Visibility into AI Agents

Alan Chan*
alan.chan@mila.quebec
Centre for the Governance of AI
Oxford, UK
Mila (Quebec AI Institute)
Montréal, Canada

Carson Ezell
Harvard University
Cambridge, USA

Max Kaufmann
Independent
London, UK

Kevin Wei
Harvard Law School
Cambridge, USA

Lewis Hammond
University of Oxford
Oxford, UK
Cooperative AI Foundation
Oxford, UK

Herbie Bradley
University of Cambridge
Cambridge, UK

Emma Bluemke
Centre for the Governance of AI
Oxford, UK

Nitarshan Rajkumar
University of Cambridge
Cambridge, UK

David Krueger
University of Cambridge
Cambridge, UK

Noam Kolt
University of Toronto
Toronto, Canada

Lennart Heim†
Centre for the Governance of AI
Oxford, UK

Markus Anderljung†
Centre for the Governance of AI
Oxford, UK

ABSTRACT

Increased delegation of commercial, scientific, governmental, and personal activities to AI agents—systems capable of pursuing complex goals with limited supervision—may exacerbate existing societal risks and introduce new risks. Understanding and mitigating these risks involves critically evaluating existing governance structures, revising and adapting these structures where needed, and ensuring accountability of key stakeholders. Information about where, why, how, and by whom certain AI agents are used, which we refer to as **visibility**, is critical to these objectives. In this paper, we assess three categories of measures to increase visibility into AI agents: **agent identifiers**, **real-time monitoring**, and **activity logging**. For each, we outline potential implementations that vary in intrusiveness and informativeness. We analyze how the measures apply across a spectrum of centralized through decentralized deployment contexts, accounting for various actors in the supply chain including hardware and software service providers. Finally, we discuss the implications of our measures for privacy and concentration of power. Further work into understanding the measures and mitigating their negative impacts can help to build a foundation for the governance of AI agents.

*Correspondence to `alan.chan@mila.quebec`
†Equal co-supervision.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658948

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Applied computing** → *Law*; • **Social and professional topics** → *Governmental regulations*.

KEYWORDS

visibility, transparency, ai agents, ai deployment, ai oversight, ai monitoring

ACM Reference Format:
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. 2024. Visibility into AI Agents. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3630106.3658948

1 INTRODUCTION

Many AI developers are creating systems with greater autonomy, access to external tools or services, and an increased ability to reliably adapt, plan, and act open-endedly over long time-horizons to achieve goals [35, 97, 106, 133, 137, 146, 158]. We will say that such systems possess relatively high degrees of **agency** and will refer to them as **(AI) agents** or **agentic systems** [35, 104, 147]. Systems with relatively low degrees of agency are those that only aid human decision-making or produce outputs without acting in the world, such as image classifiers or text-to-image models. Examples of agents could include reinforcement learning systems [132, 160] that interact extensively with the real world[1] or more capable versions of language models with tool or service access

[1]Including the physical environment but also digital environments such as online platforms.

that could, for example, plan and book a holiday or send an email on a user's behalf [27, 122, 137, 144].

Current AI agents sometimes struggle to perform even simple tasks [98, 106, 108, 130, 163, 164], but given increasing investments in AI research [61], scaling laws [15, 83, 94], pressures to develop autonomous capabilities for military use [86, 102, 142], economic applications [35], and scientific prestige [35, 67], we should not discount continued improvements in capabilities [26]. Indeed, a core goal of the AI field since its inception has been to build agents [138, 159].

As AI agents improve in capabilities, speed, and cost,[2] it may be easier and more competitive to delegate tasks currently done by humans to AI agents instead. The development and deployment of agents has surged recently [36, 124, 133, 150, 176] and could lead to the ubiquitous deployment of agents in commercial, scientific, governmental, and personal activities. Since such deployment may exacerbate existing risks and introduce new ones [35, 147], it is imperative to understand how to govern AI agents.

1.1 Risks of AI Agents

Rather than provide an exhaustive taxonomy of risks from AI agents,[3] we highlight certain agent-specific risks. In comparison to risks from other AI systems, these risks focus on the potential for agents to remove humans from the loop [35, 101]. Without a human in the loop, agents may take multiple consequential actions in rapid succession and bring about significant impacts before a human notices. The ability to remove humans from the loop also means that an agent's task performance is less limited by the expertise of its user, compared to a situation where user must guide an AI system's actions or take actions themself.

1.1.1 Malicious Use. AI agents could be a large impact multiplier for individuals or coordinated groups who wish to cause harm [147]. Existing AI systems have already assisted in malicious use, including voice cloning scams [168] and fake news generation [167]. However, more capable AI agents could automate end-to-end pipelines for complex tasks that currently require substantial human expertise and time. For untrained individuals, such agents could drastically increase the accessibility of engaging in severely harmful activities because no human in the loop would be required. For example, there is interest in building agents to execute scientific research, comprising autonomous planning and execution of scientific experiments [22, 27]. If such agents were to become as capable as human scientists, they might enable or accelerate the design and development of harmful tools (e.g., biological [140, 155], chemical [22, 27, 162]) for groups that currently lack the expertise for such production. Extremely persuasive AI agents may also enable and enhance influence campaigns [10, 78, 96].

Understanding the extent to which agents will facilitate malicious use requires information about how they are used and how they interact with external systems [173]. Moreover, when malicious users do cause harm with AI agents, regulatory enforcers will need measures to identify the users and hold them accountable.

[2]As an example of cost reduction, FLOP [82] or FLOP/s [81] per dollar could decrease at the same time as performance per FLOP increases [62].

[3]See Critch and Russell [45], Shelby et al. [149], Weidinger et al. [173] for taxonomies of risks from AI systems and [35, 147] for further discussion of risks from AI agents.

1.1.2 Overreliance and Disempowerment. Overreliance on AI agents to automate complex, high-stakes tasks could lead to severe consequences. Humans can already rely on certain automated systems more than is warranted [47, 59, 60]. More capable agents may enable automation of an increasing array of complex and useful tasks. Users—including both individuals and institutions—may rely on agents even in high-stakes situations, such as interfacing with the financial or legal systems, because human alternatives (e.g., hiring a lawyer) may become relatively slower and more expensive. At the same time, these agents may malfunction for a variety of reasons, including design flaws [117, 130, 181] or adversarial attack [12, 175, 180]. Malfunction may not be immediately apparent, especially if users lack the requisite expertise or domain knowledge. Stopping the agent may be difficult if doing so would lead to cascading failures or a competitive disadvantage for the user [147]. More broadly, profit and efficiency motives may lead to collective dependence on agents for essential societal functions, such as the provision of government services [50, 179] or the operation of essential infrastructure [17, 51]. Companies providing access to AI agents would hold substantial power [28], while malfunction of those agents could have societal-scale impacts. At minimum, societies require information about the extent of reliance upon AI agents and whether such reliance is justified.

1.1.3 Delayed and Diffuse Impacts. Potential negative impacts of AI agents may be delayed and diffuse.[4] Delayed and diffuse impacts may be difficult to manage because they may require sustained attention over long periods of time even to notice. Impacts of agents may be delayed if users give agents long-horizon goals, while diffuseness of impact may come from the widespread deployment of agents to automate complex processes. Consider an agent given the goal of continually finding and hiring job candidates who will most contribute to the company over the long-term. This agent may screen résumés [66], perform interviews, make the final hiring decision, and analyze the performance of hires. Given the time horizon over which the agent is acting and its influence over the company, any potential problems like algorithmic bias [129] could be hard to identify and become deeply entrenched. The most severe consequences of such problems may only be apparent when looking at how companies in aggregate use AI agents for hiring. AI agents could also subtly benefit their developers, akin to the self-preferencing behaviour of large-scale digital platforms [99]. Moreover, agents that mediate or even substitute for human communication [5, 107] could have diffuse and delayed psychological and social impacts [10, 90, 96], analogous to certain effects of social media [25, 110, 148]. The deployment of agents may also induce changes in market structures or workforce impacts from job displacement [6, 7]. Identifying delayed and diffuse impacts may require long-term tracking of the extent and nature of AI agent usage across a wide range of application areas.

1.1.4 Multi-Agent Risks. Interactions and dependencies between many deployed agents could lead to risks not present at the level of

[4]Roughly speaking, we consider an impact to be diffuse if it is difficult to observe and most apparent in aggregate across many individual cases.

a single system [73, 77, 128, 143]. Agents could enter into destabilising feedback loops, such as those between automated trading algorithms in the 2010 flash crash [46]. Agents partially built upon the same components—such as a particular foundation model—could have common vulnerabilities and failure modes [23, 40]; widespread deployment of such agents could risk large-scale systemic harms. More generally, there may be unpredictable behavioural changes that are characteristic of complex systems [40, 143, 153]. Competitive pressures and selection effects could lead to the development of agents that act in more anti-social ways [31, 57, 79, 177]. These potential issues motivate understanding not just individual agents, but also interactions within groups of agents.

1.1.5 Sub-Agents. Agents could instantiate more agents to accomplish (components of) a task, which may magnify several of the risks discussed so far. It may be advantageous for an agent to create potentially specialized and more efficient **sub-agents**, especially if doing so is cheap and fast. For example, an agent could call copies of itself through an API, or itself train, fine-tune, or otherwise program another agent. Sub-agents could be problematic because they introduce additional points of failure; each sub-agent may itself malfunction, be vulnerable to attack, or otherwise operate in a way contrary to the user's intentions. Stopping an agent from causing further harm might involve intervening not only on the agent, but also on any relevant sub-agents [30, 154]. Yet, this process may be difficult because we lack methods for determining when an agent has created a sub-agent. Information about the extent of sub-agent creation and operation can enable a better understanding of the significance of these risks.

1.2 The Case for Visibility into AI Agents

Addressing the risks of AI agents requires **visibility**:[5] information about where, why, how, and by whom AI agents are used. Visibility would help to evaluate existing governance structures, revise and adapt these structures where needed, and ensure accountability of key stakeholders. Regulatory oversight bodies which monitor and enforce rules on the activities of human agents and certain automated programs (e.g., trading algorithms) [14, 18, 19, 21, 32, 39, 76, 84, 85, 93, 103, 111, 165] may require additional information to understand and address harms from AI agents. For example, if agents are able to employ novel strategies for collusion [55] when carrying out economic activities, new rules and updates to investigative authority may be necessary. Furthermore, AI agents may simultaneously provide services traditionally regulated by different agencies, such as both financial and legal services. The same agent developer or deployer may thus exercise power across diverse and usually independent domains of regulation, creating additional concerns related to market consolidation and conflicts of interest.

Visibility measures also play a central role in addressing problems that arise when humans delegate to other humans or institutions [103]. The precise purpose of such regimes varies, and can include reducing information asymmetries, shaping incentive structures, and triggering enforcement actions [84, 85]. For example, employers often monitor the conduct and performance of employees through ongoing supervision and periodic performance reviews [14]. In corporate governance, shareholders monitor management through a range of institutional mechanisms, including financial audits, shareholder meetings, and company reports, buttressed by legally binding fiduciary duties (in the case of directors) and the ability (in some cases) to dismiss management if they fail to act in the collective interest of shareholders [93]. Comparable mechanisms exist to support citizens in monitoring the activities of government bodies and public officials. These mechanisms include maintaining records of government decisions, facilitating information access through freedom of information requests, and commissioning detailed public reports into government activities [19], a combination of which may ultimately inform citizens' electoral choices. Although visibility measures can be costly [63] and raise privacy concerns, they remain a necessary feature of frameworks for shaping the incentives of, and governing, agents.

We emphasize visibility into *deployed* AI agents because the scope and severity of potential impacts may not be apparent during development. By **deployed**, we mean agents that are in use, whether the agent is available to the general public, select customers, or only for internal use within the organization that develops it. Visibility into the last case may be particularly important if organizations that carry out crucial societal functions, such as banks or cloud compute providers, develop and deploy their own AI agents. We focus on deployment because pre-deployment testing [9, 152] does not account for how users or deployers may exacerbate risks [173] through fine-tuning [49], connecting to external tools or services [122, 124, 125, 146], or structuring calls to the system so as to better enable it to pursue goals [133, 146, 172, 176]. Even instances of agents that come from the same underlying system can access different tools and can be conditioned to behave differently based on prompts.

1.3 Contributions

In this paper, we assess three categories of measures to increase visibility into AI agents: agent identifiers, real-time monitoring, and activity logs. For each, we outline potential implementations that vary in intrusiveness of data collection and informativeness of the data. We analyze how our measures apply across a spectrum of centralized through decentralized deployment contexts, accounting for various actors in the supply chain including hardware and software service providers. Finally, we discuss the implications of our measures for privacy and concentration of power. Rather than advocating for immediate implementation of these measures, we emphasize the need for further understanding them and how to mitigate their negative impacts.

The measures extend existing work in deployment visibility to better account for the risks of AI agents. Agent identifiers, which indicate whether and which AI agents are involved in interactions, generalize watermarks [105, 170] because they apply to all of an agent's outputs, including the use of external tools and services, not just text, image, or audio outputs. This generalization is crucial for improving visibility if agents increasingly substitute for human

[5] We use *visibility* rather than *transparency* as we believe the former to be somewhat more common in a regulatory context. Both terms are distinguished from *explainability*, which refers to whether one can understand why an AI system generated a particular output [109]

actions. For real-time monitoring and activity logging, we assess practices that extend existing schemes [80, 91, 92, 111, 151] so as to better track complex interactions between multiple agents [45, 143], an agent's interaction with external tools or services, and delayed and diffuse effects of an agent's actions.

2 DEFINITIONS

Besides the definitions here, we also define each term when we use it for the first time in the main body. We illustrate the most common terms in Figure 1.

Agency is the degree to which an AI system acts directly in the world to achieve long-horizon goals, with little human intervention or specification of how to do so. An **(AI) agent** is a system with a *relatively high* degree of agency; we consider systems that mainly predict without acting in the world, such as image classifiers, to have relatively low degrees of agency. Examples of agents include reinforcement learning systems [132, 160] that interact extensively with the real world[6] or more capable versions of language models with tool access [27, 122, 137, 144]. We do not consider existing foundation models themselves to be agents. Our definitions compress the characterization of agency in Chan et al. [35], which points to four axes: the degree to which the system's behaviour is specified, the degree to which the system's behaviour is goal-directed, the degree to which the system has a direct impact in the world, and the degree to which the system can achieve goals over long time-horizons. For the purposes of this paper, we use agent and **agentic system** interchangeably.

Scaffolding is any method that structures the calls to an AI system so as to facilitate the pursuit of goals [36, 133, 176]. Scaffolding may include additional prompts, memory systems, access to external tools, and planning mechanisms [169]. For example, AutoGPT [133] has a language model accept a high-level goal and sequentially produce *(reasoning, plan, criticism of the plan, action)* tuples so as to achieve the goal. Scaffolding can make an AI system, such as a foundation model, more agentic.

The term **developer(s)** refers to the actor(s) involved in the construction of an AI system. While the developers of a system include those who trained the underlying machine-learning model, developers could also include those who build other components of the complete system, such as the scaffolding [133, 176].

The **user** is the human individual or group that interacts with and provides instructions to an AI system.

The **deployer** is the entity that operates an AI system and serves it to users. The deployer may not be the same as the developer(s). For example, Microsoft deploys OpenAI's systems into its products [171], but did not develop GPT-4. A deployer may provide access to an agent in one of two ways. First, the deployer may serve a foundation model which users may combine with other components to make a more[7] agentic system. For example, users may use the scaffolding framework AutoGPT [133] to chain calls to GPT-4 [123] and provide the model with tool access. Indeed, popular scaffolding frameworks depend upon an underlying foundation model [133, 176]. Second, the deployer may provide an agent or may furnish ways for users to make a provided system more agentic. For instance, OpenAI lets users build and share custom agents [124, 125] augmented with a variety of tools, including browsing, using Google Drive apps, and coding [122].

The **compute provider** is responsible for supplying and maintaining the hardware infrastructure on which an AI system operates. The compute provider could also be the same as the deployer or the developer if either runs its own infrastructure. Compute providers could be important partners for overseeing large-scale deployments of agents that are not run by deployers, which we discuss further in Section 4.

A **tool** or **service** refers to an external system or platform with which an AI agent interacts to perform its tasks. For instance, a flight booking website where the AI agent executes transactions, such as purchasing plane tickets on behalf of the user, would be a service. The **provider** of the tool or service is responsible for maintaining the system or platform. We will often use *tool* and *service* interchangeably. Agents often interact with tools or services through dedicated **APIs**, which are interfaces and protocols specifically structured for agents, rather than human users.

The **outputs** are the results or responses that an AI system generates. Some types of outputs include images, text, or actions (e.g., calling a tool). While the deployer generally has access to all the outputs, the tool or service provider's knowledge is limited to outputs relevant to their specific service (e.g., the results of an API call).

Inputs are the data that the AI agent receives from a user, a tool or service, another agent, or any other party, which inform its actions or responses. The deployer in principle has access to inputs by virtue of running the system, but may choose not to collect or store such information out of respect for user privacy.

3 MEASURES TO IMPROVE VISIBILITY

We propose three complementary categories of measures to improve visibility into AI agents. **Agent identifiers** indicate whether and which AI agents are involved in a given interaction, such as watermarks or IDs that distinguish agents in their requests to service providers. **Real-time monitoring** involves real-time analysis of an agent's activity, allowing deployers and/or service or tool providers to flag and intervene on problematic behaviour as it is occurring. **Activity logs** held by deployers and tool or service providers record certain inputs and outputs of an agent, such as interacting with external services or other agents, to facilitate post-incident attribution and forensics. See Figure 2 for an overview of the information flows for each measure.

Each category contains measures that vary in intrusiveness of data collection and informativeness. More comprehensive information collection may be justified for agents deemed to be high-risk, potentially based on the results of evaluations [9, 98, 108, 135, 137, 152] or deployment in high-risk domains [2, 45, 100]. For example, it may be desirable to subject agents involved in financial trading to monitoring requirements at least as strict as those for human traders [121]. Yet, more comprehensive data collection may have serious privacy risks, which we discuss in Section 3.4. Our goal is to provide an array of options, rather than an answer to these trade-offs and the extent to which visibility measures should be

[6] Including the physical environment but also digital environments such as online platforms.
[7] The system may not be completely autonomous since user approval may still be required for certain actions.

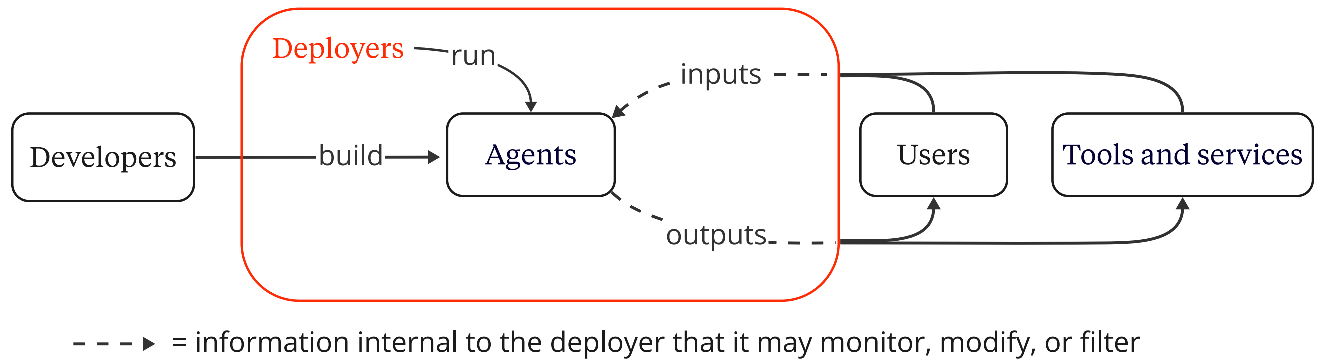


Figure 1: We illustrate how our main terms in Section 2 interact with each other. Deployers are in red and encompass the agents box to denote the fact that our paper focuses on agents that are run by deployers and served to users. Developers build agents (or an underlying system) and deployers serve instances of agents to users. Since deployers run agents, the inputs and outputs of agents are by default visible to the deployer, which facilitates the measures that we discuss in Section 3.

mandated. Finally, while we discuss potential implementations of the measures, more research is required to understand their feasibility and implications.

We focus in this section on agents run by **(agent) deployers**—entities that deploy agents, or important subcomponents like a foundation model, as a service to users. We include foundation models because many frameworks for constructing agents use a foundation model as the central component [122, 133, 176]. While deployers are unlikely to account for all agent activity, they likely constitute a substantial fraction because the most capable foundation models are only available through deployers [11, 123, 161],[8] and using a deployer may be more convenient than running a system oneself. Moreover, since deployers can already see the inputs and outputs of deployed systems, they can attach agent identifiers to outputs, perform real-time monitoring, and collect activity logs. In Section 4, we analyze how to extend our measures to decentralized deployments of agents.

3.1 Agent Identifiers

An agent identifier indicates whether and which AI agents are involved in interactions. Agent identifiers are attached to select outputs, are visible to certain actors, and may include additional information about the agent.

The ability to identify agents could be useful to several actors. A regulator could require AI agents to identify themselves as non-humans during interaction with humans [13, 65], similar to bot disclosure laws [54].[9] Members of the general public may wish to know whether they have interacted with AI agents. Summary statistics based on agent identifiers could inform governments and the general public about the extent to which agents operate in high-risk settings [2]. Identifiers for when agents send requests to tools or services providers could help to identify significant actions, such as when agents transfer sensitive information. A service provider may even reject a request absent certain guarantees attached to the identifier, such as those related to the security of the agent. Unique identifiers for each AI agent could facilitate accountability by linking an action to an AI agent and its user, developer(s), and deployer.

[8] Currently, these deployers also happen to be developers.

[9] Note that not all non-human activities may come from AI agents. For example, consider currently automated trading activity or ads auctions.

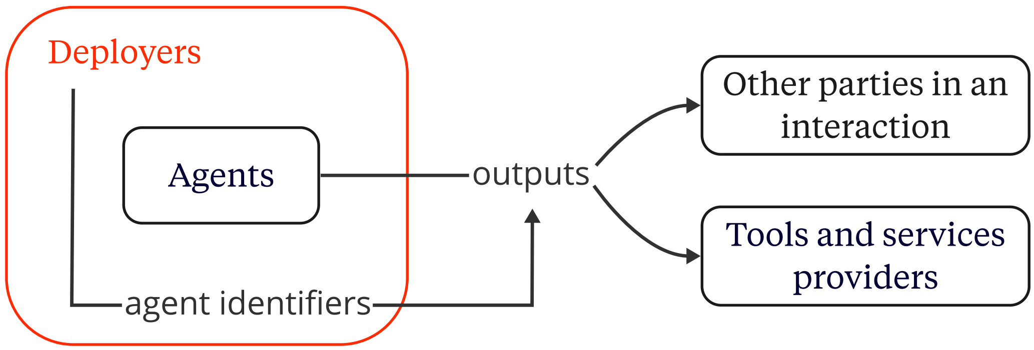
3.1.1 *Types of Agent Identifiers.*

We consider three key design decisions for an agent identifier:

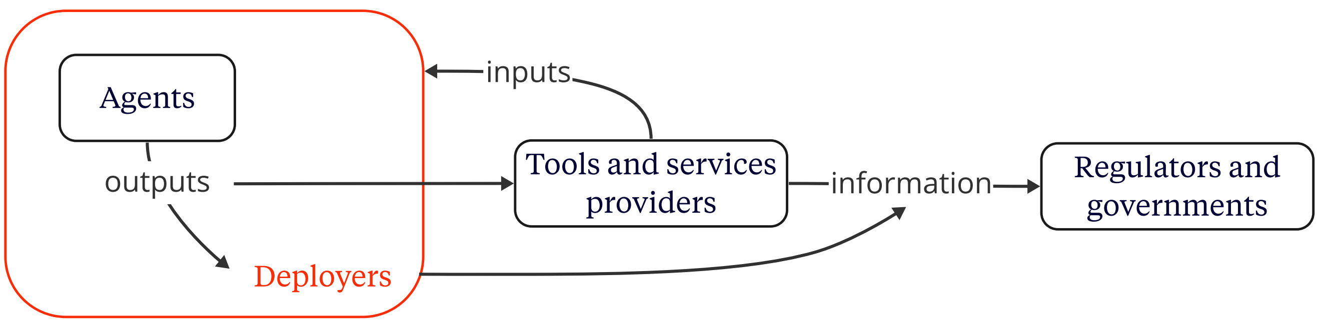
Which outputs contain the identifier? Decisions to attach an identifier may consider both the format and content of the output. By format, we mean whether the agent outputs data such as images, text, audio, or API requests to a service provider. An agent identifier's specific implementation depends upon the output format. For example, identifiers for image outputs could be watermarks [105, 170], while an identifier for an API request could be a simple header, similar to headers in HTTP requests. The difficulty of implementing identifiers varies based on the format of the output: for instance, adversarial users may easily remove watermarks [178]. Regarding content, identification may be especially important for significant outputs. Certain outputs, such as purchases made on behalf of the user, may merit identifiers by virtue of the task the agent is accomplishing. Other outputs may only be significant beyond a certain threshold, such as requests for compute resources that exceed a certain amount.

Which actors can see the identifier? An identifier may need to be visible to different actors. For example, in the context of a financial transaction, an agent identifier could be visible to any combination of the bank, the other party in the transaction, or the service provider for the bank API (which could be the bank itself). Some actors may need agent identifiers to fulfil their existing duties, such as e-commerce websites which must authenticate users and safeguard customer payment information [44]. Furthermore, facilitating the identification of multi-agent risks may require that agent identifiers be visible to other agents.

How specific is the identifier to a particular agent? An identifier could point to a particular agent, or simply denote that *some* agent was involved in the interaction. The former could facilitate incident reporting and investigation. To implement unique identifiers



(a) An agent identifier indicates to certain actors whether an AI agent is involved in an interaction. Developers (not shown) and deployers cooperate to implement agent identifiers, which the latter adds onto outputs. Here, we illustrate an agent identifier that informs other parties in a given interaction with an agent, as well as tools and services providers. If these actors know that they are interacting with an agent, they may wish to verify certain properties such as the security or robustness of the agent. See Section 3.1 for further discussion.



(b) An agent's inputs and outputs are visible to the deployer. Inputs come from tool and service providers and users (not shown). Certain outputs, such as requests to external tools and services, are also visible to tools and services providers. These actors can monitor and filter the actions in real-time (Section 3.2) or keep logs (Section 3.3) for post-incident attribution or forensics. Insights gained from real-time monitoring or from logs can inform regulators and governments.

Figure 2: We illustrate the flow of information for our measures in Section 3.

for each deployed instance of an agent, cryptographic methods—such as those used in software attestation—may be needed to assure the agent's provenance. In Section 3.1.2, we discuss additional, useful information that could be attached to an identifier.

3.1.2 Attaching Additional Information to Agent Identifiers. Additional information attached to an agent identifier may be of further use. Additional information could be specific to the instance of the agent deployed to the user, or could pertain to the **underlying system** used during agent development. Information about the former could include the goals the user has given its agent, while information about the latter could include the results of evaluations performed on the underlying system. We refer to the set of such additional information as an **agent card**, drawing inspiration from previous work on documenting AI systems [24, 68, 69, 112]. In Appendix A we provide a more comprehensive list of what could be included on an agent card, but in this section we discuss three particularly important types of information that could be included.

The underlying system. This information could include the results of evaluations [152]; previous incidents; the dependencies involved in the system's construction [24, 68, 112]; or training methods and data used. Such information could inform the decisions of actors that interact with agents. For example, tool providers may reject requests from agents that do not meet certain security standards [16].

The specific instance of the agent. This information could consist of how the agent was deployed (e.g., by its user directly or by another agent); a list of external tools or services that the agent can access (e.g., applications or software, any physical tools); the scaffolding framework (e.g. memory or planning mechanisms); the intended scope, permissions, and goal(s) of the agent [69]; or the intended sector of deployment (e.g., finance). This type of information could be useful for regulators to develop an overall understanding of the extent and significance of agent use.

The actors involved in agent development and deployment. These actors could include the user, the compute provider, developers, and other humans with whom the agent has interacted. Authentication of the agent's user may be useful when individuals are providing sensitive information to agents (e.g. agents used by banks to assist customers). Furthermore, this information helps to clarify *whom* to hold accountable in case an agent causes harm during its operation [43, 118]. A regulator could use identifying information associated to each of the actors, ideally in synergy with data from Sections 3.2 and 3.3, to help investigate and address problems. For example, a regulator may receive reports from third parties of an agent engaging in financial scams. If the transactions of the agent contain agent identifier, a regulator could match the identifier to the corresponding agent card and ask the deployer of the agent to investigate further or shut it down.

A more comprehensive agent card may be more useful, but may also be more intrusive to collect and require more substantial resources and expertise for analysis. Comprehensiveness of information should likely scale with the risk of the activity in question.

3.2 Real-Time Monitoring

Real-time monitoring involves oversight of agent activity in real-time, with the goal of flagging and potentially filtering problematic behaviour. Such monitoring is likely to be automated so as to be able to cope with the speed and scale of agent activities [74, 147]; see Naihin et al. [116] for a preliminary implementation of a real-time monitoring system for agents. Crucially, real-time monitoring does not require the collection or storage of activity logs. As an analogy, consider the difference between moderation filters and chat logs in the context of a text messaging application—moderation filters might instantly delete harmful content that trips a filter, but this measure is separate from keeping logs for subsequent analysis. This difference can be critical for users who are concerned about the privacy implications of a deployer keeping logs of user inputs, such as business customers who need to consider the security of their corporate IP. Indeed, some cloud providers already offer no-logging provisions[10] for their language model deployments to some customers, subject to real-time monitoring for abuse [37, 71, 114].

Once behaviour is flagged, a deployer can intervene on the agent and potentially inform relevant authorities. A user could be given the opportunity to request human review of any action taken to impede its agent. We leave details about the implementation challenges of automation and human review to other work [72, 95, 157, 166].

Since real-time monitoring requires an immediate decision to flag a behaviour, deployers would only be able to find problems which they can identify in real-time. Thus, real-time monitoring may be most suitable for identifying violations of clear rules such as restrictions on allowed activities of agents, limits on tools and permissions made available to agents, leakage of sensitive information, or usage caps for particular users or agents. Behaviors can also be flagged for anomalies.

Flagging behaviors might also involve combining information from multiple agents because an individual agent's behaviour may only be problematic in the context of other agents. For instance, collusion requires a group of agents, not just a single agent, to coordinate to set high prices [55]. Similarly, a single agent rapidly moving resources (e.g., data or financial investments) between entities may not be an issue in isolation, but vast numbers of agents engaging in the same behaviour might overload a given system (e.g., server outages or stock crashes). A real-time monitoring system could use agent identifiers to track which agents were involved in interactions together. This system could raise flags if indicators of the collective activity of these agents pass certain thresholds. For example, an indicator for algorithmic collusion could be correlation in price movements.

Real-time monitoring could also flag when an agent exerts significant influence in its environment [127]. Such activities could include economic transactions above a certain amount, usage of substantial compute resources, running a large number of sub-agents, or simultaneous interfacing with a large number of physical systems. While the acquisition of significant influence may not itself cause harm, the impact of an agent's failure would be magnified. A potential response to such a flagged activity could be pausing it until the deployer can verify explicit user approval, just as banks monitor transactions and temporarily suspend accounts if they suspect fraud.

3.3 Activity Logs

Activity logs are records of certain inputs and outputs of an agent. Inputs and outputs could be from both users or tools and services, as we illustrate in Figure 1. Records could also report changes in the state of the agent, such as updates to the weights. Logs can include varying levels of detail and may only record actions above a certain degree of significance, depending on the behaviour one is trying to track, privacy considerations, and storage requirements. Frameworks for building agents already implement logging of certain actions [36, 133, 176].

Activity logs can inform an understanding of the agent's impacts as well as allow one to tie behaviour to specific user choices, facilitating post-incident attribution and forensics [119]. During audits and incident investigation, investigators may need detailed information to trace the source and causes of harms arising from AI agents' actions [115, 131]. Researchers may benefit from detailed logs to improve our understanding of agents and to develop better deployment controls. Analysis of logs may also identify novel behaviours to improve real-time monitoring.

3.3.1 Agent-Specific Information. While activity logs are useful for understanding risks from AI systems in general [91, 92] and are common practice in many domains [3, 4, 41, 42, 89], they can be especially helpful for addressing the risks we identified in Section 1.1. Different categories of logs could exist, such as for tool use, internal reasoning [172], self-critique [141], memory management, etc, each of which may be useful to different actors.

Understanding the impacts of agents involves identifying which scaffolding, tools, and services they use. For example, an agent with external memory can plan and act over longer horizons. Logs of how an agent manages its memory may be helpful for a user in trying to understand the agent. More generally, a log could explicitly label when an agent has accessed an external tool and the tool's output. Both the deployer and the tool provider could engage in such logging: deployer logs may be more useful for understanding how the tool affects the overall behaviour of the agent, while tool provider logs may provide more insight into the impacts of the tool itself. Indeed, tool providers may have strong incentives to

[10] For feasibility, sometimes logs must be kept for a short amount of time. The no-logging provision of Google [71] keeps logs for 72 hours.

engage in such logging themselves: for example, tool providers can study logs to update APIs or user interfaces to prevent abuse. Tool providers can also decide to restrict services to certain AI agents with identifiers (and potentially other attached certifications), as discussed in Section 3.1.

Identifying delayed and diffuse impacts may require logs to be retained for extended periods of time. Details about the persistence of the agent could be included in the logs, such as its running time, whether it is writing to and accessing external memory, or the amount of compute used so far to run the agent. These details could inform interventions, such as limiting the lifetimes of certain agents. Yet, significant impacts may arise after the lifetime of an individual agent. The impact of the original agent could be delayed, or a user could run another agent for the same purposes, potentially even with the same inputs and memory as the original agent. Accounting for these possibilities means that logs may have to persist for a significant amount of time after the lifetime of the corresponding agent. Furthermore, logs for different agents may have to be combined if one agent can be viewed as a continuation of another.

Combining information from multiple logs may also help to understand sub-agent and multi-agent dynamics. For example, agent logs could be used to build models of how a particular malfunction might propagate through a network of agents or identifying undesirable forms of communication between agents [136].

3.3.2 Logging at Different Levels of Detail. A key design decision is the level of detail at which to record the agent's actions. Less detailed logging may only record high-level summaries of agent's behaviour or certain samples thereof. At the finest level, a regulator may require a deployer to record in detail all of an agent's behavior, especially if an agent is operating in a high-risk environment. More detailed logging is more useful, but may impose more significant costs on the deployer, require more resources and expertise for analysis, and pose more significant privacy concerns.

3.4 Risks

Privacy considerations may conflict with obtaining detailed information about agent activity. Language model deployers are increasingly offering customers, particularly business customers, privacy assurances around data collection and use. Measures to reassure customers about confidentiality include:

- Language model APIs with no logging of inputs or outputs, and the ability to turn off safety filters and moderation classifiers [37].
- Guarantees that customer data, including system outputs, will not be used to train any AI system and will be kept in a customer's cloud instance [38].
- The ability to delete logs kept by the provider after a certain amount of time [37].

Additionally, existing data protection laws, such as GDPR, impose further restrictions. Agent cards may contain identifying information about users. Agent logs may be considered personal data, such as when agents are given access to a filesystem containing personally identifiable information [64].

In general, if agents substitute for humans in a wide variety of activities, information about those agents might be tantamount to information about the users of those agents. Indeed, agent activities may be easier to monitor than human activities because deployers are a central intermediary. Governments or deployers may thus abuse their power to carry out excessive or unjustified surveillance of personal activities [70, 75]. These considerations justify limiting data collection in accordance with the risk of the agent's activities or domain of deployment. Another potential mitigation may be decentralized data custody schemes or data trusts [52, 53] whereby users or accountable representatives would make decisions about data usage.

Modulating the degree of access to collected information can also help to mitigate privacy concerns. Access can vary with respect to **granularity**, the amount of detail contained in the records, and **quantity**, the number of records that a party is allowed to access. With respect to granularity, information can be aggregated, de-identified, or identifiable. Aggregated information involves summary statistics but not individual records or logs; differentially private [20] computations of summary statistics may help to preserve the privacy of individual records. Records and logs can be de-identified with respect to individual users or identifiable. With respect to quantity, a party can have full access to all records, access based on approved search queries or filters, or access upon-request to pre-specified records to which they must provide a compelling reason for access.

The granularity and quantity of access should be the minimum necessary for the accessing party to achieve its (legitimate) objectives. When investigations pertain to specific users, identifiable information could be made available upon request given a showing of compelling need and/or after approval from a third-party adjudicator. Regulators may need logs containing identifiable information in some cases, such as oversight of certain high-risk or high-volume activities. For example, for traders transacting above a specified threshold, CFTC collects identifiable personal information to enable aggregation of data across different accounts and brokers [33].

4 DECENTRALIZED DEPLOYMENTS

Some deployments of agents may occur in a decentralized way and bypass deployers. Users, whether enterprises or individuals, may run downloadable (i.e., open release) [156] agents either on cloud compute or on their own hardware. A user may even be able to combine systems from different deployers to form an agent. Although visibility on the resulting agent may be desirable, the individual systems may not be significant enough by themselves to justify implementation of visibility measures by deployers. Indeed, a malicious actor could build and run an agent in this way so as to avoid detection by regulatory authorities. In this section, we discuss how our visibility measures may be extended to such situations, as well as the risks of doing so.

4.1 Compute Provider Oversight

Compute providers could enable oversight over deployments that involve large quantities of compute. Large-scale deployments could be concerning because they might involve vast numbers of agents, which could translate into a large impact multiplier for the user. Large-scale deployments are also noticeable because they consume significant resources. Compute providers may have cost advantages over users deploying their own hardware because of economies of scale. Indeed, using compute as a service (e.g., infrastructure as a service or cloud) is the default way for a business to deploy its IT services. If a compute provider can identify large-scale deployments and whether they correspond to agent activities, they may ask the user for proof that they have implemented certain visibility measures [56, 119].

4.2 Tool and Service Providers as Distributed Enforcement Mechanisms

The need for agents to interact with external tools offers another leverage point. By conditioning tool and service access on implementation of certain visibility measures such as agent identifiers, tool and service providers can incentivize adherence to the measures. For example, financial institutions could restrict access to AI agents without identifiers from certain trusted deployers. Such identifiers might explicitly confirm permissions to access certain services, such as performing financial transactions or accessing certain websites. This approach could also allow tool providers to minimize misuse and enable detailed analytics of AI agent interactions with their tools.

One limitation is that AI agents could circumvent APIs by directly interacting with tools in a way that mimics human behavior. The development of tools capable of detecting disguised AI activity, akin to modern CAPTCHA systems designed to differentiate between human and software interactions, may be helpful. An alternative is to require proof of human identity for high-risk actions. Certain industries perform identity verification for high-risk activities with know-your-customer protocols [56]. Similarly, tool or service providers operating in high-risk domains could require human identification. One difficulty is preventing AI agents from spoofing humans, such as through generating fake identification documents or stealing real ones. While CAPTCHA-like tests are a possibility, measures should be robust to improvements in the capabilities of agents. How to balance privacy considerations with the need for identity verification is another open question. A potential direction is to understand what mechanisms may allow humans to prove their status without identity disclosure.

While direct interaction with tools is possible, users and developers may still opt for the convenience and efficiency offered by APIs, especially if direct interaction is more complex. APIs can provide standardized interfaces, tailored services for AI agent use, and can set specific conditions like access rates or the scope of services available. This preference for API interaction could reduce the difficulty of obtaining visibility into decentralized deployments.

4.3 Risks

Extending visibility measures to decentralized deployments has serious implications for privacy and concentration of power. Compute providers that surveil deployments may be able to infer sensitive information about users. Given that a handful of compute providers dominates the market [134], monitoring users of those providers would equivalent to monitoring much of society. In addition to potential abuses of collected information, compute providers may also have lax security standards that enable attackers to gain sensitive information.

Enforcement through tool and service providers also faces similar concerns. If useful tools and services were unavaiable to agents that were not from certified deployers, users may face strong pressure to use agents from such deployers. Whether because of government demand [75] or regulatory capture [48], those deployers may have practises that are inimical to users or may not be responsive to user interests. Visibility measures for those deployers may be extremely invasive, such as excessive and unjustified logging. Agents from those deployers may not be well-suited to the user's use cases; for example, the user might require an agent to be able to operate in a low-resource language. If the market of deployers is heavily concentrated, further reliance upon them could exacerbate systemic risks [174].

One way to mitigate these risks is to explore voluntary standards for adopting visibility measures. Voluntary standards could allow experimentation to understand when and where visibility measures should be applied. Although voluntary standards might not provide visibility into malicious use or enjoy universal adherence, understanding gained from their adoption can aid their later codification.

Certain tools may aid the adoption of voluntary standards. For example, open-source frameworks to implement agent identifiers may allow users to avoid deployers and, at the same time, facilitate visibility. Even if tool and service providers reject requests from agents without identifiers, users may easily be able to add an identifier to their agents. Independent entities may be required to certify valid identifiers, similar to certificate authorities on the Internet. A source of inspiration may be Let's Encrypt, a non-profit certificate authority which provides a free, automatic certificate process and which has been instrumental in promoting the use of the more secure HTTPS [58]. Financial and technical support to develop open-source, agent identifier frameworks will likely be critical to the success of voluntary standards.

Another potential mitigation is to limit the number of tools or services that require agent identifiers. Obtaining visibility over tools involved in potentially high-risk activities, such as scientific platforms that handle pathogens or dangerous chemicals [22], may be of higher priority. Similarly, visibility into business uses of AI agents, rather than personal uses, may be more important. Regulations could require certain businesses (e.g., those above a certain size) that use AI agents to implement certain visibility measures.

Rather than mandating denial of requests from agents that do not have identifiers or that do not provide proof that certain visibility measures are implemented, one alternative is accounting for such compliance when determining the legal liability of users who deploy their own agents. Analogously, compliance with HIPAA de-identification standards can be taken into consideration to reduce regulatory fines or audits for violations [1, 120]. Similarly, the 2023 U.S. National Cybersecurity Strategy proposes to shield from private liability companies that follow cybersecurity best practices [87]. Accounting for compliance when determining liability may incentivize standards adherence.

Other potential mitigations to explore include decentralized data custody schemes for compute provider logs and enhanced transparency into both data collection practises and government requests for data [70].

5 CONCLUSION

Visibility facilitates the governance of increasingly agentic systems. We assessed three mechanisms for visibility: agent identifiers, real-time monitoring, and activity logs. Agent identifiers indicate whether and which agents are involved in an interaction. To aid accountability and incident investigation, an agent card containing additional information about the agent can be attached to an agent identifier. Real-time monitoring aims to flag problematic agent behaviour as it happens. Activity logs record certain inputs and outputs of agents so as to enable in-depth, post-hoc analysis of behaviour. We examined how to extend the visibility measures to decentralized deployments of agents, in particular through using compute providers and tools and services providers to obtain visibility. Finally, we analyzed the implications of the visibility measures on privacy and concentration of power. Rather than advocating for immediate implementation of these measures, further understanding of the measures and how to mitigate their negative impacts is required. Such understanding can help to build a foundation for the governance of AI agents.

Visibility informs actions to manage risks from the deployment of increasingly agentic systems, but is not by itself sufficient. Even with a comprehensive understanding of agent activities, those harmed by them may not have the power to intervene and reduce risks [8]. To best make use of visibility, future work could investigate increasing public influence over AI development and deployment [34, 88, 126, 145], developing a wide range of potential policy levers [9, 56, 147], and implementing infrastructure and practices to prevent or defend against harms [29, 139].

ACKNOWLEDGMENTS

We are grateful to the following people for insightful feedback and conversations over the course of writing this work: Charlotte Siegmann, Chinmay Deshpande, Nathan Barnard, Leonie Koessler, Kevin Frazier, Matthijs Maas, Micah Carroll, Yonadav Shavit, Merlin Stein, Ben Bucknall, Shalaleh Rismani, Paul Crowley, Gabriel Recchia, Janet Egan.

REFERENCES

[1] [n. d.]. 42 U.S. Code § 17941 - Recognition of security practices. https://www.law.cornell.edu/uscode/text/42/17941
[2] 2021. 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. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206
[3] 2022. 12 CFR § 1026.25 - Record retention. *Federal Register* (Jan. 2022).
[4] 2023. 14 CFR § 91.609 - Flight data recorders and cockpit voice recorders. *Federal Register* (Jan. 2023).
[5] 2023. character.ai. https://beta.character.ai/
[6] Daron Acemoglu and Pascual Restrepo. 2018. Artificial Intelligence, Automation, and Work. In *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press, 197–236. https://www.nber.org/books-and-chapters/economics-artificial-intelligence-agenda/artificial-intelligence-automation-and-work
[7] Daron Acemoglu and Pascual Restrepo. 2019. Automation and New Tasks: How Technology Displaces and Reinstates Labor. *Journal of Economic Perspectives* 33, 2 (May 2019), 3–30. https://doi.org/10.1257/jep.33.2.3
[8] Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* 20, 3 (March 2018), 973–989. https://doi.org/10.1177/1461444816676645 Publisher: SAGE Publications.
[9] Markus Anderljung, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, Ben Chang, Tantum Collins, Tim Fist, Gillian Hadfield, Alan Hayes, Lewis Ho, Sara Hooker, Eric Horvitz, Noam Kolt, Jonas Schuett, Yonadav Shavit, Divya Siddarth, Robert Trager, and Kevin Wolf. 2023. Frontier AI Regulation: Managing Emerging Risks to Public Safety. https://doi.org/10.48550/arXiv.2307.03718 arXiv:2307.03718 [cs].
[10] (Max) Hui Bai, Jan G. Voelkel, johannes C. Eichstaedt, and Robb Willer. 2024. Artificial Intelligence Can Persuade Humans on Political Issues. (Jan. 2024). https://doi.org/10.31219/osf.io/stakv Publisher: OSF.
[11] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. https://doi.org/10.48550/arXiv.2212.08073 arXiv:2212.08073 [cs].
[12] Luke Bailey, Euan Ong, Stuart Russell, and Scott Emmons. 2023. Image Hijacks: Adversarial Images can Control Generative Models at Runtime. https://arxiv.org/abs/2309.00236v2
[13] Tessa Baker. 2023. The EU AI Act: A Primer. https://cset.georgetown.edu/article/the-eu-ai-act-a-primer/
[14] Kirstie Ball. 2010. Workplace surveillance: an overview. *Labor History* 51, 1 (Feb. 2010), 87–106. https://doi.org/10.1080/00236561003654776 Publisher: Routledge _eprint: https://doi.org/10.1080/00236561003654776.
[15] Matthew Barnett and Tamay Besiroglu. 2023. The Direct Approach. https://epochai.org/blog/the-direct-approach
[16] Manish Bhatt, Sahana Chennabasappa, Cyrus Nikolaidis, Shengye Wan, Ivan Evtimov, Dominik Gabi, Daniel Song, Faizan Ahmad, Cornelius Aschermann, Lorenzo Fontana, Sasha Frolov, Ravi Prakash Giri, Dhaval Kapil, Yiannis Kozyrakis, David LeBlanc, James Milazzo, Aleksandar Straumann, Gabriel Synnaeve, Varun Vontimitta, Spencer Whitman, and Joshua Saxe. 2023. Purple Llama CyberSecEval: A Secure Coding Benchmark for Language Models. https://doi.org/10.48550/arXiv.2312.04724 arXiv:2312.04724 [cs].
[17] David Biagioni, Xiangyu Zhang, Dylan Wald, Deepthi Vaidhynathan, Rohit Chintala, Jennifer King, and Ahmed S. Zamzam. 2022. PowerGridworld: a framework for multi-agent reinforcement learning in power systems. In *Proceedings of the Thirteenth ACM International Conference on Future Energy Systems (e-Energy '22)*. Association for Computing Machinery, New York, NY, USA, 565–570. https://doi.org/10.1145/3538637.3539616
[18] Julia A. Bielicki, Xavier Duval, Nina Gobat, Herman Goossens, Marion Koopmans, Evelina Tacconelli, and Sylvie van der Werf. 2020. Monitoring approaches for health-care workers during the COVID-19 pandemic. *The Lancet. Infectious Diseases* 20, 10 (Oct. 2020), e261–e267. https://doi.org/10.1016/S1473-3099(20)30458-8
[19] Patrick Birkinshaw. 2006. Freedom of Information and Openness: Fundamental Human Rights? *Administrative Law Review* 58, 1 (2006), 177–218. https://www.jstor.org/stable/40712007 Publisher: American Bar Association.
[20] Emma Bluemke, Tantum Collins, Ben Garfinkel, and Andrew Trask. 2023. Exploring the Relevance of Data Privacy-Enhancing Technologies for AI Governance Use Cases. https://doi.org/10.48550/arXiv.2303.08956 arXiv:2303.08956 [cs].
[21] Board of Governors of the Federal Reserve System. 2021. Proactive Monitoring of Markets and Institutions. https://www.federalreserve.gov/financial-stability/proactive-monitoring-of-markets-and-institutions.htm
[22] Daniil A. Boiko, Robert MacKnight, and Gabe Gomes. 2023. Emergent autonomous scientific research capabilities of large language models. http://arxiv.org/abs/2304.05332 arXiv:2304.05332 [physics].
[23] Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang. 2022. Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization? https://doi.org/10.48550/arXiv.2211.13972 arXiv:2211.13972 [cs].
[24] Rishi Bommasani, Dilara Soylu, Thomas I. Liao, Kathleen A. Creel, and Percy Liang. 2023. Ecosystem Graphs: The Social Footprint of Foundation Models. https://doi.org/10.48550/arXiv.2303.15772 arXiv:2303.15772 [cs].
[25] Robert M. Bond, Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. 2012. A 61-million-person experiment in social influence and political mobilization. *Nature* 489, 7415 (Sept. 2012), 10.1038/nature11421. https://doi.org/10.1038/nature11421
[26] Samuel Bowman. 2022. The Dangers of Underclaiming: Reasons for Caution When Reporting How NLP Systems Fail. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 7484–7499. https://doi.org/10.18653/v1/2022.acl-long.516
[27] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. 2023. ChemCrow: Augmenting large-language models with chemistry tools. https://doi.org/10.48550/arXiv.2304.05376

arXiv:2304.05376 [physics, stat].

[28] Jenna Burrell and Marion Fourcade. 2021. The Society of Algorithms. *Annual Review of Sociology* 47, 1 (2021), 213–237. https://doi.org/10.1146/annurev-soc-090820-020800 _eprint: https://doi.org/10.1146/annurev-soc-090820-020800.

[29] Vitalik Buterin. 2023. My Techno-Optimism. https://vitalik.eth.limo/general/2023/11/27/techno_optimism.html

[30] Ryan Carey and Tom Everitt. 2023. Human Control: Definitions and Algorithms. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*. PMLR, 271–281. https://proceedings.mlr.press/v216/carey23a.html ISSN: 2640-3498.

[31] John Cassidy. 2009. *How markets fail: The logic of economic calamities.* Farrar, Straus and Giroux.

[32] CFTC. 2023. CFTC Market Surveillance Program. https://www.cftc.gov/IndustryOversight/MarketSurveillance/CFTCMarketSurveillanceProgram/index.htm

[33] CFTC. 2023. Large Trader Reporting Program. https://www.cftc.gov/IndustryOversight/MarketSurveillance/LargeTraderReportingProgram/index.htm

[34] Alan Chan, Herbie Bradley, and Nitarshan Rajkumar. 2023. Reclaiming the Digital Commons: A Public Data Trust for Training Data. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*. Association for Computing Machinery, New York, NY, USA, 855–868. https://doi.org/10.1145/3600211.3604658

[35] 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. 2023. Harms from Increasingly Agentic Algorithmic Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 651–666. https://doi.org/10.1145/3593013.3594033

[36] Harrison Chase. 2022. LangChain 0.0.77 Docs. https://langchain.readthedocs.io/en/latest/modules/agents/getting_started.html

[37] ChrisHMSFT, PatrickFarley, mrbullwinkle, eric urban, and aahill. 2023. Data, privacy, and security for Azure OpenAI Service - Azure AI services. https://learn.microsoft.com/en-us/legal/cognitive-services/openai/data-privacy

[38] Google Cloud. 2023. *Generative AI, Privacy, and Google Cloud.* Technical Report. https://services.google.com/fh/files/misc/genai_privacy_google_cloud_202308.pdf

[39] Cary Coglianese, Richard Zeckhauser, and Edward Parson. 2004. Seeking Truth for Power: Informational Strategy and Regulatory Policy Making. *Minnesota Law Review* (Jan. 2004). https://scholarship.law.upenn.edu/faculty_scholarship/107

[40] Reuven Cohen and Shlomo Havlin. 2010. *Complex networks: structure, robustness and function.* Cambridge university press.

[41] Securities and Exchange Commission. 2022. 17 CFR § 240.17a-3 - Records to be made by certain exchange members, brokers and dealers. *Federal Register* (April 2022).

[42] Securities and Exchange Commission. 2022. 17 CFR § 240.17a-4 - Records to be preserved by certain exchange members, brokers and dealers. *Federal Register* (April 2022).

[43] A. Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. 2022. Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3531146.3533150

[44] PCI Security Standars Council. 2022. Payment Card Industry Data Security Standard.

[45] Andrew Critch and Stuart Russell. 2023. TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI. https://doi.org/10.48550/arXiv.2306.06924 arXiv:2306.06924 [cs].

[46] CTFC and SEC. 2010. *Preliminary findings regarding the market events of may 6, 2010.* Technical Report. U.S. Commodity Futures Trading Commission and U.S. Securities & Exchange Commission. https://www.sec.gov/sec-cftc-prelimreport.pdf tex.creationdate: 2023-10-29T16:47:57.

[47] Mary Cummings. 2004. Automation Bias in Intelligent Time Critical Decision Support Systems. In *AIAA 1st Intelligent Systems Technical Conference*. American Institute of Aeronautics and Astronautics. https://doi.org/10.2514/6.2004-6313

[48] Ernesto Dal Bó. 2006. Regulatory capture: A review. *Oxford review of economic policy* 22, 2 (2006), 203–225. Publisher: Oxford University Press.

[49] Tom Davidson, Jean-Stanislas Denain, Pablo Villalobos, and Guillem Bas. 2023. AI capabilities can be significantly improved without expensive retraining. https://doi.org/10.48550/arXiv.2312.07413 arXiv:2312.07413 [cs].

[50] Alejandro De La Garza. 2020. States' Automated Systems Are Trapping Citizens in Bureaucratic Nightmares With Their Lives on the Line. *TIME* (May 2020). https://time.com/5840609/algorithm-unemployment/

[51] Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de las Casas, Craig Donner, Leslie Fritz, Cristian Galperti, Andrea Huber, James Keeling, Maria Tsimpoukelli, Jackie Kay, Antoine Merle, Jean-Marc Moret, Seb Noury, Federico Pesamosca, David Pfau, Olivier Sauter, Cristian Sommariva, Stefano Coda, Basil Duval, Ambrogio Fasoli, Pushmeet Kohli, Koray Kavukcuoglu, Demis Hassabis, and Martin Riedmiller. 2022. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature* 602, 7897 (Feb. 2022), 414–419. https://doi.org/10.1038/s41586-021-04301-9 Number: 7897 Publisher: Nature Publishing Group.

[52] Sylvie Delacroix and Neil D Lawrence. 2019. Bottom-up data Trusts: disturbing the 'one size fits all' approach to data governance. *International Data Privacy Law* 9, 4 (Nov. 2019), 236–252. https://doi.org/10.1093/idpl/ipz014

[53] Sylvie Delacroix, Joelle Pineau, and Jessica Montgomery. 2020. Democratising the Digital Revolution: The Role of Data Governance. https://papers.ssrn.com/abstract=3720208

[54] Renee DiResta. 2019. A New Law Makes Bots Identify Themselves—That's the Problem. *Wired* (July 2019). https://www.wired.com/story/law-makes-bots-identify-themselves/ Section: tags.

[55] Florian E. Dorner. 2021. Algorithmic collusion: A critical review. https://doi.org/10.48550/arXiv.2110.04740 arXiv:2110.04740 [cs].

[56] Janet Egan and Lennart Heim. 2023. Oversight for Frontier AI through a Know-Your-Customer Scheme for Compute Providers. https://doi.org/10.48550/arXiv.2310.13625 arXiv:2310.13625 [cs].

[57] Jeffrey C. Ely and Balazs Szentes. 2023. Natural selection of artificial intelligence. (2023).

[58] Let's Encrypt. 2024. Let's Encrypt Stats - Let's Encrypt. https://letsencrypt.org/stats/

[59] M. R. Endsley and D. B. Kaber. 1999. Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics* 42, 3 (March 1999), 462–492. https://doi.org/10.1080/001401399185595

[60] Mica R. Endsley and Esin O. Kiris. 1995. The Out-of-the-Loop Performance Problem and Level of Control in Automation. *Human Factors* 37, 2 (June 1995), 381–394. https://doi.org/10.1518/001872095779064555 Publisher: SAGE Publications Inc.

[61] Epoch. 2023. Key trends and figures in Machine Learning. https://epochai.org/trends

[62] Ege Erdil and Tamay Besiroglu. 2023. Algorithmic progress in computer vision. _eprint: 2212.05153.

[63] Eugene F. Fama and Michael C. Jensen. 1983. Agency Problems and Residual Claims. *The Journal of Law & Economics* 26, 2 (1983), 327–349. https://www.jstor.org/stable/725105 Publisher: [University of Chicago Press, Booth School of Business, University of Chicago, University of Chicago Law School].

[64] Michèle Finck and Frank Pallas. 2020. They who must not be identified—distinguishing personal from non-personal data under the GDPR. *International Data Privacy Law* 10, 1 (Feb. 2020), 11–36. https://doi.org/10.1093/idpl/ipz026

[65] Kevin Frazier. 2023. The Right to Reality. https://www.lawfaremedia.org/article/the-right-to-reality

[66] Chengguang Gan, Qinghao Zhang, and Tatsunori Mori. 2024. Application of LLM Agents in Recruitment: A Novel Framework for Resume Screening. https://doi.org/10.48550/arXiv.2401.08315 arXiv:2401.08315 [cs].

[67] Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Scott Johnston, Andy Jones, Nicholas Joseph, Jackson Kernian, Shauna Kravec, Ben Mann, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Christopher Olah, Dario Amodei, and Jack Clark. 2022. Predictability and Surprise in Large Generative Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1747–1764. https://doi.org/10.1145/3531146.3533229

[68] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (Dec. 2021), 86–92. https://doi.org/10.1145/3458723

[69] Thomas Krendl Gilbert, Nathan Lambert, Sarah Dean, Tom Zick, Aaron Snoswell, and Soham Mehta. 2023. Reward Reports for Reinforcement Learning. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*. Association for Computing Machinery, New York, NY, USA, 84–130. https://doi.org/10.1145/3600211.3604698

[70] Chloe Goodwin. 2018. Cooperation or resistance?: The role of tech companies in government surveillance. 131 (2018), 1722–1722.

[71] Google. 2023. Bard Privacy Help Hub - Bard Help. https://support.google.com/bard/answer/13594961?sjid=16420951458997305974-EU&visit_id=638406643042311657-3533103185&p=bard_pntos_retention&rd=1#retention&zippy=%2Cwhy-does-google-retain-my-conversations-after-i-turn-off-bard-activity-and-what-does-google-do-with-this-data

[72] Ben Green. 2022. The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review* 45 (July 2022), 105681. https://doi.org/10.1016/j.clsr.2022.105681

[73] David Green. 2023. Emergence in complex networks of simple agents. *Journal of Economic Interaction and Coordination* 18 (May 2023), 1–44. https://doi.org/10.1007/s11403-023-00385-w

[74] Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. 2024. AI Control: Improving Safety Despite Intentional Subversion. http://arxiv.org/abs/2312.06942 arXiv:2312.06942 [cs].

[75] Glenn Greenwald and Ewen MacAskill. 2013. NSA Prism program taps in to user data of Apple, Google and others. *The Guardian* (June 2013). https://www.theguardian.com/world/2013/jun/06/us-tech-giants-nsa-data

[76] Valentina Guarnieri, Maria Moriondo, Mattia Giovannini, Lorenzo Lodi, Silvia Ricci, Laura Pisano, Paola Barbacci, Costanza Bini, Giuseppe Indolfi, Alberto Zanobini, and Chiara Azzari. 2021. Surveillance on Healthcare Workers During the First Wave of SARS-CoV-2 Pandemic in Italy: The Experience of a Tertiary Care Pediatric Hospital. *Frontiers in Public Health* 9 (2021). https://www.frontiersin.org/articles/10.3389/fpubh.2021.644702

[77] Lewis Hammond and TBD. 2024. *Multi-Agent Risks from Advanced AI.* Technical Report.

[78] Julian Hazell. 2023. Large Language Models Can Be Used To Effectively Scale Spear Phishing Campaigns. https://doi.org/10.48550/arXiv.2305.06972 arXiv:2305.06972 [cs].

[79] Dan Hendrycks. 2023. Natural Selection Favors AIs over Humans. https://doi.org/10.48550/arXiv.2303.16200 arXiv:2303.16200 [cs].

[80] Thomas Henzinger, Mahyar Karimi, Konstantin Kueffner, and Kaushik Mallik. 2023. Runtime Monitoring of Dynamic Fairness Properties. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23).* Association for Computing Machinery, New York, NY, USA, 604–614. https://doi.org/10.1145/3593013.3594028

[81] Marius Hobbhahn and Tamay Besiroglu. 2022. Trends in GPU Price-Performance. https://epochai.org/blog/trends-in-gpu-price-performance

[82] Marius Hobbhahn, Lennart Heim, and Gökçe Aydos. 2023. Trends in machine learning hardware. https://epochai.org/blog/trends-in-machine-learning-hardware

[83] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training Compute-Optimal Large Language Models. https://doi.org/10.48550/arXiv.2203.15556 arXiv:2203.15556 [cs].

[84] Bengt Holmstrom and Paul Milgrom. 1991. Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, & Organization* 7 (1991), 24–52. https://www.jstor.org/stable/764957 Publisher: Oxford University Press.

[85] Bengt Holmström. 1979. Moral Hazard and Observability. *The Bell Journal of Economics* 10, 1 (1979), 74–91. https://doi.org/10.2307/3003320 Publisher: [RAND Corporation, Wiley].

[86] Michael Horowitz and Paul Scharre. 2021. *AI and International Stability: Risks and Confidence-Building Measures.* Technical Report. Centrer for a New American Security. https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures

[87] The White House. 2023. *National Cybersecurity Strategy.* Technical Report.

[88] Saffron Huang and Divya Siddarth. 2023. Generative AI and the Digital Commons. https://cip.org/research/generative-ai-digital-commons

[89] ISO. 2022. ISO/IEC 27001. https://www.iso.org/standard/27001

[90] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23).* Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3544548.3581196

[91] Seyyed Ahmad Javadi, Richard Cloete, Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. 2020. Monitoring Misuse for Accountable 'Artificial Intelligence as a Service'. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20).* Association for Computing Machinery, New York, NY, USA, 300–306. https://doi.org/10.1145/3375627.3375873

[92] Seyyed Ahmad Javadi, Chris Norval, Richard Cloete, and Jatinder Singh. 2021. Monitoring AI Services for Misuse. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21).* Association for Computing Machinery, New York, NY, USA, 597–607. https://doi.org/10.1145/3461702.3462566

[93] Michael C. Jensen and William H. Meckling. 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3, 4 (Oct. 1976), 305–360. https://doi.org/10.1016/0304-405X(76)90026-X

[94] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. https://doi.org/10.48550/arXiv.2001.08361 arXiv:2001.08361 [cs, stat].

[95] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic Recourse: From Counterfactual Explanations to Interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21).* Association for Computing Machinery, New York, NY, USA, 353–362. https://doi.org/10.1145/3442188.3445899 event-place: Virtual Event, Canada.

[96] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T. Hancock. 2023. Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 116:1–116:29. https://doi.org/10.1145/3579592

[97] Zachary Kenton, Ramana Kumar, Sebastian Farquhar, Jonathan Richens, Matt MacDermott, and Tom Everitt. 2023. Discovering agents. *Artificial Intelligence* 322 (Sept. 2023), 103963. https://doi.org/10.1016/j.artint.2023.103963

[98] Megan Kinniment, Lucas Jun Koba Sato, Haoxing Du, Brian Goodrich, Max Hasin, Lawrence Chan, Luke Harold Miles, Tao R. Lin, Hjalmar Wijk, Joel Burget, Aaron Ho, Elizabeth Barnes, and Paul Christiano. 2023. Evaluating Language-Model Agents on Realistic Autonomous Tasks. https://evals.alignment.org/Evaluating_LMAs_Realistic_Tasks.pdf

[99] Yuta Kittaka, Susumu Sato, and Yusuke Zennyo. 2023. Self-preferencing by platforms: A literature review. *Japan and the World Economy* 66 (2023), 101191. https://doi.org/10.1016/j.japwor.2023.101191

[100] Leonie Koessler and Jonas Schuett. 2023. Risk assessment at AGI companies: A review of popular risk assessment techniques from other safety-critical industries. https://doi.org/10.48550/arXiv.2307.08823 arXiv:2307.08823 [cs].

[101] Anton Korinek and Megan Juelfs. 2022. Preparing for the (Non-Existent?) Future of Work. https://doi.org/10.3386/w30172

[102] Alexander Kott. 2018. Intelligent Autonomous Agents are Key to Cyber Defense of the Future Army Networks. *The Cyber Defense Review* 3, 3 (2018), 57–70. https://www.jstor.org/stable/26554997 Publisher: Army Cyber Institute.

[103] Jean-Jacques Laffont and David Martimort. 2002. *The Theory of Incentives: The Principal-Agent Model.* Princeton University Press. https://doi.org/10.2307/j.ctv7h0rwr

[104] Henry Lieberman. 1997. Autonomous interface agents. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems (CHI '97).* Association for Computing Machinery, New York, NY, USA, 67–74. https://doi.org/10.1145/258549.258592

[105] Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Lijie Wen, Irwin King, and Philip S. Yu. 2024. A Survey of Text Watermarking in the Era of Large Language Models. https://doi.org/10.48550/arXiv.2312.07913 arXiv:2312.07913 [cs].

[106] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. 2023. AgentBench: Evaluating LLMs as Agents. https://doi.org/10.48550/arXiv.2308.03688 arXiv:2308.03688 [cs].

[107] Brian Melley. 2024. Judges in England and Wales Given Cautious Approval to Use AI. *TIME* (Jan. 2024). https://time.com/6553030/ai-legal-opinions-england-wales/

[108] Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. GAIA: a benchmark for General AI Assistants. https://arxiv.org/abs/2311.12983v1

[109] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (Feb. 2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[110] Smitha Milli, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao, and Anca D. Dragan. 2023. Engagement, User Satisfaction, and the Amplification of Divisive Content on Social Media. https://doi.org/10.48550/arXiv.2305.16941 arXiv:2305.16941 [cs].

[111] Matti Minkkinen, Joakim Laine, and Matti Mäntymäki. 2022. Continuous Auditing of Artificial Intelligence: a Conceptualization and Assessment of Tools and Frameworks. *Digital Society* 1, 3 (Oct. 2022), 21. https://doi.org/10.1007/s44206-022-00022-2

[112] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* 220–229. https://doi.org/10.1145/3287560.3287596 arXiv:1810.03993 [cs].

[113] Meredith Ringel Morris, Jascha Sohl-dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. 2023. Levels of AGI: Operationalizing Progress on the Path to AGI. https://doi.org/10.48550/arXiv.2311.02462 arXiv:2311.02462 [cs].

[114] mrbullwinkle and eric urban. 2023. Azure OpenAI Service abuse monitoring - Azure OpenAI. https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/abuse-monitoring

[115] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2023. Auditing Large Language Models: A Three-Layered Approach. https://doi.org/10.2139/ssrn.4361607

[116] Silen Naihin, David Atkinson, Marc Green, Merwane Hamadi, Craig Swift, Douglas Schonholtz, Adam Tauman Kalai, and David Bau. 2023. Testing Language Model Agents Safely in the Wild. https://doi.org/10.48550/arXiv.2311.10538

arXiv:2311.10538 [cs].

[117] Richard Ngo, Lawrence Chan, and Sören Mindermann. 2022. The alignment problem from a deep learning perspective. https://arxiv.org/abs/2209.00626v5

[118] Helen Nissenbaum. 1996. Accountability in a computerized society. *Science and Engineering Ethics* 2, 1 (March 1996), 25–42. https://doi.org/10.1007/BF02639315

[119] Joe O'Brien, Shaun Ee, and Zoe Williams. 2023. Deployment Corrections: An incident response framework for frontier AI models. https://doi.org/10.48550/arXiv.2310.00328 arXiv:2310.00328 [cs].

[120] U.S. Department of Human and Health Services. 2022. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html Last Modified: 2023-02-22T10:17:21-0500.

[121] Division of Trading and Markets. 2008. *Guide to Broker-Dealer Registration.* Technical Report. U.S. Securities and Exchange Commission.

[122] OpenAI. 2023. ChatGPT plugins. https://openai.com/blog/chatgpt-plugins

[123] OpenAI. 2023. GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774 arXiv:2303.08774 [cs].

[124] OpenAI. 2023. Introducing GPTs. https://openai.com/blog/introducing-gpts

[125] OpenAI. 2024. Introducing the GPT Store. https://openai.com/blog/introducing-the-gpt-store

[126] Aviv Ovadya. 2023. 'Generative CI' through Collective Response Systems. https://doi.org/10.48550/arXiv.2302.00672 arXiv:2302.00672 [cs].

[127] Alexander Pan, Chan Jun Shern, Andy Zou, Nathaniel Li, Steven Basart, and others. 2023. Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark. https://doi.org/10.48550/arXiv.2304.03279 arXiv:2304.03279 [cs] Issue: arXiv:2304.03279.

[128] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–22. https://doi.org/10.1145/3586183.3606763

[129] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 469–481. https://doi.org/10.1145/3351095.3372828

[130] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The Fallacy of AI Functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 959–972. https://doi.org/10.1145/3531146.3533158

[131] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. 2022. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)*. Association for Computing Machinery, New York, NY, USA, 557–571. https://doi.org/10.1145/3514094.3534181

[132] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. 2022. A Generalist Agent. *Transactions on Machine Learning Research* (2022). https://openreview.net/forum?id=1ikK0kHjvj

[133] Toran Bruce Richards. 2023. Auto-GPT: An Autonomous GPT-4 Experiment. https://github.com/Significant-Gravitas/Auto-GPT original-date: 2023-03-16T09:21:07Z.

[134] Felix Richter. 2023. Infographic: Amazon Maintains Lead in the Cloud Market. https://www.statista.com/chart/18819/worldwide-market-share-of-leading-cloud-infrastructure-service-providers

[135] Shalaleh Rismani, Renee Shelby, Andrew Smart, Edgar Jatho, Joshua Kroll, AJung Moon, and Negar Rostamzadeh. 2023. From Plane Crashes to Algorithmic Harm: Applicability of Safety Engineering Frameworks for Responsible ML. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3544548.3581407

[136] Fabien Roger and Ryan Greenblatt. 2023. Preventing Language Models From Hiding Their Reasoning. https://doi.org/10.48550/arXiv.2310.18512 arXiv:2310.18512 [cs].

[137] Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. 2023. Identifying the Risks of LM Agents with an LM-Emulated Sandbox. https://doi.org/10.48550/arXiv.2309.15817 arXiv:2309.15817 [cs].

[138] Stuart J. Russell and Peter Norvig. 2021. *Artificial Intelligence: A Modern Approach* (4 ed.).

[139] Jonas Sandbrink, Hamish Hobbs, Jacob Swett, Allan Dafoe, and Anders Sandberg. 2022. Differential technology development: An innovation governance consideration for navigating technology risks. https://doi.org/10.2139/ssrn.4213670

[140] Jonas B. Sandbrink. 2023. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. http://arxiv.org/abs/2306.13952 arXiv:2306.13952 [cs].

[141] William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. https://doi.org/10.48550/arXiv.2206.05802 arXiv:2206.05802 [cs].

[142] Paul Scharre. 2021. Debunking the AI Arms Race Theory (Summer 2021). (2021). https://hdl.handle.net/2152/87035 Publisher: Texas National Security Review.

[143] Thomas C. Schelling. 1978. *Micromotives and Macrobehavior.* W. W. Norton & Company.

[144] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. https://doi.org/10.48550/arXiv.2302.04761 arXiv:2302.04761 [cs].

[145] Elizabeth Seger, Aviv Ovadya, Divya Siddarth, Ben Garfinkel, and Allan Dafoe. 2023. Democratising AI: Multiple Meanings, Goals, and Methods. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*. Association for Computing Machinery, New York, NY, USA, 715–722. https://doi.org/10.1145/3600211.3604693

[146] Lee Sharkey, Clíodhna Ní Ghuidhir, Dan Braun, Jérémy Scheurer, Mikita Balesni, Lucius Bushnaq, Charlotte Stix, and Marius Hobbhahn. 2023. A Causal Framework for AI Regulation and Auditing. https://static1.squarespace.com/static/6461e2a5c6399341bcfc84a5/t/654bc268049d687cecac24d8/1699463818729/auditing_framework_web.pdf

[147] Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, Katarina Slama, Lama Ahmad, Paul McMillan, Alex Beutel, Alexandre Passos, and David G. Robinson. 2023. Practices for Governing Agentic AI Systems.

[148] Rachel Sheffield and Catherine Francois. 2021. *Is Instagram Causing Poorer Mental Health Among Teen Girls? - Is Instagram Causing Poorer Mental Health Among Teen Girls? - United States Joint Economic Committee.* Technical Report. United States Congress Joint Economic Committee. https://www.jec.senate.gov/public/index.cfm/republicans/2021/12/is-instagram-causing-poorer-mental-health-among-teen-girls

[149] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*. Association for Computing Machinery, New York, NY, USA, 723–741. https://doi.org/10.1145/3600211.3604673

[150] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. https://doi.org/10.48550/arXiv.2303.17580 arXiv:2303.17580 [cs].

[151] Murtuza N. Shergadwala, Himabindu Lakkaraju, and Krishnaram Kenthapadi. 2022. A Human-Centric Perspective on Model Monitoring. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 10 (Oct. 2022), 173–183. https://doi.org/10.1609/hcomp.v10i1.21997

[152] 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. 2023. Model evaluation for extreme risks. https://arxiv.org/abs/2305.15324v2

[153] Alexander F Siegenfeld and Yaneer Bar-Yam. 2020. An introduction to complex systems science and its applications. *Complexity* 2020 (2020), 1–16. Publisher: Hindawi Limited.

[154] Nate Soares, Benja Fallenstein, Stuart Armstrong, and Eliezer Yudkowsky. 2015. Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence.*

[155] Emily H. Soice, Rafael Rocha, Kimberlee Cordova, Michael Specter, and Kevin M. Esvelt. 2023. Can large language models democratize access to dual-use biotechnology? https://doi.org/10.48550/arXiv.2306.03809 arXiv:2306.03809 [cs].

[156] Irene Solaiman. 2023. The Gradient of Generative AI Release: Methods and Considerations. https://doi.org/10.48550/arXiv.2302.04844 arXiv:2302.04844 [cs].

[157] Emily Sullivan and Philippe Verreault-Julien. 2022. From Explanation to Recommendation: Ethical Standards for Algorithmic Recourse. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. https://doi.org/10.1145/3514094.3534185

[158] Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. 2023. Cognitive Architectures for Language Agents. https://doi.org/10.48550/arXiv.2309.02427 arXiv:2309.02427 [cs].

[159] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement learning: An introduction* (second edition ed.). The MIT Press, Cambridge, Massachusetts. tex.lccn: Q325.6 .R45 2018.

[160] Adaptive Agent Team, Jakob Bauer, Kate Baumli, Satinder Baveja, Feryal Behbahani, Avishkar Bhoopchand, Nathalie Bradley-Schmieg, Michael Chang, Natalie Clay, Adrian Collister, Vibhavari Dasagi, Lucy Gonzalez, Karol Gregor,

Edward Hughes, Sheleem Kashem, Maria Loks-Thompson, Hannah Openshaw, Jack Parker-Holder, Shreya Pathak, Nicolas Perez-Nieves, Nemanja Rakicevic, Tim Rocktäschel, Yannick Schroecker, Jakub Sygnowski, Karl Tuyls, Sarah York, Alexander Zacherl, and Lei Zhang. 2023. Human-Timescale Adaptation in an Open-Ended Task Space. https://doi.org/10.48550/arXiv.2301.07608 arXiv:2301.07608 [cs].

[161] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Yujing Zhang, Ravi Addanki, Antoine Miech, Annie Louis, Laurent El Shafey, Denis Teplyashin, Geoff Brown, Elliot Catt, Nithya Attaluri, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaly Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, Hanzhao Lin, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo-yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yong Cheng, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, YaGuang Li, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Gamaleldin Elsayed, Ed Chi, Mahdis Mahdieh, Ian Tenney, Nan Hua, Ivan Petrychenko, Patrick Kane, Dylan Scandinaro, Rishub Jain, Jonathan Uesato, Romina Datta, Adam Sadovsky, Oskar Bunyan, Dominik Rabiej, Shimu Wu, John Zhang, Gautam Vasudevan, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Betty Chan, Pam G. Rabinovitch, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Sahitya Potluri, Jane Park, Elnaz Davoodi, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Chris Gorgolewski, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Paul Suganthan, Evan Palmer, Geoffrey Irving, Edward Loper, Manaal Faruqui, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Michael Fink, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marin Georgiev, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Minnie Lui, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Lam Nguyen Thiet, Daniel Andor, Pedro Valenzuela, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Sarmishta Velury, Sebastian Krause, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Tejasi Latkar, Mingyang Zhang, Quoc Le, Elena Allica Abellan, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Sid Lall, Ken Franko, Egor Filonov, Anna Bulanova, Rémi Leblond, Vikas Yadav, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Hao Zhou, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Jeremiah Liu, Mark Omernick, Colton Bishop, Chintu Kumar, Rachel Sterneck, Ryan Foley, Rohan Jain, Swaroop Mishra, Jiawei Xia, Taylor Bos, Geoffrey Cideron, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Petru Gurita, Hila Noga, Premal Shah, Daniel J. Mankowitz, Alex Polozov, Nate Kushman, Victoria Krakovna, Sasha Brown, Mohammad-Hossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Anhad Mohananey, Matthieu Geist, Sidharth Mudgal, Sertan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Quan Yuan, Sumit Bagri, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Aliaksei Severyn, Jonathan Lai, Kathy Wu, Heng-Tze Cheng, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Mark Geller, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Andrei Sozanschi, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Abhimanyu Goyal, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Sabaer Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Tao Zhu, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Dustin Tran, Yeqing Li, Nir Levine, Ariel Stolovich, Norbert Kalb, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Balaji Lakshminarayanan, Charlie Deck, Shyam Upadhyay, Hyo Lee, Mike Dusenberry, Zonglin Li, Xuezhi Wang, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem,

Summer Yue, Sho Arora, Eric Malmi, Daniil Mirylenka, Qijun Tan, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Steven Zheng, Francesco Pongetti, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Ragha Kotikalapudi, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Chenkai Kuang, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, T. J. Lu, Abe Ittycheriah, Prakash Shroff, Pei Sun, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Ishita Dasgupta, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Yuan Liu, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Ivo Penchev, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Adam Kurzrok, Lynette Webb, Sahil Dua, Dong Li, Preethi Lahoti, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Taylan Bilal, Evgenii Eltyshev, Daniel Balle, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Adams Yu, Christof Angermueller, Xiaowei Li, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, M. K. Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Kevin Brooks, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Komal Jalan, Dinghua Li, Ginger Perng, Blake Hechtman, Parker Schuh, Milad Nasr, Mia Chen, Kieran Milan, Vladimir Mikulik, Trevor Strohman, Juliana Franco, Tim Green, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2023. Gemini: A Family of Highly Capable Multimodal Models. https://doi.org/10.48550/arXiv.2312.11805 arXiv:2312.11805 [cs].

[162] Fabio Urbina, Filippa Lentzos, Cédric Invernizzi, and Sean Ekins. 2022. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence* 4, 3 (March 2022), 189–191. https://doi.org/10.1038/s42256-022-00465-9

[163] Karthik Valmeekam, Matthew Marquez, and Subbarao Kambhampati. 2023. Can Large Language Models Really Improve by Self-critiquing Their Own Plans? https://doi.org/10.48550/arXiv.2310.08118 arXiv:2310.08118 [cs].

[164] Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2023. Large Language Models Still Can't Plan (A Benchmark for LLMs on Planning and Reasoning about Change). https://doi.org/10.48550/arXiv.2206.10498 arXiv:2206.10498 [cs].

[165] Rory Van Loo. 2019. Regulatory Monitors: Policing Firms in the Compliance Era. *Columbia Law Review* 119, 2 (Jan. 2019), 369. https://scholarship.law.bu.edu/faculty_scholarship/265

[166] Suresh Venkatasubramanian and Mark Alfano. 2020. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3351095.3372876

[167] Pranshu Verma. 2023. The rise of AI fake news is creating a 'misinformation superspreader'. *Washington Post* (Dec. 2023). https://www.washingtonpost.com/technology/2023/12/17/ai-fake-news-misinformation/

[168] Pranshu Verma. 2023. They thought loved ones were calling for help. It was an AI scam. *Washington Post* (March 2023). https://www.washingtonpost.com/technology/2023/03/05/ai-voice-scam/

[169] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. 2023. A Survey on Large Language Model based Autonomous Agents. https://doi.org/10.48550/arXiv.2308.11432 arXiv:2308.11432 [cs] Issue: arXiv:2308.11432.

[170] Zihan Wang, Olivia Byrnes, Hu Wang, Ruoxi Sun, Congbo Ma, Huaming Chen, Qi Wu, and Minhui Xue. 2021. Data Hiding with Deep Learning: A Survey Unifying Digital Watermarking and Steganography. https://arxiv.org/abs/2107.09287v3

[171] Tom Warren. 2024. Microsoft's new Copilot Pro brings AI-powered Office features to the rest of us. *The Verge* (Jan. 2024). https://www.theverge.com/2024/1/15/24038711/microsoft-copilot-pro-office-ai-apps

[172] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. https://doi.org/10.48550/arXiv.2201.11903 arXiv:2201.11903 [cs].

[173] 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. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. https://doi.org/10.48550/arXiv.2310.11986 arXiv:2310.11986 [cs].

[174] David Gray Widder, Meredith Whittaker, and Sarah Myers West. 2023. Open (for Business): Big Tech, Concentrated Power, and the Political Economy of Open AI. https://ssrn.com/abstract=4543807

[175] Simon Willison. 2023. Prompt injection: What's the worst that can happen? https://simonwillison.net/2023/Apr/14/worst-that-can-happen/

[176] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework. _eprint: 2308.08155.

[177] Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. 2023. Language Agents with Reinforcement Learning for Strategic Play in the Werewolf Game. https://arxiv.org/abs/2310.18940v2

[178] Hanlin Zhang, Benjamin L. Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. 2023. Watermarks in the Sand: Impossibility of Strong Watermarking for Generative Models. https://doi.org/10.48550/arXiv.2311.04378 arXiv:2311.04378 [cs].

[179] Miri Zilka, Holli Sargeant, and Adrian Weller. 2022. Transparency, Governance and Regulation of Algorithmic Tools Deployed in the Criminal Justice System: a UK Case Study. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. https://doi.org/10.1145/3514094.3534200

[180] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and Transferable Adversarial Attacks on Aligned Language Models. https://arxiv.org/abs/2307.15043v2

[181] Remco Zwetsloot and Allan Dafoe. 2019. Thinking about risks from AI: Accidents, misuse and structure. *Lawfare. February* 11 (2019), 2019.

A POTENTIAL INFORMATION TO INCLUDE ON AN AGENT CARD

A.1 The Underlying System

- Evaluations of the system's degree of agency [106, 108, 137];
- Evaluations of the system's generality: its ability to accomplish a broad array of tasks to some specified performance threshold [113];
- Red flags, such as previous incidents or results of previous dangerous capabilities and alignment evaluations [152];
- Dependencies of the agent, such as with an ecosystem graph [24] (e.g., whether the agent is a fine-tuned variant of another model).

A.2 The Specific Instance of the Agent

- How the agent instance was created (e.g., by its user directly or by another agent?);
- The agent's goal, including both what the user specifies and what the system appears to be achieving [69];
- Any tools or services that the agent can access (e.g., spinning up another agent through an API call, physical manipulation of robotics);
- The agent's permissions (e.g., whether it has sudo access in a terminal);
- Details about the persistence of the agent, such as whether the agent has a set lifetime;
- The sector of the intended deployment environment (e.g., finance);
- The number of people the system can directly impact and the severity of such impact;
- The degree and ease of human oversight over the agent.

A.3 The Actors Involved in the Creation and Operation of the Agent

- The user;
- The compute provider;
- Developers of the underlying system, including developers of any scaffolding or component foundation models;
- Any humans with whom the agent has interacted;
- Tools and services providers that the agent can use.