]. Respond to crises.	Compliance costs, and difficulty scoping an AI incident.	Incident reporting has precedent in multiple industries [68, 70].

Societal Impacts Statement

This paper aims to increase visibility of AI's societal impacts. However, increasing visibility naturally raises privacy concerns. In this paper, the most privacy-sensitive policy we discuss is the analysis of users' chat logs to help understand AI usage (other metrics we discuss are usually aggregated and/or carry no personal information, only high-level statistics about usage). Analysis of usage is already conducted by foundation model providers for misuse monitoring, and is usually highly automated (meaning few humans require access to chat logs). Monitoring usage information should be carried out using best practice developed for those purposes, with a minimal set of employees able to view personal data. When considering data sharing agreements, governments and other actors should: follow data protection laws in the relevant jurisdiction(s) at a minimum; ensure data sharing agreements are clear and transparent to users; and take every effort to conceal or remove personal data using privacy-preserving technologies.

Acknowledgments and Disclosure of Funding

Authors' salaries whilst conducting this work were provided by Connor Dunlop's employment at the Ada Lovelace Institute, Jamie Bernardi's fellowship position at the Institute for AI Policy and Strategy, and Merlin Stein's PhD position at the University of Oxford.

For comments on this work, we'd like to thank Rishi Bommasani, Simon Mylius, Tommy Shaffer Shane, Kevin Wei, Zoe Williams.

References

[1] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. Gpts are gpts: An early look at the labor market impact potential of large language models, 2023.

[2] Sarah Wang and Shangda Xu. 16 changes to the way enterprises are building and buying generative ai. Technical report, Andreessen Horowitz, 2024.

[3] Innovation Department for Science and Technology. Artificial intelligence (ai) opportunities action plan: terms of reference. Technical report, UK Government, 2024. Accessed 2024-09-06.

[4] Irene Pietropaoli. Use of artificial intelligence in legal practice, 2023. Accessed: 2024-09-06.

[5] Michael Cross. News focus: Artificial intelligence debuts at the old bailey, 2023. Accessed: 2024-09-06.

[6] Peter Lee, Sebastien Bubeck, and Joseph Petro. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239, 2023.

[7] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from language models, 2021.

[8] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. Harms from increasingly agentic algorithmic systems. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23. ACM, June 2023.

[9] Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, Seliem El-Sayed, Sasha Brown, Canfer Akbulut, Andrew Trask, Edward Hughes, A. Stevie Bergman, Renee Shelby, Nahema Marchal, Conor Griffin, Juan Mateos-Garcia, Laura Weidinger, Winnie Street, Benjamin Lange, Alex Ingerman, Alison Lentz, Reed Enger, Andrew Barakat, Victoria Krakovna, John Oliver Siy, Zeb Kurth-Nelson, Amanda McCroskery, Vijay Bolina, Harry Law, Murray Shanahan, Lize Alberts, Borja Balle, Sarah de Haas, Yetunde Ibitoye, Allan Dafoe, Beth Goldberg, Sébastien Krier, Alexander Reese, Sims Witherspoon, Will Hawkins, Maribeth Rauh, Don Wallace, Matija Franklin, Josh A. Goldstein, Joel Lehman, Michael Klenk, Shannon Vallor, Courtney Biles, Meredith Ringel Morris, Helen King, Blaise Agüera y Arcas, William Isaac, and James Manyika. The ethics of advanced ai assistants, 2024.

[10] Leon Yin, Davey Alba, and Leonardo Nicoletti. Openai's gpt is a recruiter's dream tool. tests show there's racial bias, March 2024.

[11] Kweilin Ellingrud, Saurabh Sanghvi, Gurneet Singh Dandona, Anu Madgavkar, Michael Chui, Olivia White, and Paige Hasebe. Generative ai and the future of work in america. Technical report, McKinsey & Company, 2023.

[12] Kathryn Bonney, Cory Breaux, Catherine Buffington, Emin Dinlersoz, Lucia Foster, Nathan Goldschlag, John Haltiwanger, Zachary Kroff, and Keith Savage. Tracking firm use of ai in real time: A snapshot from the business trends and outlook survey. Technical report, US Census Bureau, 2024.

[13] Richard Fang, Rohan Bindu, Akul Gupta, Qiusi Zhan, and Daniel Kang. Teams of llm agents can exploit zero-day vulnerabilities, 2024.

[14] Gabriel Nicholas. Towards researcher access to ai usage data, 2024.

[15] Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, Lewis Ho, Divya Siddarth, Shahar Avin, Will Hawkins, Been Kim, Iason Gabriel, Vijay Bolina, Jack Clark, Yoshua Bengio, Paul Christiano, and Allan Dafoe. Model evaluation for extreme risks, 2023.

[16] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. Sociotechnical safety evaluation of generative ai systems, 2023.

[17] Tom Davidson, Jean-Stanislas Denain, Pablo Villalobos, and Guillem Bas. Ai capabilities can be significantly improved without expensive retraining, 2023.

[18] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA, 2022. Association for Computing Machinery.

[19] Leonie Koessler and Jonas Schuett. Risk assessment at agi companies: A review of popular risk assessment techniques from other safety-critical industries, 2023.

[20] Leonie Koessler, Jonas Schuett, and Markus Anderljung. Risk thresholds for frontier ai, 2024.

[21] Merlin Stein and Connor Dunlop. Safe before sale: Learnings from the fda's model of life sciences oversight for foundation models. Technical report, Ada Lovelace Institute, December 2023.

[22] Charitini Stavropoulou, Catherine Doherty, and Paul Tosey. How effective are incident-reporting systems for improving patient safety? a systematic literature review. *Milbank Quarterly*, 93(4):826–866, Dec 2015.

[23] Ken Goekcimen, René Schwendimann, Yvonne Pfeiffer, Giulia Mohr, Christoph Jaeger, and Simon Mueller. Addressing patient safety hazards using critical incident reporting in hospitals: A systematic review. *Journal of Patient Safety*, 19(1):e1–e8, Jan 2023.

[24] Eric Fielding, Andrew W. Lo, and Jian Helen Yang. The national transportation safety board: A model for systemic risk management. *Journal of Investment Management*, 9(1):17–49, First Quarter 2011.

[25] Iva Nenadić, Elda Brogi, and Konrad Bleyer-Simon. Structural indicators to assess effectiveness of the eu's code of practice on disinformation. RSC Working Paper 2023/34, European University Institute, Robert Schuman Centre for Advanced Studies, 2023.

[26] European Parliament and of the Council. Regulation (eu) 2022/2065 of the european parliament and of the council of 19 october 2022 on a single market for digital services (digital services act), 2022. Accessed: 2024-09-09.

[27] Matt Davies Julia Smakman. New rules? `https://www.adalovelaceinstitute.org/report/new-rules-ai-regulation/`, 2024. [Accessed 16-11-2024].

[28] Merlin Stein, Milan Gandhi, Theresa Kriecherbauer, Amin Oueslati, and Robert Trager. Public vs private bodies: Who should run advanced ai evaluations and audits? a three-step logic based on case studies of high-risk industries. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1401–1415, 2024.

[29] Ashwin Acharya-Researcher. Ai-relevant regulatory precedents. *IAPS*, 2024.

[30] Sean McGregor. Preventing repeated real world ai failures by cataloging incidents: The ai incident database. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):15458–15463, May 2021.

[31] Lara Groves, Jacob Metcalf, Alayna Kennedy, Briana Vecchione, and Andrew Strait. Auditing work: Exploring the new york city algorithmic bias audit regime. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1107–1120, 2024.

[32] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. Outsider oversight: Designing a third party audit ecosystem for ai governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 557–571, 2022.

[33] OpenAI. Disrupting deceptive uses of ai by covert influence operations, May 2024. Accessed: 2024-09-09.

[34] OpenAI. Disrupting malicious uses of ai by state-affiliated threat actors, February 2024. Accessed: 2024-09-09.

[35] Jacob Wulff Wold. Openai sitting on tool to watermark ai-generated content, 2024. Accessed: 2024-09-09.

[36] Cade Metz. Openai hack, 2024. Accessed: 2024-09-09.

[37] Merlin Stein and Connor Dunlop. Safe beyond sale: post-deployment monitoring of ai. Technical report, Ada Lovelace Institute, 2024.

[38] Rishi Bommasani, Kevin Klyman, Sayash Kapoor, Shayne Longpre, Betty Xiong, Nestor Maslej, and Percy Liang. The foundation model transparency index v1. 1: May 2024. *arXiv preprint arXiv:2407.12929*, 2024.

[39] Multiple. International scientific report on the safety of advanced ai. Technical report, AI Seoul Summit, 2024.

[40] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024.

[41] Innovation Department for Science and Technology. Emerging processes for frontier ai safety, 2023. Accessed: 2024-09-11.

[42] The White House. Voluntary ai commitments, 2023.

[43] OpenAI. Model behavior feedback. Accessed: 2024-09-09.

[44] Reva Schwartza, Jonathan Fiscusa, Kristen Greenea, Gabriella Watersb, Rumman Chowdhuryb, Theodore Jensena, Craig Greenberga, Afzal Godila, Razvan Amironeseia, Patrick Hallb, and Shomik Jaina. Nist assessing risks and impacts of ai (aria) pilot evaluation plan. Pilot evaluation plan, National Institute of Standards and Technology, Information Technology Laboratory and Associate, August 2024. Contact: aria_inquiries@nist.gov.

[45] Siddarth Srinivasan. Detecting ai fingerprints: A guide to watermarking and beyond. Technical report, Brookings, 2024.

[46] Coalition for Content Provenance and Authenticity. C2pa technical specification. Technical report, Coalition for Content Provenance and Authenticity, 2021.

[47] Silen Naihin, David Atkinson, Marc Green, Merwane Hamadi, Craig Swift, Douglas Schonholtz, Adam Tauman Kalai, and David Bau. Testing language model agents safely in the wild. In *Socially Responsible Language Modelling Research*, 2023.

[48] Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, Lennart Heim, and Markus Anderljung. Visibility into ai agents. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 958–973, New York, NY, USA, 2024. Association for Computing Machinery.

[49] Rosco Hunter, Richard Moulange, Jamie Bernardi, and Merlin Stein. Monitoring human dependence on ai systems with reliance drills, 2024.

[50] Rijul Chaturvedi, Sanjeev Verma, Ronnie Das, and Yogesh K Dwivedi. Social companionship with artificial intelligence: Recent trends and future avenues. *Technological Forecasting and Social Change*, 193:122634, 2023.

[51] Julian De Freitas, Ahmet K Uguralp, Zeliha O Uguralp, and Puntoni Stefano. Ai companions reduce loneliness. *arXiv preprint arXiv:2407.19096*, 2024.

[52] Steffen Roth and Anton Schutz. Ten systems: Toward a canon of function systems. *Cybernetics & Human Knowing*, 22(4):11–31, 2015.

[53] John Bordeaux, Aaron Strong, Jhacova Williams, Benjamin Boudreaux, Michael JD Vermeer, Shannon Prier, Jonathan W Welburn, Zev Winkelman, Cheryl K Montemayor, Jessica Welburn Paige, et al. Identifying and prioritizing systemically important entities: Advancing critical infrastructure security and resilience, 2023.

[54] Eric Boyd. Microsoft and mistral ai announce new partnership to accelerate ai innovation and introduce mistral large first on azure, February 2024. Accessed: 2024-09-09.

[55] Joe O'Brien, Shaun Ee, and Zoe Williams. Deployment corrections: An incident response framework for frontier ai models, 2023.

[56] Girish Sastry, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O'Keefe, Gillian K. Hadfield, Richard Ngo, Konstantin Pilz, George Gor, Emma Bluemke, Sarah Shoker, Janet Egan, Robert F. Trager, Shahar Avin, Adrian Weller, Yoshua Bengio, and Diane Coyle. Computing power and the governance of artificial intelligence, 2024.

[57] Sara Hooker. On the limitations of compute thresholds as a governance strategy, 2024.

[58] Heidy Khlaaf. Toward comprehensive risk assessments and assurance of ai-based systems. Technical report, Trail of Bits, 2023.

[59] Jamie Bernardi, Gabriel Mukobi, Hilary Greaves, Lennart Heim, and Markus Anderljung. Societal adaptation to advanced ai, 2024.

[60] Emma Bluemke, Tantum Collins, Ben Garfinkel, and Andrew Trask. Exploring the relevance of data privacy-enhancing technologies for ai governance use cases. *ArXiv*, abs/2303.08956, 2023.

[61] Angela Spelsberg, Christof Prugger, Peter Doshi, Kerstin Ostrowski, Thomas Witte, Dieter Hüsgen, and Ulrich Keil. Contribution of industry funded post-marketing studies to drug safety: survey of notifications submitted to regulatory agencies. *BMJ*, 356, 2017.

[62] Noam Kolt, Markus Anderljung, Joslyn Barnhart, Asher Brass, Kevin Esvelt, Gillian K. Hadfield, Lennart Heim, Mikel Rodriguez, Jonas B. Sandbrink, and Thomas Woodside. Responsible reporting for frontier ai development, 2024.

[63] David J. Teece, Gary Pisano, and Amy Shuen. Dynamic capabilities and strategic management. *Strategic Management Journal*, 18(7):509–533, 1997.

[64] Joel Lexchin and Barbara Mintzes. Transparency in drug regulation: Mirage or oasis? *CMAJ*, 171(11):1363–1365, 2004.

[65] Emilia Korkea-aho and Päivi Leino. Who owns the information held by eu agencies? weed killers, commercially sensitive information and transparent and participatory governance. *Common Market Law Review*, 54(4):1059–1091, 2017.

[66] Food and Drug Administration. Medwatch: The fda safety information and adverse event reporting program. Accessed: 2024-09-09.

[67] Federal Aviation Authority. Accident & incident data. Accessed: 2024-09-09.

[68] Ren Bin Lee Dixon and Heather Frase. An argument for hybrid ai incident reporting, 2024.

[69] Matt Davies and Michael Birtwistle. Regulating ai in the uk. Technical report, Ada Lovelace Institute, 2023.

[70] Tommy Shaffer Shane. Ai incident reporting: Addressing a gap in the uk's regulation of ai. Technical report, Centre for Long Term Resilience, 2024.

[71] John Croxton, David Robusto, Satya Thallam, and Doug Calidas. Establishing an ai incident reporting system, 2024. Accessed: 2024-09-09.

[72] Policy Observatory. Ai incidents monitor (aim).

[73] Charlie Pownall. Ai, algorithmic, and automation incidents and controversies.

[74] Christina P. Walker, Daniel S. Schiff, and Kaylyn Jackson Schiff. Merging ai incidents research with political misinformation research: Introducing the political deepfakes incidents database. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21):23053–23058, 2024.

[75] Nahema Marchal, Rachel Xu, Rasmi Elasmar, Iason Gabriel, Beth Goldberg, and William Isaac. Generative ai misuse: A taxonomy of tactics and insights from real-world data, 2024.

[76] Edyta Bogucka, Marios Constantinides, Julia De Miguel Velazquez, Sanja Šćepanović, Daniele Quercia, and Andrés Gvirtz. The atlas of ai incidents in mobile computing: Visualizing the risks and benefits of ai gone mobile, 2024.

[77] Mengyi Wei and Kyrie Zhou. Ai ethics issues in real world: Evidence from ai incident database. In *56th Hawaii International Conference on System Sciences*, 08 2022.

[78] Michael Feffer, Nikolas Martelaro, and Hoda Heidari. The ai incident database as an educational tool to raise awareness of ai harms: A classroom exploration of efficacy, limitations, & future improvements. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, New York, NY, USA, 2023. Association for Computing Machinery.

[79] Kevin Wei and Lennart Heim. Designing incident reporting systems for harms from ai (forthcoming), 2024.

[80] MITRE. Common weakness enumeration database. Accessed: 2024-09-09.

[81] UK Government. A pro-innovation approach to ai regulation.

[82] Elliot Jones. Keeping an eye on ai. Technical report, Ada Lovelace Institute, 2024.

[83] Tommy Shaffer Shane. Ai incident reporting: Addressing a gap in the uk's regulation of ai. Technical report, Centre for Long-Term Resilience, 2024.

[84] UK Government Office for National Statistics. Economic activity and social change in the uk, real-time indicators, 2023.

[85] Vincent Manancourt and Tom Bristow. British pm rishi sunak secures 'landmark' deal on ai testing, 2024. Accessed: 2024-09-09.

[86] Ina Fried. Anthropic and openai models proposed for us ai safety institute, 2024. Accessed: 2024-09-09.

[87] Cary Coglianese, Richard Zeckhauser, and Edward A. Parson. Seeking truth for power: Informational strategy and regulatory policy making. *Minnesota Law Review*, 89:277, 2004.

[88] Office for National Statistics. Relevant legislation. Technical report, Office for National Statistics, 2021.

[89] European Parliament. Chapter ix, artificial intelligence act (regulation (eu)). 2024/1689, Official Journal version of 13 June 2024, 2024. Accessed: 2024-09-09.

[90] SimilarWeb. Chatgpt. Technical report, SimilarWeb, 2024.

[91] UK AI Safety Institute, 2024. Accessed: 2024-09-09.

[92] European Commission. Status report on mechanisms for researcher access to online platform data. Technical report, European Commission, 2024. Accessed: 2024-09-12.

[93] Julien Nioche. The environmental impact of the cloud – the common crawl case study. Technical report, Common Crawl, 2024. Accessed: 2024-09-06.

[94] Jessica Howington. The ai gender gap: Exploring variances in workplace adoption, 2024.

[95] Suzanne Vergnolle. Putting collective intelligence to the enforcement of the digital services act: Report on possible collaborations between the european commission and civil society organisations. *Cnam Paper Series*, 2023.

[96] C. K. Sanders and E. Scanlon. The digital divide is a human rights issue: Advancing social inclusion through social work advocacy. *Journal of Human Rights and Social Work*, 6:130–143, 2021.

[97] Financial Conduct Authority. Reporting requirements: payment service providers and e-money issuers, 2024. Accessed: 2024-09-09.

[98] U.S. Food and Drug Administration. Fys 2013 - 2017 regulatory science report: Analysis of generic drug utilization and substitution, 2024. Accessed: 2024-09-09.

[99] L. Nannini, E. Bonel, D. Bassi, et al. Beyond phase-in: assessing impacts on disinformation of the eu digital services act. *AI Ethics*, 2024.