Penn Carey Law
UNIVERSITY of PENNSYLVANIA

Public Law and Legal Theory Research Paper Series
Research Paper No. 25-04

# Leashes, Not Guardrails: A Management-Based Approach to Artificial Intelligence Risk Regulation

**Cary Coglianese**
*UNIVERSITY OF PENNSYLVANIA CAREY LAW SCHOOL*

**Colton R. Crum**
*UNIVERSITY OF NOTRE DAME*

# Leashes, Not Guardrails:

## A Management-Based Approach to Artificial Intelligence Risk Regulation

*Cary Coglianese[†] and Colton R. Crum[‡]*

Abstract. *Calls to regulate artificial intelligence (AI) have sought to establish "guardrails" to protect the public against AI going awry. Although physical guardrails can lower risks on roadways by serving as fixed, immovable protective barriers, the regulatory equivalent in the digital age of AI is unrealistic and even unwise. AI is too heterogeneous and dynamic to circumscribe fixed paths along which it must operate—and, in any event, the benefits of the technology proceeding along novel pathways would be limited if rigid, prescriptive regulatory barriers were imposed. But this does not mean that AI should be left unregulated, as the harms from irresponsible and ill-managed development and use of AI can be serious. Instead of*

[†] Edward B. Shils Professor of Law and Political Science and Director of the Penn Program on Regulation, University of Pennsylvania.

[‡] Ph.D. Candidate in Computer Science & Engineering, University of Notre Dame.

We want to thank the two anonymous reviewers and the journal's editors for their invaluable feedback on an earlier draft of this article. We also gratefully acknowledge exceedingly helpful comments provided by Richard Berk, Chris Callison-Burch, and Kevin Werbach. We would also like to express appreciation to Andrew Coopersmith and Mallika Kulkarni for the excellent assistance they contributed during our preparation of this article. The views expressed here should not necessarily be attributed to anyone else, and we alone are responsible for any errors or omissions. This paper has been accepted for publication in the journal *Risk Analysis*.

*"guardrails," though, policymakers should impose "leashes." Regulatory leashes imposed on digital technologies are flexible and adaptable—just as physical leashes used when walking a dog through a neighborhood allow for a range of movement and exploration. But just as a physical leash only protects others when a human retains a firm grip on the handle, the kind of leashes that should be deployed for AI will also demand human oversight. In the regulatory context, a flexible regulatory strategy known in other contexts as management-based regulation will be an appropriate model for AI risk governance. In this article, we explain why regulating AI by management-based regulation—a "leash" approach—will work better than a prescriptive or "guardrail" regulatory approach. We discuss how some early regulatory efforts include management-based elements. We also elucidate some of the questions that lie ahead in implementing a management-based approach to AI risk regulation. Our aim is to facilitate future research and decision-making that can improve the efficacy of AI regulation by leashes, not guardrails.*

Many policy discussions surrounding artificial intelligence (AI) commonly refer to a need for regulation to provide necessary "guardrails" that will protect the public from the risks of AI technology. Given the intense public attention to the type of AI represented by large language models (LLMs)—such as those that empower OpenAI's ChatGPT (OpenAI et al., 2023), Meta's Llama (Touvron et al., 2023), and Google's Gemini (Gemini Team et al., 2023)—the use of the term guardrail has broadened, especially in the realm of governmental policy, to encompass regulatory controls that seek to provide ex-ante safety protection from the risks posed by AI. This term has proliferated into ordinary usage when discussing virtually any proposal for AI

governance. Policymakers argue that "we need *guardrails*" (Warren, 2024)—and specifically "government-imposed *guardrails*" (Schumer, 2023)—and that "without significant *guardrails* in place, these algorithms…are going to accelerate the problems that we've got" (Warren, 2024). Policy analysts and advocates similarly call for stringent controls on AI—or even, at times, moratoria on AI research or use (Future of Life Institute, 2023; Claypool & Hunt, 2023)—by emphasizing the need for new regulatory "guardrails" (Working Group on Artificial Intelligence, 2024; Hewitt, 2023; Casovan & Shankar, 2023). Likewise with academic researchers who have used the term "guardrails" in other regulatory contexts (Farber, 2022; Livermore & Revesz, 2020) but have also done so specifically to refer both to technical AI best practices (Shamsujjoha et al., 2024) and regulatory controls placed on the development and use of AI tools (Eom, Newman, Brossard & Scheufele 2024; Gasser & Mayer-Schönberger, 2024).

Despite seemingly universal acceptance of the term, fixed regulatory guardrails are ill-fitted as the principal strategies for overseeing such a dynamic, heterogenous technology such as AI. On roadways, guardrails establish boundaries, clearly demarcating a course of travel and ensuring that travelers remain on the course deemed safe—sometimes literally keeping them and their vehicles from going over a cliff. In other words, guardrails lower risk by keeping behavior—namely, the control of vehicles—on a predetermined path. The language of "guardrails," when used to refer to regulation of technology risks in general, and with AI in particular, implies something similar: a fixed set of rules that remain in place and keep the technology on an acceptable course that ensures that members of the public remain safe and that society does not go over a metaphorical cliff. As Rouse (2023) notes, "AI guardrails are a lot like

highway guardrails—they are both created to keep people safe and guide positive outcomes."[1]

The regulatory design most akin to a physical guardrail is a fixed bright-line rule that ensures that regulated behavior, including the development of risky technologies, remains on a socially acceptable course. When rules function as guardrails, they spell out approved or prohibited courses of action or specify mandatory levels of performance for technologies and their development.[2] This is most evident for rules that simply ban all use of a specific technology, as Italian data protection regulators did in initially prohibiting the use of ChatGPT within their country (Satariano, 2023).[3] It is also exemplified in proposals for bans on certain kinds of AI uses, such as in processing employment applications, or the use of certain kinds of AI-assisted tools altogether, such as facial recognition.[4] Guardrails also tell developers or users exactly how their technologies must be developed or how they can be used. The European Union's AI Act,

---

[1] A report issued by McKinsey & Company (2024) begins by noting: "You know about guardrails on the highway: barriers along the edge of the road that protect vehicles from veering off course and into danger. With the advent of generative AI (gen AI), the concept of guardrails also applies to systems designed to ensure that a company's AI tools, especially large language models (LLMs), work in alignment with organizational standards, policies, and values."

[2] The language used to describe different types of regulatory designs varies widely. Following a framework adopted by the National Academies of Sciences, Engineering, and Medicine (NASEM), what we consider to be "guardrails" would be regulatory designs classified as "micro-means" or "micro-ends" (NASEM, 2018).

[3] Of course, instead of serving just as "guardrail," a complete ban on AI technology might be better analogized to a roadblock or a "do not enter" sign.

[4] Article 5 of the European Union's Artificial Intelligence Act contains a host of "prohibited AI practices" (EU AI Act, 2024, Art. 5).

for example, contains a provision imposing a mandatory standard on the "statistical properties" for the data that developers of certain AI tools may use in training their models (EU AI Act, 2024, Art. 10). Rules such as these, captured under the metaphor of "guardrails," generally allow for some technological momentum, but only within the predetermined confines of predetermined standards for the design or operation of technologies or of specific outcomes to be avoided.

The problem with a traditional guardrail approach to conceptualizing AI regulation is that, with such a dynamic, variable technology, the path of any metaphorical roadway, alongside which regulatory guardrails might be installed, cannot easily be determined in advance. Putting fixed regulatory guardrails in the form of prescriptive standards requires a regulator to know exactly the appropriate actions and outcomes to incorporate into regulations. Putting in place a strict ban on the use of all racial variables by AI developers, for example, might seem laudable but will not prove effective in eliminating unjust discrimination from AI tools if AI models still "learn" to discriminate from other variables that may correlate with race, such as zip codes (Kearns & Roth, 2020). Such blunt regulatory approaches can prove counterproductive if they prevent AI developers from identifying and addressing ways that their models may be inadvertently discriminatory or risky. Regulators not only can find themselves at an informational disadvantage vis-à-vis AI developers (Coglianese, Zeckhauser & Parson, 2004; Stephenson, 2011), but AI technology is highly heterogeneous such that it is moving along more roadways than even the most well-informed regulators could demarcate in advance (Marchant, 2011). Moreover, mapping out these roadways, and fixing rigid railings between permissible and impermissible algorithmic designs or permissible or impermissible outcomes, can unduly deny society the benefits of innovation from AI technologies and their applications. If society is to gain from innovation, innovators need to be able to blaze new paths and explore new roadways,

free from rigid regulatory barriers understood as guardrails.

In this article, we explain why "guardrails" are generally the wrong way to think about AI risk regulation. Instead, we propose that a better conception of regulation in this realm is for one that calls for AI firms to put "leashes" on their design and use of AI tools. The concept of a regulatory "leash"—that is, a law or regulation that is flexible but mandates a vigilant human oversight of AI—will better account for AI's mounting heterogeneity and rapid dynamism in both design and use. Conceiving of regulation as leash-based not only recognizes the virtue of flexibility but also the need for ongoing human oversight—much like pet owners walking their dogs are expected to keep a hand on the other end of their leashes. We argue that AI risk governance as a "leash" is best associated with the form of regulation known as management-based regulation (Coglianese & Lazer, 2003)—or sometimes called macro-means regulation (NASEM 2018). Management-based regulation of AI holds several key advantages over prescriptive "guardrails" in that it better responds to AI's novel uses and problems and better allows for technological exploration, discovery, and change. Nevertheless, management systems required by regulation still provide a tethered structure that, like a leash, can help prevent AI from "running away." Management-based regulation emphasizes the need to maintain human oversight across all stages of the AI pipeline (training, validation, and testing stages), and it reinforces the need for vigilance and continuous human oversight and responsibility.

## 1. The Misfit Between Prescriptive "Guardrails" and AI Risks

Despite pressing societal concerns for AI's safe use, the widespread use of the metaphor of "guardrails" fails to fit with AI's characteristics and risks. Regulation conceived as "guardrails"

implies the imposition of well-defined, pre-determined limits that have demarcated areas where AI can travel and where it cannot. But such an approach fails to accommodate AI's tremendous versatility in conquering new terrain—where humans are otherwise unable to navigate—which subsequently creates novel and unique paths that can be positive for society (Coglianese, 2023). In addition to prohibiting any new, "off-road" discovery, a guardrails approach to regulation associates rules with immovable structures that lack the flexibility required to protect the public from a dynamic technology. It also implies an emphasis on the deployment of fixed safeguards in the use of AI tools rather than ongoing precautions taken throughout all phases of AI training, development, *and* deployment. In addition, on roadways, once guardrails are put in place, they provide protection without any further human involvement. The use of "guardrails" to describe regulatory approaches to AI risks suggests a lack of need for continual human involvement once regulatory standards have been put in place. But in reality, ongoing vigilance by AI firms (i.e., their managers and engineers) is needed, which is why mandatory "leashes" are needed rather than "guardrails."

One of the greatest challenges associated with the regulation of AI stems from the heterogeneity of this technology and the risks associated with its use (Coglianese 2023). AI is, in fact, a suite of different technologies, more than a single technology per se (Davenport & High, 2024). Some AI tools are built for specific ultimate purposes, such as to detect skin cancer or make movie recommendations. Others, such as the well-known LLMs, are generative or "general purpose" AI tools that can be used for many different purposes, including research, coding, drafting, and synthesizing text (Zao-Sanders, 2024). These latter tools are sometimes referred to as foundation models because they can be adapted by users or integrated into larger automated systems designed for more specific (but still varied) purposes. AI is also being built

into digital agents that can solve problems, generate reasons, and perform other tasks on behalf of humans (Acharya et al., 2025; Shavit, 2023).

Considering all these various tools together, AI's heterogeneity stems from more than just their different purposes but also from the varied internal properties associated with algorithmic and training of neural networks, which serve as the bedrock of contemporary AI. Neural networks consist of millions (or billions) of neurons that process input signals and rely on their outputs to other subsequent "layers" of neurons. Nonlinearities introduced within these layers allow the neurons collectively to learn powerful representations of high-dimensional data such as images, video, and audio. Nuances within the ordering of neurons, their configurations, and nonlinear functions (otherwise known as activation functions) ultimately determine an eclectic assortment of neural network architectures suited for various potential tasks and uses. In addition to heterogeneity across AI models themselves, heterogeneity is introduced at the training of these models or networks, which involves tuning each individual neuron with respect to learning the overall task, represented by the "loss" (otherwise known as the "objective" or "cost") function. Tuning the neurons during training involves making an assortment of changes to algorithmic configurations within a given solution space, any of which can contribute to wild fluctuations in model outcomes—a heterogeneity or variability in performance. Furthermore, the training of an AI requires an enormous amount of data representative of what the model will encounter during the evaluation stage. Sources of data and the methods used to prepare them, such as file types, shuffling, and training splits, may have a substantial impact on an AI tool's outputs downstream.

Naturally, this level of technical heterogeneity complicates regulatory efforts. Regulation operates by making generalizations (Schauer & Zeckhauser, 2007), which assumes a relatively stable class of regulated entities or objects that produce an identified suite of problems that have

reasonably well-understood causes. This allows regulators to impose standards for how to design risky products or what specific safety tests they need to pass. With AI, however, the "products" are technically diverse, in part because they are applied in a wide range of settings for an exceedingly broad range of uses. They are also changing over time—often quite rapidly—whether through the reconfiguration of models or the processing of new data. Moreover, the regulatory problems that AI tools can create can be highly varied and can stem from a variety of different sources, including hardware, software, data, and algorithm training—or any combination of these or other factors. Once these multiple dimensions of heterogeneity in the underlying risk profile of AI are recognized, it becomes clearer how unrealistic it is to think that AI overall can be regulated simply by putting in place a fixed set of "guardrails" in the form of one-size-fits-all design standards or even uniform performance tests.[5]

AI exhibits a vast heterogeneity in uses: social media, autonomous vehicles, chatbots, precision medicine, fintech investment advisors, and much more. The latest foundation models that support generative AI can be deployed for any number of uses, limited mainly by the imagination of the individuals who prompt these models to perform tasks (Zao-Sanders, 2024).

[5] This is not to deny that specific applications of AI tools could not be expected to meet performance tests applicable to specific purposes or uses when AI tools are designed to support narrow, well-defined tasks or are embedded within a larger technological system. It would presumably still be reasonable, for example, to expect an AI-assisted braking system in an automobile to meet a standard test for the performance of any braking system (even a non-AI one). The application of performance tests to generative AI or foundation models, however, will be more difficult owing to the inherent heterogeneity or variability in the uses of these forms of AI (Coglianese & Crum, 2025).

AI's ability to support a wide variety of uses gives rise to numerous benefits for society, often in circumstances where humans are lacking due to their own shortcomings, such as cognitive biases or decision fatigue. When AI is used for cancer screening, for example, it can find evidence of tumors that even well-trained radiologists can miss (Satariano & Metz, 2023). AI tools have also been shown to identify cases of asymptomatic COVID with considerable accuracy, even without any medical testing (Bastani et al., 2021; Castro & Mclaughlin, 2019). AI tools can and are increasingly deployed in numerous applications and exceed human capabilities in the accuracy and speed of making predictions or completing other tasks.

But this is not to deny that AI can also produce a variety of distinct harms. Indeed, taken as a class, the harms posed by AI can be highly varied, if for no other reason than that AI can support such a wide variety of uses. Consider just three examples of AI harms from three distinct uses: autonomous vehicles, social media, and classification software.

*AI in Autonomous Vehicles: Collison Risks*. Self-driving cars are made possible due to advanced sensing capabilities and AI computing. But on a dark Arizona night in 2018, the AI in an autonomous vehicle appears to have contributed to the death of a woman crossing the street—an accident that likely might not have occurred if the vehicle had not been equipped with autonomous driving technology (NTSB, 2019). The AI detected the woman 5.6 seconds before the crash but failed to determine whether she was a bicyclist, pedestrian, or unknown object—ultimately unsure if their paths would eventually collide (NTSB, 2019). Overseeing the AI system was a human "driver," tasked with monitoring the system and taking command if the system failed. But while the AI was deciphering what exactly lurked ahead in the seconds before impact, the human was distracted, allegedly watching TV on her phone when the deadly collision occurred (Riess & Sottile, 2023). Other accidents associated with autonomous vehicle

technology have raised questions about whether this application of AI technology can sometimes contribute to accidents that would not have occurred had the relevant vehicles needed to rely solely on human drivers (NTSB, 2020; NTSB 2018).[6]

*AI in Social Media: Suicide Risks*. The proliferation of social media platforms that have attracted users around the world are powered by sophisticated algorithms, many of them that would now be classified as AI. In October 2023, a London coroner blamed social media as a contributing factor in the suicide of a 14-year-old girl. In the months prior to taking her life in November 2017, the girl had apparently viewed algorithm-recommended content about suicide and self-harm provided through several social media platforms, including Instagram and Pinterest. A seasoned child psychologist who viewed the content to prepare expert testimony delivered in the coroner's proceeding said the material "was so 'disturbing' and 'distressing' that it caused him to lose sleep for weeks" (Satariano, 2022). A senior Pinterest executive reportedly said the app was "not safe" when the girl had used it (Milmo, 2022). Others suggested that the graphic content that had been displayed was inconsistent with the social media companies' own content policies (Satariano, 2022). Even though it can be difficult, of course, to determine a precise causal relationship in any individual case, concern over a systemic propensity for social media to accentuate risks of self-harm and other mental health problems is endemic. In 2024, the

[6] This is not to say that these or other automobile accidents would not have occurred anyway, even in the absence of the AI tool. A key question for policymakers is whether AI technologies used in automobile transport perform better in terms of risk reduction than the status quo, which relies on admittedly flawed human operators. For an industry study indicating that AI can lower the risks from automobile accidents, see Di Lillo et al., 2024.

U.S. Surgeon General called for mandating a “warning label on social media platforms” that would indicate “that social media is associated with significant mental health harms for adolescents” (Murthy, 2024).

*AI in Classification Software: Risks of Bias and Discrimination.* One of the broadest categories of tasks that AI performs involves sorting and classifying, such as of words, images, or videos. As much as AI tools can classify with considerable speed and accuracy, they can also make mistakes—ones that humans would not make and ones that can be quite serious. AI used to sort and categorize images in Google Photos, for example, was once reportedly shown to classify Black people as “gorillas” (Metz, 2021). An experimental hiring algorithm developed by Amazon was abandoned by the company after it became clear it was disproportionately recommending the hiring of male job applicants, the bias stemming from having trained the algorithm on historical data that contained an extreme lopsidedness of male applicants (Oppenheim, 2018). Similar concerns about bias have been expressed with other AI tools (Crawford, 2016). Early versions of ChatGPT apparently made associations with Muslims or people from countries such as Syria or Yemen in response to prompts related to terrorism (Biddle, 2022). Microsoft’s overhaul of Bing featured a dominant chat feature powered by Open AI. The generative algorithm was easily steered off track with 15 or more questions (Microsoft, 2023), making it turn to dark language that promoted sexism, racism, and even blackmailing users (Perrigo, 2023).

*Extensions and Implications*. These three examples—in vehicles, social media, and classification software—highlight just some of the varied harms that can arise from different applications of AI. Many other problems have been associated with other AI uses (Bengio et al., 2025). Indeed, researchers at Massachusetts Institute of Technology have created a dynamic

repository of over 1,000 risks associated with AI (Slattery et al., 2024).[7]

Once one considers the full range of uses of AI that are rapidly being introduced across virtually every sector of the economy and continuously being adapted (Coglianese, 2023) and then takes into account the broad range of associated risks from these uses, it should become evident that any regulatory equivalent of fixed roadway guardrails is unrealistic. The "roads" that AI travels down are too numerous and, with generative AI especially, they are constantly changing. As a result, the harms associated with AI can be highly variable and dynamic. Specifying precise algorithmic designs or brute force engineering or programming filtering would tax any regulator's abilities. Even designating well-specified performance outputs will often be difficult, as the outputs of concern can be too numerous, hard to define, or practically impossible for any outside regulator to monitor (Coglianese & Crum, 2025).

Moreover, the fundamental prowess of AI stems from the fact that engineers are alleviated from prescriptive programming, which is fragile, weak, and less accurate in real-world practice. Prescriptive regulatory guardrails are, in an important sense, inherently a reversion from the core innovative power of AI, in purely technical terms. Fixed guardrails are akin to hard coding, whereas a leash approach to regulation is more akin to AI itself, as it promotes "living" governance ecosystems that are attentive and adaptive to novel problems. Ultimately, the better way to conceive of risk governance of AI technology is through "leashes" not "guardrails"—that is, through regulation that is both flexible yet strong when needed and that keeps humans engaged in oversight.

[7] The repository can be accessed online at: https://airisk.mit.edu/.

## 2. Leashes: A Management-Based Approach to AI Regulation

Management-based regulation induces firms to keep AI technologies on a leash. The management-based model not only fits theoretically with the heterogeneous nature of AI technologies and their varied and dynamic uses and problems, but it is also emerging as the principal approach in new laws and directives around the world. If management-based regulation does become the dominant mode of AI risk regulation, this makes all the more important the need to be clear that it is providing protection in the form of a leash—not a guardrail. Management-based regulation does not function in the same way as a solid, immovable guardrail, which is a risk management response that implies that human overseers are reprieved of their responsibilities once a guardrail is put into place. On the contrary, just like a leash affixed at one end around a large dog in a neighborhood protects children because a human is gripping its other end, AI risk management depends on ongoing human vigilance.[8] Management-based regulation seeks to ensure that the firms that develop and use AI in ways that pose risks to consumers and society will exercise that continuous oversight—keeping a firm grip on the leash, even as the leash's flexibility affords room for exploration and innovation.

### *A. Management-Based AI Regulation*

Management-based regulatory strategies can be found in other domains whenever a blanket, "one size fits all" legal solution does not fit, such as toxic pollution prevention (Bennear 2006), food safety (Coglianese & Lazer, 2003), accident prevention (Kleindorfer, 2006), and chemical

[8] For an illuminating treatment of similarities between AI and animal intelligence, see Darling (2021).

facility security (Coglianese & Starobin, 2020). Because of the extreme diversity of products such as chemicals and foods, the facilities that produce or process them are highly heterogeneous in how they operate—much too variable for regulators to be assured that they can provide a uniform set of rules that will be sufficient to mitigate the specific risks that each facility poses. Under a management-based approach, the regulated entity is required to develop an internal plan to identify and monitor risks, produce and implement protective procedures, and document changes to address those risks (Coglianese & Lazer, 2003). These internal plans may be subject to auditing and require continual updates to prevent risk management decay. It eliminates regulators from having to have the same knowledge about AI for every firm and does not set unrealistic barriers to developing advances in technology. Management-based approaches allow for scalable reporting methods, detailed disclosure, and product testing—something that would be difficult, if not impossible, for an outside regulator to specify on a uniform basis in a prescriptive regulation. In this way, management-based approaches are more flexible—regulatory leashes, not guardrails.

Internal planning required under management-based AI regulation would involve standardized documentation of the AI systems' current uses, a description of the AI's goals, and a thorough identification of possible AI failures and harms. Management-based regulation would require firms and their engineers to follow internal practices that accord with a Deming Cycle of "plan-do-check-act."[9] In the AI context, this is sometimes called an AI impact assessment (Coglianese & Shaikh, 2025). In addition, firms could be required to identify and use

[9] These required management practices are considered a form of "macro-means" regulation—another term used for management-based regulation (NASEM 2018).

management practices such as “red-teaming” or adversarial testing during an algorithm’s training process.

To help ensure responsible management of AI, firms that develop or deploy these tools would be required to develop their own internal routines and risk management procedures that their data scientists and engineers would need to follow to identify and reduce AI risks. In addition, firms could be required to allow regulators to look at this documentation as well as all the relevant data and training information assembled as part of a company’s ongoing risk management system. A management-based AI regulation could also provide for external auditing of firms’ management plans and procedures. Perhaps in the case of the Arizona pedestrian, a requirement for input from a regulator or third-party auditor providing a “peer” review of the company’s plan would have suggested additional night-time testing that could have averted the crash (NTSB, 2019).

In the case of social media, management-based regulation could call for these firms to anticipate the range of possible harms from the use of their platforms, including creating or exacerbating mental health problems, and then to create processes to monitor and address these problems. Such processes might include the creation of user feedback systems or other reporting channels to keep track of harms that emerge even after AI tools are in use so that methods can be established to respond to them. Similar reporting measures have been used as part of other management-based regulatory regimens. The Federal Aviation Administration (FAA), for example, has created an anonymous reporting system for anyone witnessing potentially dangerous behavior or impairment from pilots. The feasibility of requiring similar feedback strategies in the AI context is demonstrated by their voluntary adoption. The AI, Algorithmic, and Automation Incidents and Controversies (AIAAIC) is an independent private organization

that maintains a repository of AI-related incidents.[10] Moving toward a mandatory, centralized reporting system along these lines would allow both firms and regulators to track down various leads and consider whether these issues are related to other known technical problems.

It is important to note that, in practice, "leashes" and at least some "guardrails" need not be entirely mutually exclusive. The firms that are subject to a leash requirement would themselves presumably install certain internal guardrails. Moreover, in many regulatory contexts, management-based regulation is compatible with other, more prescriptive forms of regulation deployed on a targeted basis (NASEM 2018). In some instances, certain technical best practices or standards related to specific regulatory problems or causal pathways to problems can be so apparent that they can be reflected in more prescriptive "micro-means" or "micro-ends" standards (NASEM 2018). For example, in the case of generative AI, when a user prompts a large language model for a task that poses a well-specified risk (e.g., asking a chatbot to provide instructions for committing an act of self-harm), a regulation might conceivably require that the model be programmed to respond in a specified manner (e.g., provide information about a mental health crisis hotline). But such limited examples of best practices would not at all be incompatible with requiring a developer of the large language model to establish and implement an AI risk management system and rely on that system for the continual scanning of potential problems, including those covered by any specified micro-means or micro-ends regulations. A management-based leash will always be needed to address the overall varied and rapidly changing risks associated with different forms of AI.

[10] This repository can be found at: https://www.aiaaic.org/aiaaic-repository.

*B. The Emerging Management-Based Paradigm in AI Risk Regulation*

Despite widespread calls for regulatory "guardrails" as the solution to AI problems, early regulatory efforts in the European Union and the United States have recognized the vital role for a management-based approach to AI risk governance. In key respects, AI policies around the world have not been functioning like guardrails that are fixed, firm, and statically protective. Rather, important features of emerging law and policy have aimed to be protective by being more flexible and adaptable—and, importantly, by calling for ongoing human vigilance about AI risks. These emerging approaches—although often misleadingly characterized as "guardrails"—are best thought of as requiring the use of leashes.

*EU AI Act*. The European Union's AI Act amounts to the most comprehensive AI legislation to date. Although the Act contains elements that are guardrail-like, the use of AI in high-risk systems, where AI can cause serious harm or is deemed by the Act to violate fundamental human rights, is subject to a set of leash-like requirements for a risk management system that is both comprehensive and faithfully implemented. The Act stresses the need for a "continuous iterative process" of risk management throughout an AI tool's entire life cycle (EU AI Act, 2024, Art. 9, para. 2). Such a process must comprise the characterization of foreseeable risks, evaluation of those risks under intended purposes (and even foreseeable misuse), identification of new risks post-market launch (system monitoring), and adoption of risk management methods. Although most of these provisions are straightforward, the EU AI Act (2024, Art. 9) also stresses risks associated with the end-user, calling for "due consideration … to the technical knowledge,

experience, education, the training to be expected by the deployer, and the presumable context in which the system is intended to be used."

Executive Order 14,110. Within the United States, similar measures had been called for under Executive Order 14,110, signed by President Joseph Biden in October 2023. Although this order never directly imposed binding requirements on AI firms (and has now been rescinded by President Donald Trump), it called for management standards for the use of AI, especially by the federal government. The executive order specifically called for the use of risk management measures with respect to foundational AI models, including "any ongoing or planned activities related to training, developing, or producing [these] models, including the physical and cybersecurity protections taken to assure the integrity of that training process against sophisticated threats" (E.O. 14,100, 2023). In addition, in response to Executive Order 14,110, the Office of Management and Budget published a memorandum calling upon federal agencies that use AI to conduct impact assessments, specifying the AI's intended purpose, potential risks of using AI, and the quality and appropriateness of relevant data. Section D of this memorandum called for ongoing monitoring and periodic human review, including "testing for performance of the AI in a real-world context … at least annually, and after significant modifications to the AI or to the conditions or context in which the AI is used" (E.O. 14,110, 2023).

*Summary*. Notwithstanding frequent calls for regulatory action in the form of "guardrails," both the EU AI Act and Executive Order 14,110 exhibit core elements of a management-based or leashing approach to AI governance. Admittedly, as we have noted, Article 5 of the EU AI Act does impose, in guardrail fashion, some general prohibitions on certain uses related to AI, such as for "social scoring" by the government or the scraping of images for facial recognition tools. Yet in dealing generally with the plethora of permissible but still risky uses of AI, the EU

AI Act, as with the Biden-era executive order, represents a more flexible approach by calling for systematic management practices that can be adapted to different contexts, depending on AI's varying uses and risks.

## 3. Issues in Management-Based AI Risk Regulation

Management-based regulation in other settings has been shown to result in expected risk reductions, such as in reduced toxic air pollution in states that have adopted prevention planning laws (Bennear, 2007) and decreases in foodborne illnesses following the adoption of so-called HACCP food safety planning requirements (Minor & Parrett, 2017). Positive results from mandatory leashes, though, are never guaranteed. In any context, the adoption of management-based regulation raises at least three implementation issues: the threshold for regulatory application (*when should a leash be required*); the specific management measures required (*what constitutes an appropriate leash*); and, finally, the role for human oversight (*responsibility of the human at the end of the leash*).

### *A. Thresholds for Regulatory Application: When Should a Leash be Required?*

With respect to the first issue, just as physical leashes would be more important for walking a large German shepherd than a tiny Cavapoo, and more important when taking *any* breed of dog for a walk through a day care center's playground than through a wilderness area, AI leashing will be more appropriate when an AI tool can potentially cause significant harm to others. Causing potential harm to others depends on (a) the tool itself and how it functions, as well as

(b) the dangers associated with the surrounding environment within which it is deployed. One might ask, as an initial heuristic, whether a human performing a similar task would warrant some kind of risk regulation. If so, then the imposition of management-based requirements on the development and use of AI for the task should be at least presumptively justified.

The reverse scenario may be less clear. If a human performing the exact same task does not warrant the same imposed managerial regimen, policymakers might still ask whether the AI tool introduces any *new* risk that might merit the imposition of management-based regulation. Although this question will be largely context-specific, to consider whether an AI tool might introduce heightened risk, one might ask at least initially: "What's the worst that can happen *because* of the AI?" The answers will vary. For an AI tool that makes movie recommendations, the worst might be that customers terminate their streaming service. For chatbots, the worst might entail the AI tool spitting out racial slurs or threatening customers. And for AI-operated robotic pizza delivery cars, the worst could be accidents on roads near college campuses.

As with other forms of regulation, conventional forms of risk analysis and benefit-cost analysis would be appropriate tools to deploy to anticipate whether certain uses are more likely to be problematic such that the imposition of management-based requirements would be justified. Such analyses for AI should be attentive not only to the consequences of deploying an AI tool, but also to the risks associated with the entire machine-learning pipeline, including data acquisition or training. Requirements for leashes will be more likely to be justified, for example, for AI tools that derive from unethical practices, sensitive web-scraping, or training on unauthorized or copyrighted materials.

Furthermore, any assessment of AI risks should proceed with the status quo in mind. It will inevitably be the case, for example, that AI tools make mistakes. The question will be how

serious those mistakes are compared to those made by existing systems that depend on humans, with their own well-known limitations of perception and decision-making. This question may not always be easy to answer, but when management-based regulation is justified, it could demand that testing be conducted to find such an answer before deploying an AI tool.

*B. Specification of Required Management Measures: What Constitutes an Appropriate Leash?*

Just because a management-based regulation might be justified, it remains to be determined exactly what management measures ought to be required—such as internal risk analysis, auditing, red-teaming, data acquisition, personnel training, monitoring, documentation, or reporting. In other words, what constitutes an appropriate leash? How demanding should management requirements be? How clearly specified should they be demarcated in a regulation?

Often, the general parameters of different AI management-based regulations will follow a systemic plan-do-check-act model. This is the approach reflected, for example, in the National Institute of Standards and Technology's Risk Management Framework for AI (Slattery et al., 2024). But just as the physical leash for a large dog will be thicker and sturdier than that needed for a small dog, so too will AI leashes need to vary in their scope, stringency, and specificity.

Specification of required risk management measures for the development and use of AI tools will necessitate a refinement by the regulator of an overall problem statement, asking questions such as: What is the range of possible harms to users and the public from a given type of AI? What is the firm's overall goal for the AI tool, and are these objectives aligned with the regulator's objectives? Are these goals being integrated with care into the training of the AI, such as through a well-defined loss function? In other words, are suitable objectives

meaningfully conveyed to the AI tool by the firm during training?

With social media, for example, the regulator should consider whether the AI tool has an objective that simply seeks to *maximize* metrics such as watch time, engagement, or impressions. Or does the algorithm's objective seek to *minimize* relative exposure to offensive content? Or is the objective one that balances multiple values in ways that align with overall social welfare?

These kinds of questions can help indicate whether: (a) the firm that develops an AI tool has malevolent goals and the AI tool is adhering to those goals (think: a dog is trained to bite others); (b) there is a basic alignment between the firm's and regulator's goals, but a misalignment between these goals and the AI tool's objective (the dog is poorly or incorrectly trained); or (c) the AI tool is defiant of both the regulator's and firm's goals (the dog refuses to obey its handler's orders).

The strength of a needed leash will also be reflective of past performance. Any dog that has previously acted aggressively towards children should not be taken to a public playground without a very strong leash. Likewise, if a given training set, architecture, or training configuration has been explicitly known to regurgitate or leak sensitive information, then stronger leashes may be necessary. This may mean imposing requirements for more frequent monitoring of the AI tools, greater disclosure of testing results, or even approval of the AI firm's management plan and its operation after periodic regulatory reviews.

The strength of a regulatory leash should also be appropriate for the potential risks related to the AI tool's tasks. In other words, specific management measures should be compatible with the potential harms of the AI tool or the tool's functioning. General purpose or foundational models have broader functions compared with simpler AI models designed for well-specified tasks. Consequently, the required leash should reflect the broader range of tasks the tool is expected to

perform and their associated potential harms (Coglianese & Crum, forthcoming).

Challenges for the regulator arise when AI use is associated with ill-defined, hard-to-specify, or even hidden harms. The harms associated with AI may also be highly varied, even from the use of the same tool. But this is precisely where management-based regulation can be valuable—and often it may be the only practical option. It can help force AI firms to define more clearly what is ill-defined, scan for a broad range of possible harms, and then take internal managerial and technical steps to assess and manage risks to users and others.

### *C. The Role of Human Oversight in Leashing*

Within a firm's process of leashing its AI tools, as required by management-based regulation, humans can be engaged in various ways, such as through (a) auditing and information disclosure, (b) human training guidance, and (c) human-AI teaming. The regulator will need to decide how a firm's managers and engineers should oversee their AI tools and then how to specify that form of oversight in the text of any management-based rule. In other contexts, for example, management-based rules have contained provisions that mandate specific protocols for periodic auditing, documentation, and peer review.

In the domain of AI design and development, relevant research on how firmly humans should maintain a grip on their leash over AI tools has arisen in the context of AI research dedicated to model explainability, interpretability, and trustworthiness (Smith-Renner et al., 2020). This research aims to gather information about the models' internal operations and convey those findings in a human-explainable and interpretable format so that developers and end users can be informed of any potential threats (Dimanov et al., 2020). Early firm-level

attempts to systematize information collection and reporting have taken the form of "model cards" (Mitchell et al., 2019)—also called Datasheets (Gebru et al., 2021) or FactSheets (Richards et al., 2020)—which contain basic information about an AI tool, such as the model type and dataset used. Although disclosures of these model cards have, to date, been voluntary, a key policy question is whether and to what extent they should be required in their present or some modified form—and to whom testing results and other information about algorithms should be given: just the regulators or to the entire public. Some experts have questioned the adequacy of current disclosures, with reasonable concerns made about whether they give humans enough to be assured that AI companies have been responsibly managing their AI risks.

Another way to manage AI risks will be to integrate human expertise knowledge into the training of AI. Relevant, task-specific information collected from human experts can be injected into the model during training (Boyd, Tinsley, Bowyer & Czajka, 2023). Human-guided training can improve model performance and create more expert interpretable outputs (Linsey et al., 2018). An example of human-guided models can be found with AI tools used to screen for cancer. These AI tools work alongside radiologists to catch potential errors and aid in the diagnostic process. As noted previously, models can latch onto spurious features and produce their own errors. For example, a model used to predict anomalies in chest x-rays might use timestamps located in the corners of the scan to classify samples instead of features located in a patient's thoracic cavity. Although such a model has effectively *learned* the task, it has done so in a way misaligned with the radiologist's diagnostic methods to the potential detriment of the patient's health. Instead of allowing the model to learn only from the available data, adding to model training supplemental information provided by radiologists, such as their eye-tracking patterns, can help guide the model towards expert-salient features.

Of course, human-guided models are not without their downsides, and management-based regulation would not dictate that firms deploy these models as much as consider and explain whether and how they should be deployed. After all, obtaining human salient information is costly and its efficacy is not always guaranteed. In addition, there is no clear answer as to how many annotations are necessary to train human-guided models. Human-guided models are still fragile and prone to some of the same problematic issues found in conventionally trained models. Nonetheless, human-guided models seem especially promising candidates to be considered when: (a) the AI is being deployed in areas that impact human safety or rights (so-called high-risk domains); (b) human expertise offers distinctive insights; (c) model outputs must be aligned with human experts; or (d) natural shortcomings exist in obtaining sufficient training samples.

Relatedly, if AI could benefit from human oversight but human-guided training is impractical, then human-AI teaming could be considered as an element of a firm's AI risk management plan. The aim would be to leverage the benefits of humans and AI tools in conjunction with each other to solve a given task. Sources of communication between AI and humans through alerts, descriptions, and visualizations are under continual technical development and should be considered. Human-AI teaming will also necessitate continual human training, similar to training for commercial pilots when cockpit technology gets updated. Just as with other applications of management-based regulation, then, personnel training would almost certainly be a component of any required AI risk management plan.

In addition to defining the roles for humans within AI firms' management practices, the AI regulator will itself need to oversee firms and their oversight of their AI tools. It will not be enough simply to require that firms undertake responsible management practices. Without

ongoing regulatory oversight, compliance with management-based requirements can be prone to slippage or practices that simply "go through the motions," whereby organizations produce required management analyses and plans only in the most pro forma manner (NASEM, 2018). Auditing of some sort will thus be needed to ensure that firms take seriously their required management practices. This means that a role for human auditing by regulators or other third parties will need to be defined. The information to be collected and made available to auditors will also need to be specified. After all, algorithms cannot be successfully audited or examined if no information about the algorithm's training or its operation is disclosed to a human.

## 4. Conclusion

Discussion of AI safety regulation has to date centered around the establishment of "guardrails," implying that fixed, prescriptive rules can be established to protect the public from harm from AI. We have argued for a different way to conceptualize AI risk regulation, namely, as governance that relies on "leashes." Leashes are flexible. They permit AI tools to explore new domains without regulatory barriers getting in the way. Yet this flexibility comes along with active human oversight, too. The goal of AI risk management should not be simply to establish guardrails and let AI tools operate within a fixed space unsupervised—at least until significant harms occur. Rather, regulation should promote active human oversight and management of AI—keeping human overseers at the other end of a leash, ready and capable of steering AI away from danger as needed. Under a leashes approach—specifically, management-based regulation—AI governance seeks to promote ongoing human oversight through sound management practices that identify and respond to risks early or address them promptly when

they emerge. AI risks can be more appropriately assessed—and societies can better safeguard themselves from these risks—through attentive reliance on regulatory leashes, not by hoping that some kind of rigid guardrails can be established.

References


Acharya, D. B. et al. (2025). Agentic AI: Autonomous intelligence for complex goals–A comprehensive survey. *IEEE Access, 13*, 18912–18936. https://ieeexplore.ieee.org/document/10849561.

Bastani, H. et al. (2021). Efficient and targeted COVID-19 border testing via reinforcement learning. *Nature*, *599*, 108–113.

Bengio, Y. et al. (2025). *International AI safety report 2025* (DSIT research paper series number 2025/001). Department for Science, Innovation and Technology. https://www.gov.uk/government/publications/international-ai-safety-report-2025.

Bennear, L. S. (2007). Are management-based regulations effective? Evidence from state pollution prevention programs. *Journal of Policy Analysis and Management, 26*(2), 327–348.

——— (2006). Evaluating management-based regulation: A valuable tool in the regulatory toolbox? In C. Coglianese & J. Nash (Eds.), *Leveraging the Private Sector: Management-based Strategies for Improving Environmental Performance*. Routledge.

Biddle, S. (2022, December 8). The internet's new favorite AI proposes torturing Iranians and surveilling mosques. *The Intercept*. https://theintercept.com/2022/12/08/openai-chatgpt-ai-bias-ethics/.

Boyd, A., Tinsley, P., Bowyer, K., & Czajka, A. (2023). CYBORG: Blending human saliency

into the loss improves deep learning-based synthetic face detection. *IEEE/CVF Winter Conference on Applications of Computer Vision*, 6097–6106.

Casovan, A., & Shankar, V. (2023, April 4). Generative AI needs guardrails, not a pause. *Responsible Artificial Intelligence Institute*. https://www.responsible.ai/generative-ai-needs-guardrails-not-a-pause/.

Castro, D., & Mclaughlin, M. (2019, February 4). Ten ways the precautionary principle undermines progress in artificial intelligence. *Information Technology*. https://itif.org/publications/2019/02/04/ten-ways-precautionary-principle-undermines-progress-artificial-intelligence/.

Claypool, R., & Hunt, C. (2023, April 18). "Sorry in advance!" Rapid rush to deploy generative A.I. risks a wide array of automated harms. *Public Citizen*. https://www.citizen.org/article/sorry-in-advance-generative-ai-artificial-intellligence-chatgpt-report/.

Coglianese, C. (2023). Regulating machine learning: The challenge of heterogeneity. *TechReg Chronicle, (February 2023), 17-23*, https://scholarship.law.upenn.edu/faculty_scholarship/2921/.

Coglianese, C., & Crum, C. R. (forthcoming). On leashing (and unleashing) AI innovation.

——— (2025). Regulating multifunctionality. In Philipp Hacker, Andreas Engel, Sarah Hammer and Brent Mittelstadt (Eds.), *The Oxford Handbook on the Foundations and Regulation of Generative AI*. Oxford University Press.

Coglianese, C., & Lazer, D. (2003). Management-based regulation: Prescribing private management to achieve public goals. *Law & Society Review, 37, 691–730.*

Coglianese, C., & Shaikh, N. (2025). Management-based oversight of the automated state:

Emerging standards for AI impact assessment and auditing in the public sector. In E. Yaghmaei et al. (Eds.), *Global Perspectives on AI Impact Assessment*. Oxford University Press.

Coglianese, C., & Starobin, S. M. (2020). Management-based regulation. In K. R. Richards & J. Zeben (Eds.), *Policy Instruments in Environmental Law, 292-307*. Edward Elgar.

Coglianese, C., Zeckhauser, R., & Parson, E. (2004). Seeking truth for power: Informational strategy and regulatory policymaking. *Minnesota Law Review, 89, 277*.

Crawford, K. (2016, June 25). Artificial intelligence's white guy problem. *The New York Times*. https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html.

Darling, K. (2021) *The new breed: What our history with animals reveals about our future with robots*. New York: Henry Holt and Co.

Davenport, T. H., & High, P. (2024, December 13). How gen AI and analytical AI differ and when to use each. *Harvard Business Review*. https://hbr.org/2024/12/how-gen-ai-and-analytical-ai-differ-and-when-to-use-each.

Di Lillo, L. et. al. (2024). Do autonomous vehicles outperform latest-generation human-driven vehicles? A comparison to Waymo's auto liability insurance claims at 25.3 million miles. *Waymo*. https://waymo.com/research/do-autonomous-vehicles-outperform-latest-generation-human-driven-vehicles-25-million-miles/.

Dimanov, B. et al. (2020). You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. *Proceedings of the 24th European Conference on Artificial Intelligence, 2473–2480*.

Eom, D., Newman, T., Brossard, D., & Scheufele, D. A. (2024). Societal guardrails for AI?

Perspectives on what we know about public opinion on artificial intelligence. *Science and Public Policy, 51(5), 1004–1013*.

Executive Order (E.O.) 14,110. (2023, November 1). *Federal Register, 88*, 75191–75226. https://www.govinfo.gov/content/pkg/FR-2023-11-01/pdf/2023-24283.pdf.

European Parliament and the Council of the European Union (EU AI Act). (2024). *Artificial Intelligence Act*. https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

Farber, D. (2022, March-April). Staying within the guardrails. *The Environmental Forum, 40-47*.

Future of Life Institute. (2023). FLI on "a statement on AI risk" and next steps. *Future of Life Institute*. https://futureoflife.org/ai-policy/fli-on-a-statement-on-ai-risk-and-next-steps/.

Gasser, U., & Mayer-Schönberger, V. (2024). *Guardrails: Guiding human decisions in the age of AI*. Princeton University Press.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64, 86–92*.

Gemini Team et al. (2023). Gemini: a family of highly capable multimodal models. *arXiv*. http://arxiv.org/abs/2312.11805.

Hewitt, D. T. (2023). The new invisible hand: The impact of algorithms on competition and consumer rights, 118th U.S. Congress. https://www.judiciary.senate.gov/imo/media/doc/2023-12-13_pm_-_testimony_-_hewitt.pdf

Kearns, M., & Roth, A. (2020). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Kleindorfer, P. R. (2006). The risk management program rule and management-based regulation. In C. Coglianese & J. Nash (Eds.), *Leveraging the Private Sector: Management-based Strategies for Improving Environmental Performance*. Routledge.

Linsley, D. et al. (2019). Learning what and where to attend. In Proceedings at the *17th International Conference on Learning Representations*.

Livermore, M. A., & Revesz, R. L. (2020). *Reviving rationality: Saving cost-benefit analysis for the sake of the environment and our health*. Oxford University Press.

Mac, R., & Kang, C. (2021, October 3). Whistle-blower says Facebook "chooses profits over safety." *The New York Times*. https://www.nytimes.com/2021/10/03/technology/whistle-blower-facebook-frances-haugen.html.

Marchant, G. E. (2011). Addressing the pacing problem. In G. E. Marchant, B. R. Allenby & J. R. Herkert (Eds.), *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight*. Springer Dordrecht.

McKinsey & Company. What are AI guardrails? (2024, November 14). https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-are-ai-guardrails#/.

Metz, C. (2021, March 15). Who Is Making Sure the A.I. Machines Aren't Racist? *The New York Times*. https://www.nytimes.com/2021/03/15/technology/artificial-intelligence-google-bias.html.

Microsoft. (2023, February 15). The new Bing & Edge: Learning from our first week. *Microsoft Bing Blogs*. https://blogs.bing.com/search/february-2023/The-new-Bing-Edge-–-Learning-from-our-first-week.

Milmo, D. (2022, September 22). Pinterest executive admits platform 'not safe' when Molly Russell used it. *The Guardian*. https://www.theguardian.com/technology/2022/sep/22/pinterest-executive-deeply-regrets-content-viewed-by-molly-russell.

Minor, T., & Parrett, M. (2017). The economic impact of the Food and Drug Administration's

final juice HACCP rule. *Food Policy, 68, 206–213*.

Mitchell, M. et al. (2019). Model cards for model reporting. In Proceedings at the *Conference on Fairness, Accountability, and Transparency, 220-229*.

Murthy, V. H. (2024, June 17). Surgeon General: Why I'm calling for a warning label on social media platforms. *The New York Times*. https://www.nytimes.com/2024/06/17/opinion/social-media-health-warning.html.

National Academies of Sciences, Engineering, and Medicine (NASEM). (2018). *Designing safety regulations for high-hazard industries*. The National Academies Press.

National Transportation Safety Board (NTSB). (2020). Collision between a sport utility vehicle operating with partial driving automation and a crash attenuator, Mountain View, California, March 23, 2018. *National Transportation and Safety Board, Highway Accident Report NTSB/HAR-20/01.*

——— (2019). Collision between vehicle controlled by developmental automated driving system and pedestrian, Tempe, Arizona, March 18, 2018. *National Transportation and Safety Board, Highway Accident Report NTSB/HAR-19/03.*

——— (2018). Rear-end collision between a car operating with advanced driver assistance systems and a stationary fire truck, Culver City, California, January 22, 2018. *National Transportation Safety Board, Highway Accident Brief NTSB/HAR-19/03.*

OpenAI et al. (2023). GPT-4 technical report. *arXiv*. http://arxiv.org/abs/2303.08774.

Oppenheim, M. (2018, October 11). Amazon scraps "sexist AI" recruitment tool. *Independent*. https://www.independent.co.uk/tech/amazon-ai-sexist-recruitment-tool-algorithm-a8579161.html.

Perrigo, B. (2023, February 17). Bing's AI is threatening users: That's no laughing matter. *Time*.

https://time.com/6256529/bing-openai-chatgpt-danger-alignment/.

Richards, J. et. al. (2020). A Methodology for Creating AI FactSheets. *IBM Research*. http://arxiv.org/abs/2006.13796.

Riess, R., & Sottile, Z. (2023, July 29). Uber self-driving car test driver pleads guilty to endangerment in pedestrian death case. *CNN*. https://www.cnn.com/2023/07/29/business/uber-self-driving-car-death-guilty/index.html.

Rouse, M. (2023, December 24). AI guardrail. *Techopedia*. https://www.techopedia.com/definition/ai-guardrail.

Satariano, A. (2022, October 1). British Ruling Pins Blame on Social Media for Teenager's Suicide. *The New York Times*. https://www.nytimes.com/2022/10/01/business/instagram-suicide-ruling-britain.html.

Satariano, A. (2023, March 31). ChatGPT is banned in Italy over privacy concerns *The New York Times*. https://www.nytimes.com/2023/03/31/technology/chatgpt-italy-ban.html.

Satariano, A., & Metz, C. (2023, March 5). Using A.I. to Detect Breast Cancer That Doctors Miss. *The New York Times*. https://www.nytimes.com/2023/03/05/technology/artificial-intelligence-breast-cancer-detection.html.

Schauer, F., & Zeckhauser, R. (2007). Regulation by generalization. *Regulation & Governance, 1, 68–87*. https://doi.org/10.1111/j.1748-5991.2007.00003.x.

Schumer, C. (2023, October 26). At the Washington Post AI summit, majority leader Schumer talks on bipartisan AI insight forums, the future of AI, and AI regulation. *Senate Democratic Leadership*. https://www.democrats.senate.gov/newsroom/press-releases/transcript-at-the-washington-post-ai-summit-majority-leader-schumer-talks-on-bipartisan-ai-insight-forums-the-future-of-ai-and-ai-regulation.

Shamsujjoha, M. et al. (2024). A taxonomy of multi-layered runtime guardrails for designing foundation model-based agents: Swiss cheese model for AI safety by design. *arXiv*. https://arxiv.org/abs/2408.02205.

Shavit, Y. et al. (2023). *Practices for governing agentic AI systems*. OpenAI. https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf.

Slattery, P. et al. (2024). The AI risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. *arXiv*. https://arxiv.org/abs/2408.12622.

Smith-Renner, A. et al. (2020). No explainability without accountability: An empirical study of explanations and feedback in interactive ML. *Proceedings of the CHI Conference on Human Factors in Computing Systems, 1–13*. https://doi.org/10.1145/3313831.3376624.

Stephenson, M. C. (2011). Information acquisition and institutional design. *Harvard Law Review, 124, 1422*.

Touvron, H. et al. (2023). LLaMA: Open and Efficient Foundation Language Models. *arXiv*. http://arxiv.org/abs/2302.13971.

Warren, E. (2024, February 9). ICYMI: At hearing, Warren calls out Medicare Advantage insurers for using AI to deny care and boost profits. *Senator Elizabeth Warren Newsroom.* https://www.warren.senate.gov/newsroom/press-releases/icymi-at-hearing-warren-calls-out-medicare-advantage-insurers-for-using-ai-to-deny-care-and-boost-profits.

Working Group on Artificial Intelligence. (2024). AI as a public good: Ensuring democratic control of AI in the information space. *Forum on Information & Democracy*.

Zao-Sanders, M. (2024, March 19). How people are really using genAI. *Harvard Business Review*. https://hbr.org/2024/03/how-people-are-really-using-genai.