An FDA for AI? Pitfalls and Plausibility of Approval Regulation for Frontier Artificial Intelligence

Daniel Carpenter[1], Carson Ezell[1]

[1]Harvard University
dcarpenter@gov.harvard.edu, cezell@college.harvard.edu

Abstract

Observers and practitioners of artificial intelligence (AI) have proposed an FDA-style licensing regime for the most advanced AI models, or 'frontier' models. In this paper, we explore the applicability of *approval regulation* – that is, regulation of a product that combines experimental minima with government licensure conditioned partially or fully upon that experimentation – to the regulation of frontier AI. There are a number of reasons to believe that approval regulation, simplistically applied, would be inapposite for frontier AI risks. Domains of weak fit include the difficulty of defining the regulated product, the presence of Knightian uncertainty or deep ambiguity about harms from AI, the potentially transmissible nature of risks, and distributed activities among actors involved in the AI lifecycle. We conclude by highlighting the role of policy learning and experimentation in regulatory development, describing how learning from other forms of AI regulation and improvements in evaluation and testing methods can help to overcome some of the challenges we identify.

1 Introduction

Massive leaps in the scale, performance, and apparent risks of artificial intelligence (AI) have led practitioners and observers to call for various forms of regulation. Many proposals have focused on adapting AI regulations to account for the novel risks introduced by 'frontier' AI (Anderljung et al. 2023a; Schuett et al. 2024). Based on recent trends, we characterize frontier AI systems as usually having (1) general-purpose functionality, (2) more costly R&D processes (Whittaker 2021; Ahmed, Wahed, and Thompson 2023), (3) dual-use capabilities that pose misuse risks, and (4) systemic or structural risk (Zwetsloot and Dafoe 2019; Council of the European Union 2024). Some regulatory proposals for frontier AI are analogs of regulatory regimes already in existence, including that of the U.S. Food and Drug Administration (FDA). For example, at a recent congressional hearing, emeritus professor Gary Marcus stated that among the "many guardrails and regulations I would suggest," one was "Creating an FDA-like regulatory regime for AI that evaluates large-scale deployment, balancing risks and benefit" (Marcus 2023). There have been several proposals arguing for an explicit licensing regime

based upon FDA-style approval regulation, including licensing that would apply to model development, deployment, or the operation of large datacenters (Stein and Dunlop 2023; Matheny 2023; Microsoft 2023; Encode Justice and Future of Life Institute 2023; Smith 2024; Malgieri and Pasquale 2024; Allen et al. 2024). Other aspects of FDA regulation have also served as analogies. For example, the National Artificial Intelligence Advisory Committee (NAIAC) has called for an adverse events reporting system, wherein the FDA system of the same title (it has been known for decades as the AERS) figures as a reference point (The National Artificial Intelligence Advisory Committee 2023).

The possibility of an FDA-like regulatory regime for frontier AI has since occasioned considerable debate. AI-related regulatory frameworks and legislative proposals have involved various forms of licensing (Blumenthal and Hawley 2023; Center for AI Policy 2024), and critics have identified many limitations of a licensing regime (Guha et al. 2023; Wheeler 2024), including a range of libertarian organizations and writers who quickly aligned against the idea (Bailey 2023; Thierer and Chilson 2023). Thus far, the mapping of FDA-like analogies to AI regulation has largely proceeded by means of vague metaphors – understandable for an early stage of public debate, but not desirable as actual policies are discussed. Indeed, there are properties of AI and its risks that would call for reconsideration of some aspects of FDA-style regulation as it has been traditionally practiced.

In this paper, we argue that there is need for careful consideration of institutional and organizational forms before any regime, much less an "FDA-like regulatory regime," could be adopted. We proceed through four general claims. First, at its essential core, FDA-style regulation is a form of *approval regulation* linking mandatory testing with a regulatory veto over part or all of a firm's R&D process. Second, this regime of regulation makes specific assumptions about the product and firm that are being regulated, the measurability of risks from the product, the observability of a firm's actions (e.g. development and testing), and the enforceability of rules that prevent certain unapproved activities from taking place. Third, there are aspects of frontier AI that do not conform to these assumptions. Finally, policy experimentation and learning are essential to addressing some limitations of approval regulation, including from

other forms of AI regulation and developments in model evaluation methods.

In Section 2, we describe approval regulation, including its properties that make it unique from other forms of regulation. In Section 3, we describe the conditions that facilitate approval regulation in the FDA context. In Section 4, we consider the applicability of approval regulation to the frontier AI context. In Section 5, we overview the role of policy experimentation and learning in the development an an approval regulation regime.

Before proceeding, two prefatory notes. First, we make no judgment here about whether approval regulation is optimal or efficient in the spaces in which it has been applied, especially in the area of biomedical innovation. Second, it is important to consider the possible complementarity or substitutability of different regulatory policies. Much of the argument from libertarian voices suggests that it is possible to rely upon self-regulation, intellectual property regulation, fiduciary or "duty of care" standards, or tort liability regimes to regulate AI harms (Thierer 2023). These arguments may be on the mark, but it is worth noting that in many areas of regulation – and not just biomedical innovation – forms of approval regulation co-exist with these and other forms of governance. To say that they co-exist is *not* to assert that they do so without friction or inefficient cross-subsidization of activities. The point is that the desirability or plausibility of one form of regulatory institution does not, *ipso facto*, rule out the possible desirability or plausibility of another. Considering the optimal portfolio of institutions is exactly where research is needed, and it is unlikely that any such portfolio will be designed *ex nihilo* but will evolve.

2   "FDA-Style" Approval Regulation

Commentators referring to an "FDA-like" or "FDA-style" regime are usually referencing the FDA's regulation of new biomedical products, a function which is now global and exercised by dozens of national and regional regulators (the European Medicines Agency (EMA), for instance). At their core, "FDA-style" regimes rest upon structures of *approval regulation* (Carpenter 2004; Carpenter and Ting 2007; Carpenter, Grimmer, and Lomazoff 2010; Henry and Ottaviani 2019; Henry, Loseto, and Ottaviani 2022; Ottaviani and Wickelgren 2023), which we define here as a regime in which a regulator requires a firm to engage in testing before conducting subsequent activities (e.g. releasing a product), and in which this testing generates data that is used by the regulator to decide whether part or all of the subsequent activities can be conducted.

So defined, approval regulation gives the regulator a "veto" over stages of product development and release, but approval regulation is far more than a mere gatekeeping function or a veto. Any number of other regulatory mechanisms can regulate "entry" (Djankov et al. 2002), including mechanisms that are already being used in AI governance. For example, the EU AI Act requires ex-ante assessments of conformity with standards for high-risk AI deployments (Council of the European Union 2024).

Furthermore, many regulatory mechanisms can encourage or require testing, audits, and/or information disclo-sure without linkage to approval mechanisms. For example, Executive Order 14110 (EO 14110) in the U.S. requires that certain testing results for "dual-use foundation models" are reported to regulators (Executive Office of the President 2023). Furthermore, various standards-setting bodies, including the National Institute of Standards and Technology (NIST) in the U.S. (U.S. Department of Commerce 2023) and CEN (European Committee for Standardization)/CENELEC (European Committee for Electrotechnical Standardization) in the E.U. (Laux, Wachter, and Mittelstadt 2024), are developing further AI standards and best practices, including testing guidelines. Subsequent regulations can require or incentivize developers to follow these standards without a connection to approval mechanisms.

The essential properties of approval regulation were outlined in a series of mathematical models before 2010 (Carpenter 2004; Carpenter and Ting 2007; Carpenter, Grimmer, and Lomazoff 2010), and subsequent research has led to a more detailed understanding of how these regimes develop and operate. The history of approval regulation institutions has been the subject of studies in history and political science (Marks 1997; Carpenter 2010). Models from the economic theory and management science literature have examine general properties of regulation and veto institutions and consider issues such as optimal timing of entry and regulation, the structure of costly experimentation in persuasion, and the relationship between *ex ante* and *ex post* regulation (Henry and Ottaviani 2019; Henry, Loseto, and Ottaviani 2022; Ottaviani and Wickelgren 2023). Recent models have also explored a range of alternative institutional arrangements and the potential tradeoffs or complementarities among them (Henry and Ottaviani 2019; McClellan 2022; Bates et al. 2024). While all of these models are simplifications, they are nonetheless essential for understanding the critical operative structure and incentive-based kernels of these regimes, especially when modelers pay appropriate attention to the institutional context.

The *combination* of testing and veto in approval regulation is essential to differentiating these institutions from other institutions that erect entry barriers. These two powers reinforce each other, create particular incentives, and complement a range of other regulatory policies implemented and enforced by agencies such as the FDA and EMA.

The ability of regulators to write new rules governing testing depends heavily upon gatekeeping. Regulators verify that tests are conducted according to particular practices because specific testing results are necessary for them to perform their gatekeeping function. For example, required labeling for biomedical products incorporates information from required experiments, and the proposed labeling is an important part of the pre-market review. While regulators are influential in shaping and standardizing best practices in testing, their views of these best practices are also significantly influenced by developments that are exogenous to regulation. For example, the standards of pre-market review at the FDA developed hand-in-hand with changes in pharmacological and experimental standards (Marks 1997; Carpenter 2010). In terms of phased experimentation, developments in oncology (especially at the National Cancer Insti-

tute) were critical to the FDA's view of phased experiment (Carpenter 2010; Keating and Cambrosio 2014).

Approval regulation also creates particular experimental and long-range behavioral incentives. First, the fact that the regulator likely has a higher bar for converting R&D into product launch than does the firm itself means that firms have incentives to conduct more testing than they otherwise would (Carpenter and Ting 2007; Henry and Ottaviani 2019), and adhere more closely to practices specified by the regulator. Notably, the primary costs associated with gatekeeping regulation are not the agency's decision itself but the set of experiments that come before, which are directly observed and regulated.[1] Second, a single company likely has a range of products, some that are already released and others that are under development. A key property of the biomedical marketplace is that there is more profit to be made from the newest products than the older ones, due in part to patents (Carpenter 2004; Carpenter et al. 2010). This means that even a profitable firm has strong incentives to behave "well" in front of the approval regulator, as its profitability depends heavily upon a stream of new molecules to be authorized in the future. Being perceived well by the regulator can both increase the chance of approval and reduce the expected time to approval.

Of course, the EMA and FDA do many things other than require testing and decide upon the marketability of new biomedical products. These agencies inspect production facilities, require firms to conduct experiments after regulatory authorization, require firms and other actors to generate reports on "adverse events" associated with the product, monitor other data (a form of observational epidemiology), consider revisions to labels and warnings, and also regulate advertising and marketing practices. How can we consider these in relation to approval regulation? It is useful to differentiate here between the set of things that happen before a product is authorized – *ex ante* regulation – and the set of things that happen afterwards – *ex post* regulation (Carpenter 2010; Henry, Loseto, and Ottaviani 2022). The basic structure of phased experiment – Phase I trials for basic toxicity in non-diseased individuals, Phase II and III trials for examination of safety and efficacy in diseased populations – occurs before marketing authorization (the "veto"). Yet important regulatory tools are available after regulatory marketing authorization. The regulator can require or request changes in labeling, can remove the product from the market (making the initial approval reversible at least in fact) and can, on its own volition, monitor a range of other data on the evolving risks of the approved product.

---

[1]In the model of Carpenter and Ting (2007), the firm possesses a more precise prior on the state variable of the regulated product – the asymmetric information is not absolute – but all experiments are publicly observed. Later approval regulation models have a similar structure, and while there are aspects of this assumption that are violated in the real world (such as when a regulatory sponsor has access to certain aspects of Phase III trial records that the regulator does not), this simplification captures much of the actual operation of approval regulation regimes.

3 Conditions for Approval Regulation

As it has developed in the area of biomedical innovation (Marks 1997; Carpenter 2010), approval regulation assumes a particular form. A firm develops a molecule and then begins to test it, first upon non-human animals and then upon humans in a series of clinical trials.[2] The regulator observes these trials and their results on roughly the same schedule – though not, simultaneously, with the same precision – as does the firm. The firm then collects data and documentation from these experiments and other tests (such as manufacturing data) and submits a "new drug application" or "dossier" to the regulator. The dossier is massive and is the basis for the regulator's decision of whether or not to authorize/release for marketing of the drug. After regulatory approval, the regulator often mandates further experiments (often called "postmarketing trials" or "Phase IV trials") and also monitors the risk profile of the molecule through a combination of inspections, adverse event reports and survey of databases. Medical device regulation carries forward many principles and institutions from drug regulation. In both molecules and devices, the dominant regulatory regimes for the FDA include mandatory pre-market experimentation and then an approval decision *based upon those experiments*.

The set of assumptions and enabling structures undergirding these regulatory regimes is considerable. It includes:

- **Identifiability of a regulated unit**. In examining any regulatory policy, we should ask what is the thing to be regulated, to be governed? In the case of biopharmaceutical regulation, it is the molecule even more than the firm. More specifically and germanely, approval regulation in biopharmaceuticals generally possesses an identifiable object of regulation. This is not exogenous to regulation but is defined in part by the law itself, in the concepts of Investigational New Drug and New Molecular Entity or New Therapeutic Biological Products, or in the case of medical devices, Class III devices.

- **Identifiability of firms engaging in regulated activities**. In part because biomedical innovation is exogenously costly, in part because the costs associated with regulation itself, and in part because of the incentives stemming from patent systems (an agent must claim intellectual property rights over the molecule in order to enjoy patent protection upon its marketing authorization), the production of new therapeutic molecules and the agents or organizations that produce them and conduct experiments upon them is often well known. This assumption holds even in innovation markets with highly secondary and tertiary markets for contracting and subcontracting.

---

[2]Importantly, at the EMA and FDA, the relevant regulated organization (the "firm") is not necessarily the one that "discovered" the product (molecule) but its rather the "sponsor," the firm that prepares and submits the regulatory dossier. As detailed in Carpenter (2010, Chapter 10), the structure of approval regulation at the FDA and related agencies is such that regulatory sponsorship is now an established, if not pivotal, component of biopharmaceutical firms.

- **Testing methods to identify and measure risk.** In biopharmaceutical regulation, two facts about the data used in evaluation are that (1) the adverse events to which probabilities are assigned are often known and detectible and (2) well-known probability models can be developed to describe the risk of these adverse events, such that these probability models are consulted directly in product evaluation. While in theory the set of things that could go wrong is infinite, in practice it is usually quite manageable.[3] For instance, a vast amount of research has been conducted on the risk of hepatotoxicity associated with the ingestion of biopharmaceuticals, as many of these products place heavy demands upon the liver and their therapeutic properties often depend upon metabolization there. An entire set of measurements and statistics are available for measuring these risks and assigning probabilities or severity measures to them. The "set of things that could go wrong" is often well known and regulators know where to look for most of (perhaps not all of) the risk. Beyond this, the tests conducted by developers and required by regulators make it more likely that adverse events will be potentially observable at sufficient frequency that large-sample properties of statistical inference can be applied.

- **Observability of the fact of testing, once mandated**. In biopharmaceutical regulation, the event that "the firm conducts a test upon its product" is highly observable, and in models of approval regulation (Carpenter 2004; Carpenter and Ting 2007; Henry and Ottaviani 2019), this fact is perfectly observable and at a cost known to the regulator as well as the firm. This fact is in part endogenous to institutions, including regulatory institutions, because the molecule is registered with the FDA (all drugs under study in the United States must have an approved status of Investigational New Drug (IND)) as well as professional institutions (funding agencies such as the National Institute of Health, research clinics and hospitals that are regulated by professions and by numerous levels of government). Furthermore, groups of professional scientists and statisticians are routinely consulted in the design, pre-registration and analysis of these experiments. This observability assumes that regulators and experts external to the firm are given access to information about the product and the experiments conducted upon it.

- **Observability of the fact of development**. In biopharmaceutical regulation, it is difficult for actors to conceal the development, release, and marketing of new therapeutic products. For example, if a consumer wants insurance to pay for health services, they will need to come from a licensed or recognize provider, and relatedly, the product prescribed to the consumer will need to be listed on some kind of formulary. In the market for human medical services as well as the market for therapeutic commodities (pharmaceuticals or devices), the vast insurance market serves as a *de facto* regulator of illegal develop-

ment and provision. It is not impossible, however, and substantial activity prevails at the margins of the regulated marketplace, either with known but unregulated products that are consumed (but not legally marketed) with believed health effects in mind, such as nutritional supplements, or with non-ethical drug use for health-related purposes (those who grow their own cannabis and who use it for self-ascribed health improvement reasons). In related forms of regulation, such as the regulation of new dams or nuclear reactors, the ability of an actor to "innovate" (create a new product) outside the bounds of regulation is again quite limited. In the field of molecules, this fact is also not exogenous to institutions, as a range of drug enforcement agencies at various levels of government monitor and enforce laws against unauthorized production of chemical substances.

- **An industrial structure and social institutions that facilitate the previous assumptions.** The identifiability of firms, the ability of the regulator (or other agents) to observe these firms' behavior, and the observability of the fact of testing (a kind of compliance) are greatly facilitated in the biopharmaceutical industry by the fact that the number of firms, while large, is not so large as to defy manageability. Once we consider the fact that the field for evaluating risk in biopharmaceuticals is often bounded by the extent of a diseased population, it is further the case that the number of firms and laboratories active in a particular disease market is far smaller than the set of all biopharma firms generally. While there is no mathematical or empirical proof of the hypothesis, there may be reason to believe that the feasibility of approval regulation depends in part upon an oligopolistic industrial structure. Beyond this, much of FDA governance in molecules and medical devices is assisted by, relies upon the science and professional standards of, and assumes the enforcement of physicians and other medical and health professions.

A final note. Some observers might quibble, and fairly, with this simplified description of the biopharmaceutical world to which "FDA-style" approval regulation has been applied. Our point is that these stylized facts have characterized something of the "steady state" of the biopharmaceutical world, even as it is an incredibly dynamic domain with massive amounts of investment and innovation. Entire modes of innovation, from early forms of model-assisted drug development to the important role that AI itself now plays in drug development, have changed. And yet some of the institutional and contextual features of the system are quite stable, and not only because of approval regulation.

4 Potential Pitfalls for AI Approval Regulation

Given these stylized conditions that facilitate approval regulation, especially in the biopharmaceutical realm, we now turn to the emerging field of frontier AI development and assess the extent to which its characteristics are conducive to FDA-style regulation. Whether the facts adumbrated in the previous section apply to frontier AI regulation is an empirical question. It is possible that the conditions for applicabil-

---

[3]This is even true with the transmissible risk from biologics, as in many cases infectious disease specialists know at least some, if not many, of the "red flags" to look for.

ity of approval regulation to biopharmaceuticals, which are shown in Table 1, are not *yet* satisfied in the area of frontier AI, but that they could be in the future, given policies or forms of industrial evolution, so nothing in this section should be construed as an impossibility result. Another way of putting the matter is that *the potential fit between models of approval regulation and frontier AI is a fruitful research agenda in institutional design as well as applied governance*.

4.1 Scope

Defining Regulated Units Approval gates are intended to regulate risky products or activities, so they rely on clear definitions of what counts as risky. However, frontier AI poses several challenges to such demarcations.

First, there are not clear metrics to characterize the risk posed by AI systems, in part due to their generality and complexity. While definitions of foundation models that are dual-use or systemically risky have been used as the basis of regulatory action in the U.S. and E.U. (Executive Office of the President 2023; Council of the European Union 2024), these definitions have been criticized for being overinclusive or underinclusive of certain types of systems that might be developed (Schuett 2023; Bommasani 2023).

Furthermore, actors throughout the distribution chain can make a vast array of modifications to AI systems which can alter their risk profile (Davidson et al. 2023), exacerbating the problem of defining what counts as a 'new' unit subject to gatekeeping. For example, fine-tuning or other modifications to parameters can alter a system's behavior or capabilities, and scaffolding frameworks in which AI systems are embedded can also alter their risk (Sharkey et al. 2024). There is an open question of the extent to which this problem can be addressed by emerging methods that make it more difficult to modify models to introduce undesirable behaviors (Deng et al. 2024; Sheshadri et al. 2024). Another consideration is that a new foundation model with similar or identical properties to an existing model (e.g. architecture, data, etc.) might not count as a new system for the purpose of regulation.

The identifiability of homogeneous regulated units is not merely important for determining when a new system is developed that is subject to approval, but also for aggregating data about risk to enable more informed assessments (Bommasani et al. 2022). If new data about an AI system (e.g. from testing or incident reports) leads to an updated risk profile, regulators rely upon an understanding of how applicable the new findings are to other systems. However, this task becomes more difficult as heterogeneity becomes more complicated. The FDA has more established methods for data aggregation. For example, when examining a large dataset of chemical assays of a molecule, or the experience of thousands of patients with that molecule, or the mechanical properties of a hip implant, or the experiences of thousands of patients with said device, both the product and the experience have to sufficiently comparable (or "commensurable" as to be able to aggregated).

In addition, the development of a model itself introduces risks. In biomedical innovation, there are many products and

experiments that the public or regulators generally do not see or do not observe as thoroughly, and this is especially so for the products that "fail" in the sense of not having achieved market launch (Hwang et al. 2016). These products sit on the "shelf" and there is not likely much of a risk of their being seized and deployed for other uses.[4] In the world of algorithms there seems little, beside strong intellectual property and cybersecurity protections, to prevent others from unsafely using them (Guha et al. 2023; Nevo et al. 2023), including through stealing model weights or developing a similar model of their own. This raises the question of whether approval should be required for certain forms of development activities prior to any deployment, which may include the conduct of a large training run, certain forms of modification after pre-training, or the operation of a large datacenter where regulated models are trained and stored.

Regulated Entities The originators of foundation models are, for the moment and in general, well known. However, compared to a range of other regulated entities – say bank holding companies regulated by the Federal Reserve and other national bank regulators, or biopharmaceutical and medical device companies as regulated by the FDA or EMA – there is far less known about the industrial structure of the AI industry. This fact stems in part from the novelty of the industry and its rapid rise, but also from the fact of its non-regulation. Regulation often stipulates certain organizational forms be taken by a regulated organization (a compliance department, or a regulatory affairs department) that must then function as a liaison between the organization and the relevant regulatory agency. These sub-organizations produce considerable data and fulfill reporting requirements. They function as a translator for the agency and make the regulated firm and its products more "observable."

It is unclear whether the industrial organization of foundation model development will lead to an industrial structure with these properties. There are reasons to think that the future of foundation model development will be characterized by high-cost research and development and by a small number of dominant firms whose models not only outcompete the models of other firms on a performance basis, but also learn about the strengths and weaknesses of those rival models and adapt. Indeed, the compute cost of training a frontier model is increasing rapidly (Whittaker 2021; Sevilla et al. 2022; Cottier 2023). As with many other capital-intensive industries, then, the number of operative firms would be reduced. Large and well-resourced companies or laboratories would more likely have the organizational and financial capacity to comply with intensive reporting requirements. At least OpenAI and Anthropic have established internal positions or teams responsible for documenting the implementation of catastrophic risk assessment practices for frontier models (OpenAI 2023; Anthropic 2023).

However, several factors could enable many smaller organizations to be involved in the development of new foundation models, including reductions in the cost of development

---

[4]In some sense, intellectual property regimes address some of this risk, but in most regulated markets they address the risk of illegal appropriation for profit, not for misuse.

Category	Considerations
Scope	**Regulated Units:** What characteristics demarcate products or activities that are subject to approval regulation?
	Regulated Entities: How conducive are the organizational forms of entities involved in frontier AI development to facilitating oversight and complying with requirements?
Observability	**Testing Requirements:** What evaluation tools and tests are available for measuring risks and informing approval decisions?
	Oversight Mechanisms: What oversight mechanisms are available for regulators to verify firms' compliance and ensure the rigor of model evaluation/testing or other forms of risk assessment?
Enforceability	**Control of Unregulated Activities:** To what extent do conditions enable unreported activities subject to regulation to persist, included unreported domestic activities or foreign activities that undermine the efficacy of domestic approval regulation?

Table 1: Conditions for the applicability of approval regulation to frontier AI, based on experiences from biomedical regulation.

of frontier models from improvements in algorithmic efficiency (Ho et al. 2024) or meaningful alterations via post-training enhancements (Qi et al. 2023). These organizations are less likely to have the resources to engage in as rigorous compliance and reporting activities.

Another problem arises from the fact that the set of organizations that deploy foundation models may differ materially and appreciably from the set of labs that create them. If deployers are making consequential decisions that impact the risk profile, including implementing their own usage monitoring or other safety guardrails, regulators may have an interest in granting approval for actions by deployers rather than, or in addition to, upstream developers (Stein and Dunlop 2023). However, deployers may lack the ability to conduct as rigorous testing and reporting as upstream developers due to having less expertise, resources, and, perhaps most importantly, access to proprietary information about the system that is maintained by upstream developers (Bommasani et al. 2023a; Hacker, Engel, and Mauer 2023; Anderljung et al. 2023b; Casper et al. 2024).

In addition, deployers would likely lack the institutional forms that facilitate observability and verification of compliance, especially where model weights are openly released (Seger et al. 2023; Kapoor et al. 2024). The general principle here is that approval regulation in the biomedical realm depends upon a set of social and economic institutions that developed alongside and somewhat separably from approval regulators like the FDA or EMA. In the biomedical realm, the secondary market for the "deployment" of approved technologies is regulated by the professionalization of prescribers and, more implicitly but no less consequentially, by the tort system. Similar structures and institutional forms are still in their early stages for AI development, deployment, and usage (Solaiman 2023; Eiras et al. 2024; Gorwa and Veale 2024; Shevlane 2024). Yet this raises the question for AI regulation of what social and economics structures – professionals that regulate use, tort systems that impose liability constraints, concentrated industrial structure that enhances the prospect for compliance capacity – will emerge in foundation models.

4.2 Observability

Experimentation Requirements In an important observation, Knight (1921) described a form of "uncertainty" in which events can be enumerated but probabilities cannot be assigned to them. In a recent paper, Sunstein (2023) reviews the postulates of this concept and argues that regulatory policy development must take account of this ineluctable fact.

Whether probabilities can be assigned to the various risk events that we encounter with the development of AI to form the basis of effective regulatory decisions is not known. The complexity of the deployment environment means that model behaviors and their resulting effects are difficult to anticipate (Weidinger et al. 2023).

But even if Knightian uncertainty did not exist in this world, another problem would: deep ambiguity or what Kay and King (2020) call "radical uncertainty." Compared to most regulated domains, the AI domain seems replete with potential risks and rewards that are, almost by forcible extension from the promise and pitfalls of artificial intelligence, hard to imagine. A related concern is what Taleb (2014) has called *the Lucretius problem*, namely the tendency to believe that the past contains the full set of harms that could occur and that nothing worse than what is in that (memory) set could possibly occur in the future. This problem is exacerbated by increasingly capable models that can create previously unknown pathways for risk (Shevlane et al. 2023; OpenAI 2023). Or the auxiliary risks from diffusive bioweapons, proliferating nuclear weapons, or interconnectedness may exacerbate the harm that could happen from an otherwise stable process governing risks from foundation models (Zwetsloot and Dafoe 2019). This makes risk evaluation and risk management not merely a difficult proposition but also requires those who would regulate frontier AI to consider scenarios that have never before occurred *and have not yet been imagined*, either by machine or by human.

Recent regulatory developments suggest that one kind of testing that is and will be conducted upon foundation models is 'red teaming' (Ganguli et al. 2022; Perez et al. 2022; Rando et al. 2022; Casper et al. 2023; Feffer et al. 2024), which EO 14110 defines as a "structured testing effort" that usually involves "adversarial methods to identify flaws and vulnerabilities, such as harmful or discriminatory outputs from an AI system, unforeseen or undesirable system behav-

iors, limitations, or potential risks associated with the misuse of the system" (Executive Office of the President 2023). The mapping from AI red teaming to risk assessment is akin to financial stress-testing, where red teaming exercises can provide insights into harms that can arise in various scenarios or contexts. Indeed, stress testing in the financial sector involves considering worst-case scenarios and contexts involving systemic risk. While previous AI red-teaming efforts have been limited (Feffer et al. 2024), more rigorous red-teaming exercises may involve providing red-teamers with greater access to models to assess worst-case behaviors (Kinniment et al. 2023; Casper et al. 2024). Advances in methods to identify a wider array of undesirable behaviors can enhance the effectiveness of red-teaming. However, red-teaming shares a property of stress testing that tests are biased towards studying scenarios that humans have already imagined. Indeed, stress tests were conducted before the 2008 financial crisis, but they did not imagine and test for a scenario with sufficient stress from a decline in housing prices (Frame, Gerardi, and Willen 2015).

Behavioral testing methods, such as AI red-teaming as it is often practiced, are also prone to producing misleading results (Casper et al. 2024), including due to a poor understanding of training dynamics (Schaeffer, Miranda, and Koyejo 2023) or data contamination (Golchin and Surdeanu 2023; Oren et al. 2023). A related concern is that the training process might encourage advanced models to behave well during behavioral testing in contrast to actual deployment (Berglund et al. 2023; Cohen et al. 2024; Ngo, Chan, and Mindermann 2024; Hubinger et al. 2024).

Testing can involve complementary approaches beyond behavioral evaluations. For example, interpretability methods focused on studying model internals can gain insights into model reasoning (Wang et al. 2022; Li et al. 2024a). While researchers often use these methods to analyze small models (Elhage et al. 2022), recent work raises the question of whether emerging interpretability methods can produce valuable insights about the behavior of large models (Cunningham et al. 2023; Marks et al. 2024; Templeton et al. 2024). "In-the-wild" testing (Naihin et al. 2023), including in sandbox environments (Park et al. 2023), can also produce insights that are difficult to produce in more controlled settings. Ecosystem-wide documentation (Bommasani et al. 2023b; Chan et al. 2024) can also be directly consulted to inform risk assessment. In general, various forms of testing can complement each other when assessing the risk profile of a model.

The question of risk from frontier AI is not merely the question of considering various pathways and their likelihoods, but also *the potential costs incurred once that barrier is ruptured* (or ruptured with sufficient severity that serious human costs occur). To be clear, any regulatory regime that makes decisions based in part on imagined worst-case scenarios would have to avoid implementing the most naïve decision rules. Just because an exercise can produce a horrific imagined result – the end of the world – should not imply that the most restrictive regulatory response should be adopted (Sunstein 2009). Any speculative exercise that included the worst possible scenario would also need to consider humanity's likely best response in addition to regulatory options.

Despite ongoing experimentation and recent progress in AI risk assessment, the risks are currently far better known in biomedical innovation, in part because they have been known descriptively for decades or even a century or more. We can and do measure the risk of liver damage or hepatotoxicity from drugs, but beyond that, there is abundant community knowledge about where such risks can lead and the likely profile of costs that can be imposed. In oncology, for instance, there is an entire subfield dedicated to studying the cardiac risks of oncologic therapies, including cytotoxic and immunotherapeutic interventions (Herrmann et al. 2022; Lyon et al. 2020). The "event" (hepatotoxicity, cardiotoxicity) can be defined, as can its attendant sequellae that imposes costs upon the human person (conditional probability or likelihood of dysfunction requiring a transplant, or mortality). Or in disaster insurance, there are entire industries dedicated to modeling the aggregate effects of a hurricane or tornado cluster. In short, there are a set of questions that any implementable risk science would need to be addressed in any risk-benefit analysis of a foundation model.

Oversight Mechanisms If tests are required, what is the enforcement regime for ensuring that they are carried out? Even in the area of biomedical regulation, many pivotal trials are not reported and many post-approval trials are neither commenced, completed nor fully reported (Carpenter 2010; Moore and Furberg 2014; Hwang, Kesselheim, and Bourgeois 2014; Wallach et al. 2018). One descriptive study of new drugs approved by the FDA in 2008 found that five years later (2013) "26 of 85 (31%) of the postmarketing study commitments had been fulfilled, and 8 (9%) [of those studies] had been submitted for agency review" (Moore and Furberg 2014; see also Carpenter 2014).

The regulated organization would be responsible for carrying out testing or permitting government or third-party observers access to the resources with which they could be performed. In the case of financial stress tests, the regulated organization is often one of the most heavily regulated and well-documented organizations on the planet. Consider, for example, the kinds of data that the Federal Reserve carries and published on commercial banks or bank holding companies (https://www.federalreserve.gov/data.htm). On a quarterly basis, regulators observe hundreds if not thousands of indicators on the operation of each entity they regulate. In the case of bank holding companies, for instance, this incudes a regular statement of their consulting, advising and external legal expenses (Libgober and Carpenter 2024). And as of May 2022, different government agencies employ over 60,000 bank examiners.[5]

A similar degree of oversight does not exist for frontier AI developers. Many developers are hesitant to share information about their proprietary models (Bommasani et al. 2023a) and have incentives to limit information sharing (Casper et al. 2024; Kolt et al. 2024). Furthermore, doc-

---

[5]See the data adduced by the Bureau of Labor Statistics, which decomposes the bank examiner population into several professional types; https://www.bls.gov/oes/current/oes132061.htm.
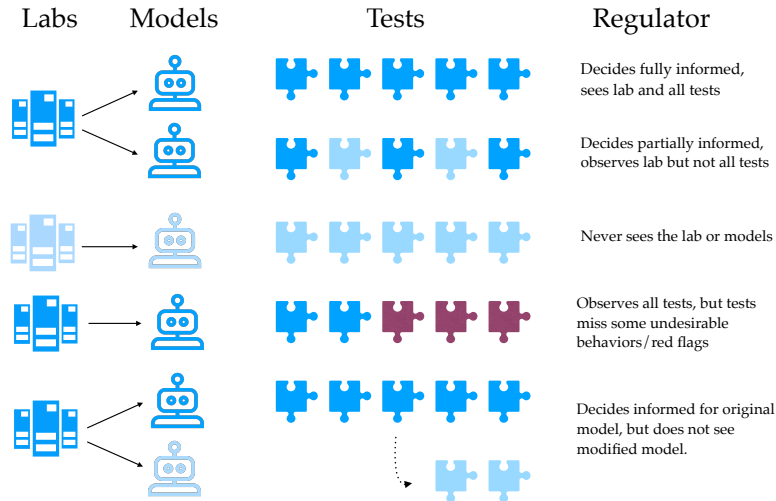
Figure 1: Possible scenarios where regulators lack complete information about frontier AI model development or testing. Arrows show which models were trained by which labs. Dark blue icons reflect regulators having complete information, and faded blue icons reflect a lack of regulatory visibility. Purple icons represent misleading or uninformative test results.

umentation and disclosure practices among frontier model developers are inconsistent and incomplete (Kolt et al. 2024; Pal, Bau, and Miller 2024). This inhibits an understanding of model behaviors and risks even when information is shared, and some relevant information (e.g. detailed documentation of internal testing) might not be documented at all. In Figure 1, we show several scenarios which could result in regulators lacking complete information about models subject to approval regulation, including testing data.

4.3 Enforceability

Control of Unregulated Activities Approval regulation and any other kind of licensing or entry regulation depends upon institutions of detection. Furthermore, approval regulation is reinforced by the existence of an organization that could be sanctioned for illegally marketing or distributing an unapproved product, or that could potentially be fined for failure to observe regulatory requirements.

Direct regulation of innovators is becoming a standard feature of policy proposals in the AI domain. This is the direction in which the Biden Administration in the United States (Executive Office of the President 2023) as well as the European Union are moving. The question becomes how enforceable such requirements are. The applicability of approval regulation to frontier AI governance depends, again, upon the existence, whether designed or co-evolved, of an industry structure that permits detection of R&D, violation of regulatory requirements, and feasible compliance activities.

If firms or labs do not wish to announce the development of a new model, or if there are many small labs capable of producing new foundation models, then it may be more dif-

ficult for any third-party agent to observe many acts of a new foundation model being developed or deployed.

However, foundation model development that is not reported by developers may be observable to compute providers because of the scale of frontier AI model training runs (Sastry et al. 2024). As new AI models are developed and deployed, they often require massive utilization of computing power (and, relatedly, monetary investments to purchase relevant equipment, processing time and concomitant utilization of energy), so they are trained in large datacenters (Pilz and Heim 2023). If these expenditures can be measured by regulators or third parties, then development of new foundation models may be detectible (Shavit 2023; Heim et al. 2024). Another possibility is that the expense of new model development may be so high as to induce exogenous barriers to entry and a small number of dominant firms or labs. Then as with the earlier problem of regulated organizations, industrial structure – something like an oligopoly – may reduce the set of regulable players to a manageable number.

Still, some risk may come from the fact that organizations in less well-regulated jurisdictions may wish to invest in laboratories or compute infrastructure to develop their own AI capabilities – this may include state-sponsored organizations. In addition, regulatory settings with lower-cost regulatory requirements might well attract more laboratory activity from abroad and compute infrastructure development. Because foundation models present the prospect of highly diffusive and contagious risk, the globalization or "harmonization" of regulatory requirements would likely be far more important in regulating AI than it would in regulating biomedical products (Ho et al. 2023; Trager et al. 2023).

Hence, an effective approval regulation regime would still be constrained without legal uniformity (at least for regulatory minima) for foundational model developers (Cihon 2019), as well as international coordination on monitoring development (Trager et al. 2023; Shavit 2023).

5 Developing Approval Regulation through Experimentation and Learning

The upshot of these considerations might be that the conditions do not yet exist for the implementation of an effective approval regulation regime. However, it would be premature to conclude that the obstacles we laid out could not be overcome, especially through policy experimentation and learning. Any regulatory policy must be considered in a dynamic context, which means that *the status quo must always be regarded as at least partially an experiment from which lessons can be drawn and to which adaptations can be made*. The longer history of approval regulation in molecules has taken the better part of a century (in devices, a half-century at least) to evolve, and decades- or century-long time horizons have characterized the evolution of regulation in other domains such as antitrust, anti-collusion, consumer product safety and systemic finance.

Consider that molecular regulation in therapeutics started without a regulatory veto for therapeutic drugs—the 1906 Pure Food and Drugs Act gave the federal government post-market inspection and product removal power (though note that the very first vaccines *did* have something like a gatekeeping institution in the 1902 Biologics and Vaccines Act). Regulatory development depended heavily upon coincident developments in pharmacology, statistics and the study of clinical trials and cancer therapeutics (Keating and Cambrosio 2014). It was these developments, combined with particular regulatory crises, that led to a new regime of regulatory pre-market review in the 1930s and the subsequent Kefauver-Harris Amendments of 1962 (Carpenter 2010, Chapter 3), which mandated proof of "effectiveness".

Furthermore, regulatory experience at the FDA spurred scientific findings that have led to various transformations of its approval models. For example, in some areas of therapeutics, while there is an abiding debate about the merits of such programs (Fleming 2005; Moore and Furberg 2014; Carpenter 2014; Budish, Roin, and Williams 2015; Naci, Smalley, and Kesselheim 2017), most or all new drugs are now approved on the basis of surrogate endpoints (Yu et al. 2015). The basic idea is that what society most cares about is mortality and morbidity, but that stand-in correlates of these core variables (tumor growth in solid tumors, say, or A1C reduction in diabetes medications) can be observed or measured earlier in the experimentation process, and may be sufficient for making decisions about the marketability of a new product.

5.1 Learning from Other AI Regulations

One possibility, and a scenario that has some historical experience to support its plausibility, is that "lighter" and more inchoate forms of regulation may generate lessons applicable to regulatory reform. A range of governance proposals and regimes have already emerged for AI and foundation models. Regulation of foundation models is trending toward the adoption of registration and reporting requirements, and there are many aspects of these regimes, too, that suffer from adaptability and feasibility problems (Guha et al. 2023). Several of the challenges are similar and may offer applicable lessons, including establishing definitions of regulated systems (Schuett 2023; Bommasani 2023), conducting informative tests, establishing mechanisms for oversight of development, and limiting non-compliant activities. In some sense, there is relevant experimentation right now.

In addition, early forms of foundation model regulation can precipitate the development of similar regimes in other jurisdictions. Such patterns have already emerged with institutions focused at least in part on frontier AI governance—the establishment of the UK AI Safety Institute (Department for Science, Innovation & Technology 2024) was followed by the establishment of similar institutions in at least the U.S. (U.S. Department of Commerce 2023), Japan (Shimbun 2023), and Canada (Cass-Beggs 2024). Indeed, with pharmaceutical regulation, there have been adaptations of FDA-like regulatory frameworks adopted across national and regional settings. These patterns can increase international coordination and facilitate more experimentation, although more experimentation across national contexts trades off against harmonization and creates risks of regulatory 'races to the bottom'.

European societies were long accustomed to apply less stringent approval regulation to pharmaceuticals than in the United States. The reduced stringency took several forms: (1) weaker experimental standards entailing less costly experiments that observed fewer dimensions of efficacy and risk, (2) weaker requirements on dossiers such that experimental data were summarized and not fully reported, and, finally, (3) easier approval standards. Counter-intuitively from the perspective of regulatory "races to the bottom," it is Europe that moved in the direction of the United States, not vice versa (Carpenter 2010, Chapter 12). Many observers now consider European biopharmaceutical regulation to be more stringent than in the United States.

5.2 Learning from AI Evaluation and Testing

Methods for model evaluation and testing are being developed and iterated upon both in the context of emerging regulatory regimes, as well as within the AI research community more broadly (Chang et al. 2024; Birhane et al. 2024). These developments include novel benchmarks (Srivastava et al. 2023; Li et al. 2024b), methodologies and tools (Kinniment et al. 2023; Ojewale et al. 2024; Hubinger et al. 2024), taxonomies of risks (Weidinger et al. 2023; OpenAI 2023; Critch and Russell 2023; Shevlane et al. 2023; Anthropic 2023), and documentation practices for communicating results (Gilbert et al. 2023; Kolt et al. 2024; Clymer et al. 2024). Thus far, evaluation practices for frontier models have been largely unstandardized (Feffer et al. 2024), but they have produced key learnings (Ganguli et al. 2023), and there are nascent efforts to increase standardization of evaluation practices (METR 2024; U.S. Department of Commerce 2024; AI Safety Institute and Department for Science,

Innovation & Technology 2024).

One possibility is that a set of potentially governable risks and tools to measure them might be adduced as they emerge in either experimentation or in real-world behavior. This is in the spirit of reporting requirements and incident reporting systems (McGregor 2021; The National Artificial Intelligence Advisory Committee 2023). The many decades of experience with adverse event reporting systems in biomedical innovation suggest that it will take considerable time and institutional investment to develop standardized frameworks for evaluation.

The scope of this nascent evaluation and risk detection industry is beyond the ambit of this paper. An important question for those proposing approval regulation regimes (Stein and Dunlop 2023), a variety of "FDA-like" institutions (Tutt 2016) or even "adverse event reporting systems" (The National Artificial Intelligence Advisory Committee 2023), however, is whether a standardized framework for threat detection and risk evaluation can emerge from these scattered efforts. One may wish for a less standardized approach, but a true "system"-based approach to regulation will, sooner or later, seek to aggregate across different datasets and analyses.[6] In the biomedical regulation world, there have been decades of calls for "harmonization" of regulatory requirements and standards across nations. The prima facie logic inspiring these proposals seems defensible, but given that federalism is itself a form of experimentation (Volden, Ting, and Carpenter 2008; Callander and Harstad 2015), one worries that learning value is surrendered when regulatory harmonization develops into strong uniformity.

One final problem with an experimental and incremental approach to regulation is that the materialization of the most severe risks may create conditions from which it is hard to escape. The most "catastrophic" risks from foundation model development may call for more stringent regulation in the first place (Stein and Dunlop 2023; Weil 2024; Cohen et al. 2024).

6   Conclusion

This paper joins other calls for circumspection in the application of regulatory models to generative artificial intelligence, in particular calling for more careful consideration of the feasibility of "FDA-like" approval regulation regimes to the regulation of frontier AI models and the catastrophic risks they may pose. The greatest impediments to such a model, in our judgment, are (1) enforceability of rigorous testing requirements and development/deployment restrictions and, perhaps most important, (2) the inapposite mapping between AI evaluation and the world of large samples and well-defined risks in which approval regulation operates, due to the lack of well-established indicators of catastrophic risk. However, we propose viewing these obstacles through the lens of policy learning, where the emergence

---

[6]Another way of putting the question here is whether any unified regulatory regime should exist at all, as opposed to a range of less centralized arrangements operating in communication, but not stringent coordination, with each other. This is quite different from calls for self-regulation or no regulation at all.

of a regulatory regime that achieves fit within its domain depends upon adaptation and incorporation of new information from both regulatory experience and exogenous factors.

Regulatory change, of course, implies neither regulatory evolution in a "fitness" sense nor monotonic improvement. Yet in a range of domains, it is at least plausible that regulation has been transformed due to criticism, scientific analysis, benefit-cost analysis and more rational forms of political oversight (McCraw 1986). This may not rise to the level of the culture championed by Greenstone (2009), but that does not mean that useful information cannot be yielded by such learning, nor does it mean that a less formally experimental approach is worse. Learning about policies from prospectively designed experiments alone may be difficult over the long run, and recent arguments suggest that a purely experimental approach may be wrong for optimization of policies in different domains (Stevenson 2023). Whatever the preferred mode of policy learning, it would be essential to approach such inferences prospectively and retrospectively, and to consider hybrid forms of regulation, given the rapidly changing nature of foundation models in AI and the often unquantifiable nature of their dangers.

Our point is not that approval regulation is a necessary component or end point of a comprehensive regulatory regime, or that other forms of regulation are necessarily insufficient or instrumental. There is, of course, no law that stipulates (and certainly no evidence consistent with any law that suggests) that regulation evolves in any monotonic fashion from less to more efficient. Yet regulatory reform and deregulation have occurred in many domains (Greenstone 2009). There is no unidirectionality to regulation. Nor is there any systematic historical or empirical evidence for any such unidirectionality.

References

Ahmed, N.; Wahed, M.; and Thompson, N. C. 2023. The growing influence of industry in AI research. *Science*. Publisher: American Association for the Advancement of Science.

AI Safety Institute; and Department for Science, Innovation & Technology. 2024. AI Safety Institute releases new AI safety evaluations platform. https://www.gov.uk/government/news/ai-safety-institute-releases-new-ai-safety-evaluations-platform. Accessed: 2024-07-25.

Allen, D.; Hubbard, S.; Lim, W.; Stanger, A.; Wagman, S.; and Zalesne, K. 2024. A Roadmap for Governing Ai: Technology Governance and Power Sharing Liberalism. In *Harvard Ash Center for Democratic Governance and Innovation*.

Anderljung, M.; Barnhart, J.; Korinek, A.; Leung, J.; O'Keefe, C.; Whittlestone, J.; Avin, S.; Brundage, M.; Bullock, J.; Cass-Beggs, D.; Chang, B.; Collins, T.; Fist, T.; Hadfield, G.; Hayes, A.; Ho, L.; Hooker, S.; Horvitz, E.; Kolt, N.; Schuett, J.; Shavit, Y.; Siddarth, D.; Trager, R.; and Wolf, K. 2023a. Frontier AI Regulation: Managing Emerging Risks to Public Safety. arXiv:2307.03718.

Anderljung, M.; Smith, E. T.; O'Brien, J.; Soder, L.; Bucknall, B.; Bluemke, E.; Schuett, J.; Trager, R.; Strahm, L.; and Chowdhury, R. 2023b. Towards Publicly Accountable Frontier LLMs: Building an External Scrutiny Ecosystem under the ASPIRE Framework. arXiv:2311.14711.

Anthropic. 2023. Anthropic's Responsible Scaling Policy, Version 1.0. https://www.anthropic.com/news/anthropics-responsible-scaling-policy.

Bailey, R. 2023. OpenAI chief Sam Altman wants an FDA-style agency for artificial intelligence. https://reason.com/2023/05/16/openai-chief-sam-altman-wants-an-fda-style-agency-for-artificial-intelligence/. Section: Science & Technology.

Bates, S.; Jordan, M. I.; Sklar, M.; and Soloff, J. A. 2024. Incentive-Theoretic Bayesian Inference for Collaborative Science. arXiv:2307.03748.

Berglund, L.; Stickland, A. C.; Balesni, M.; Kaufmann, M.; Tong, M.; Korbak, T.; Kokotajlo, D.; and Evans, O. 2023. Taken out of context: On measuring situational awareness in LLMs. arXiv:2309.00667.

Birhane, A.; Steed, R.; Ojewale, V.; Vecchione, B.; and Raji, I. D. 2024. AI auditing: The Broken Bus on the Road to AI Accountability. 612–643. IEEE Computer Society. ISBN 9798350349504.

Blumenthal, R.; and Hawley, J. 2023. Bipartisan Framework for U.S. AI Act. https://www.blumenthal.senate.gov/imo/media/doc/09072023bipartisanaiframework.pdf. Accessed: 2024-03-16.

Bommasani, R. 2023. Drawing Lines: Tiers for Foundation Models. https://crfm.stanford.edu/2023/11/18/tiers.html.

Bommasani, R.; Creel, K. A.; Kumar, A.; Jurafsky, D.; and Liang, P. S. 2022. Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization? *Advances in Neural Information Processing Systems*, 35: 3663–3678.

Bommasani, R.; Klyman, K.; Longpre, S.; Kapoor, S.; Maslej, N.; Xiong, B.; Zhang, D.; and Liang, P. 2023a. The Foundation Model Transparency Index. arXiv:2310.12941.

Bommasani, R.; Soylu, D.; Liao, T. I.; Creel, K. A.; and Liang, P. 2023b. Ecosystem Graphs: The Social Footprint of Foundation Models. arXiv:2303.15772.

Budish, E.; Roin, B. N.; and Williams, H. 2015. Do Firms Underinvest in Long-Term Research? Evidence from Cancer Clinical Trials. *American Economic Review*, 105(7): 2044–2085.

Callander, S.; and Harstad, B. 2015. Experimentation in Federal Systems *. *The Quarterly Journal of Economics*, 130(2): 951–1002.

Carpenter, D. 2010. *Reputation and Power: Organizational Image and Pharmaceutical Regulation at the FDA*. Princeton: Princeton University Press.

Carpenter, D. 2014. Can Expedited FDA Drug Approval Without Expedited Follow-up Be Trusted? *JAMA Internal Medicine*, 174(1): 95–97.

Carpenter, D.; Grimmer, J.; and Lomazoff, E. 2010. Approval regulation and endogenous consumer confidence: Theory and analogies to licensing, safety, and financial regulation. *Regulation & Governance*, 4(4): 383–407. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1748-5991.2010.01091.x.

Carpenter, D.; Moffitt, S. I.; Moore, C. D.; Rynbrandt, R. T.; Ting, M. M.; Yohai, I.; and Zucker, E. J. 2010. Early Entrant Protection in Approval Regulation: Theory and Evidence from FDA Drug Review. *The Journal of Law, Economics, and Organization*, 26(3): 515–545.

Carpenter, D.; and Ting, M. M. 2007. Regulatory Errors with Endogenous Agendas. *American Journal of Political Science*, 51(4): 835–852. Publisher: [Midwest Political Science Association, Wiley].

Carpenter, D. P. 2004. Protection without Capture: Product Approval by a Politically Responsive, Learning Regulator. *American Political Science Review*, 98(4): 613–631.

Casper, S.; Ezell, C.; Siegmann, C.; Kolt, N.; Curtis, T. L.; Bucknall, B.; Haupt, A.; Wei, K.; Scheurer, J.; Hobbhahn, M.; Sharkey, L.; Krishna, S.; Von Hagen, M.; Alberti, S.; Chan, A.; Sun, Q.; Gerovitch, M.; Bau, D.; Tegmark, M.; Krueger, D.; and Hadfield-Menell, D. 2024. Black-Box Access is Insufficient for Rigorous AI Audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 2254–2272. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Casper, S.; Lin, J.; Kwon, J.; Culp, G.; and Hadfield-Menell, D. 2023. Explore, Establish, Exploit: Red Teaming Language Models from Scratch. arXiv:2306.09442.

Cass-Beggs, D. 2024. A Welcome Voice for Canada on the Future of AI | TechPolicy.Press. https://techpolicy.press/a-welcome-voice-for-canada-on-the-future-of-ai. Accessed: 2024-05-02.

Center for AI Policy. 2024. Model Legislation: Responsible Advanced AI Act. https://www.aipolicy.us/work/model.

Chan, A.; Ezell, C.; Kaufmann, M.; Wei, K.; Hammond, L.; Bradley, H.; Bluemke, E.; Rajkumar, N.; Krueger, D.; Kolt, N.; Heim, L.; and Anderljung, M. 2024. Visibility into AI Agents. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 958–973. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; Ye, W.; Zhang, Y.; Chang, Y.; Yu, P. S.; Yang, Q.; and Xie, X. 2024. A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, 15(3): 39:1–39:45.

Cihon, P. 2019. Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development. https://cdn.governance.ai/Standards_-FHI-Technical-Report.pdf. Accessed: 2024-05-14.

Clymer, J.; Gabrieli, N.; Krueger, D.; and Larsen, T. 2024. Safety Cases: How to Justify the Safety of Advanced AI Systems. arXiv:2403.10462.

Cohen, M. K.; Kolt, N.; Bengio, Y.; Hadfield, G. K.; and Russell, S. 2024. Regulating advanced artificial agents. *Science*, 384(6691): 36–38. Publisher: American Association for the Advancement of Science.

Cottier, B. 2023. Trends in the Dollar Training Cost of Machine Learning Systems. https://epochai.org/blog/trends-in-the-dollar-training-cost-of-machine-learning-systems.

Council of the European Union. 2024. 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://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf. Accessed: 2024-03-16.

Critch, A.; and Russell, S. 2023. TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI. arXiv:2306.06924.

Cunningham, H.; Ewart, A.; Riggs, L.; Huben, R.; and Sharkey, L. 2023. Sparse Autoencoders Find Highly Interpretable Features in Language Models. arXiv:2309.08600.

Davidson, T.; Denain, J.-S.; Villalobos, P.; and Bas, G. 2023. AI capabilities can be significantly improved without expensive retraining. arXiv:2312.07413.

Deng, J.; Pang, S.; Chen, Y.; Xia, L.; Bai, Y.; Weng, H.; and Xu, W. 2024. SOPHON: Non-Fine-Tunable Learning to Restrain Task Transferability For Pre-trained Models. 250–250. IEEE Computer Society. ISBN 9798350331301. ISSN: 2375-1207.

Department for Science, Innovation & Technology. 2024. Introducing the AI Safety Institute. https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute. Accessed: 2024-05-02.

Djankov, S.; La Porta, R.; Lopez-de Silanes, F.; and Shleifer, A. 2002. The Regulation of Entry*. *The Quarterly Journal of Economics*, 117(1): 1–37.

Eiras, F.; Petrov, A.; Vidgen, B.; Schroeder, C.; Pizzati, F.; Elkins, K.; Mukhopadhyay, S.; Bibi, A.; Purewal, A.; Botos, C.; Steibel, F.; Keshtkar, F.; Barez, F.; Smith, G.; Guadagni, G.; Chun, J.; Cabot, J.; Imperial, J.; Nolazco, J. A.; Landay, L.; Jackson, M.; Torr, P. H. S.; Darrell, T.; Lee, Y.; and Foerster, J. 2024. Risks and Opportunities of Open-Source Generative AI. arXiv:2405.08597.

Elhage, N.; Hume, T.; Olsson, C.; Schiefer, N.; Henighan, T.; Kravec, S.; Hatfield-Dodds, Z.; Lasenby, R.; Drain, D.; Chen, C.; Grosse, R.; McCandlish, S.; Kaplan, J.; Amodei, D.; Wattenberg, M.; and Olah, C. 2022. Toy Models of Superposition. *Transformer Circuits Thread*.

Encode Justice; and Future of Life Institute. 2023. AI Licensing for a Better Future: On Addressing Both Present Harms and Emerging Threats. https://futureoflife.org/open-letter/ai-policy-for-a-better-future-on-addressing-both-present-harms-and-emerging-threats/.

Executive Office of the President. 2023. Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence. Accessed: 2024-03-16.

Feffer, M.; Sinha, A.; Lipton, Z. C.; and Heidari, H. 2024. Red-Teaming for Generative AI: Silver Bullet or Security Theater? arXiv:2401.15897.

Fleming, T. R. 2005. Surrogate endpoints and FDA's accelerated approval process. *Health Affairs (Project Hope)*, 24(1): 67–78.

Frame, W. S.; Gerardi, K.; and Willen, P. 2015. The Failure of Supervisory Stress Testing: Fannie Mae, Freddie Mac, and OFHEO. SSRN:2593492.

Ganguli, D.; Lovitt, L.; Kernion, J.; Askell, A.; Bai, Y.; Kadavath, S.; Mann, B.; Perez, E.; Schiefer, N.; Ndousse, K.; Jones, A.; Bowman, S.; Chen, A.; Conerly, T.; Das-Sarma, N.; Drain, D.; Elhage, N.; El-Showk, S.; Fort, S.; Hatfield-Dodds, Z.; Henighan, T.; Hernandez, D.; Hume, T.; Jacobson, J.; Johnston, S.; Kravec, S.; Olsson, C.; Ringer, S.; Tran-Johnson, E.; Amodei, D.; Brown, T.; Joseph, N.; McCandlish, S.; Olah, C.; Kaplan, J.; and Clark, J. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. arXiv:2209.07858.

Ganguli, D.; Schiefer, N.; Favaro, M.; and Clark, J. 2023. Challenges in evaluating AI systems. https://www.anthropic.com/index/evaluating-ai-systems.

Gilbert, T. K.; Lambert, N.; Dean, S.; Zick, T.; Snoswell, A.; and Mehta, S. 2023. Reward Reports for Reinforcement Learning. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 84–130. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702310.

Golchin, S.; and Surdeanu, M. 2023. Time Travel in LLMs: Tracing Data Contamination in Large Language Models.

Gorwa, R.; and Veale, M. 2024. Moderating Model Marketplaces: Platform Governance Puzzles for AI Intermediaries. arXiv:2311.12573.

Greenstone, M. 2009. Toward a culture of persistent regulatory experimentation and evaluation. In Project, T., ed., *New perspectives on regulation*, 116–19. New York: Cambridge University Press), 111.

Guha, N.; Lawrence, C.; Gailmard, L. A.; Rodolfa, K.; Surani, F.; Bommasani, R.; Raji, I.; Cuéllar, M.-F.; Honigsberg, C.; Liang, P.; and Ho, D. E. 2023. AI Regulation Has Its Own Alignment Problem: The Technical and Institutional Feasibility of Disclosure, Registration, Licensing, and Auditing. SSRN:4634443.

Hacker, P.; Engel, A.; and Mauer, M. 2023. Regulating Chat-GPT and other Large Generative AI Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 1112–1123. New York,

NY, USA: Association for Computing Machinery. ISBN 9798400701924.

Heim, L.; Fist, T.; Egan, J.; Huang, S.; Zekany, S.; Trager, R.; Osborne, M. A.; and Zilberman, N. 2024. Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation. arXiv:2403.08501.

Henry, E.; Loseto, M.; and Ottaviani, M. 2022. Regulation with Experimentation: Ex Ante Approval, Ex Post Withdrawal, and Liability. *Management Science*, 68(7): 5330–5347. Publisher: INFORMS.

Henry, E.; and Ottaviani, M. 2019. Research and the Approval Process: The Organization of Persuasion. *American Economic Review*, 109(3): 911–955.

Herrmann, J.; Lenihan, D.; Armenian, S.; Barac, A.; Blaes, A.; Cardinale, D.; Carver, J.; Dent, S.; Ky, B.; Lyon, A. R.; López-Fernández, T.; Fradley, M. G.; Ganatra, S.; Curigliano, G.; Mitchell, J. D.; Minotti, G.; Lang, N. N.; Liu, J. E.; Neilan, T. G.; Nohria, A.; O'Quinn, R.; Pusic, I.; Porter, C.; Reynolds, K. L.; Ruddy, K. J.; Thavendiranathan, P.; and Valent, P. 2022. Defining cardiovascular toxicities of cancer therapies: an International Cardio-Oncology Society (IC-OS) consensus statement. *European Heart Journal*, 43(4): 280–299.

Ho, A.; Besiroglu, T.; Erdil, E.; Owen, D.; Rahman, R.; Guo, Z. C.; Atkinson, D.; Thompson, N.; and Sevilla, J. 2024. Algorithmic progress in language models. arXiv:2403.05812.

Ho, L.; Barnhart, J.; Trager, R.; Bengio, Y.; Brundage, M.; Carnegie, A.; Chowdhury, R.; Dafoe, A.; Hadfield, G.; Levi, M.; and Snidal, D. 2023. International Institutions for Advanced AI. arXiv:2307.04699.

Hubinger, E.; Denison, C.; Mu, J.; Lambert, M.; Tong, M.; MacDiarmid, M.; Lanham, T.; Ziegler, D. M.; Maxwell, T.; Cheng, N.; Jermyn, A.; Askell, A.; Radhakrishnan, A.; Anil, C.; Duvenaud, D.; Ganguli, D.; Barez, F.; Clark, J.; Ndousse, K.; Sachan, K.; Sellitto, M.; Sharma, M.; DasSarma, N.; Grosse, R.; Kravec, S.; Bai, Y.; Witten, Z.; Favaro, M.; Brauner, J.; Karnofsky, H.; Christiano, P.; Bowman, S. R.; Graham, L.; Kaplan, J.; Mindermann, S.; Greenblatt, R.; Shlegeris, B.; Schiefer, N.; and Perez, E. 2024. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv:2401.05566.

Hwang, T. J.; Carpenter, D.; Lauffenburger, J. C.; Wang, B.; Franklin, J. M.; and Kesselheim, A. S. 2016. Failure of investigational drugs in late-stage clinical development and publication of trial results. *JAMA Internal Medicine*, 176(12): 1826–1833.

Hwang, T. J.; Kesselheim, A. S.; and Bourgeois, F. T. 2014. Postmarketing trials and pediatric device approvals. *Pediatrics*, 133(5): e1197–1202.

Kapoor, S.; Bommasani, R.; Klyman, K.; Longpre, S.; Ramaswami, A.; Cihon, P.; Hopkins, A.; Bankston, K.; Biderman, S.; Bogen, M.; Chowdhury, R.; Engler, A.; Henderson, P.; Jernite, Y.; Lazar, S.; Maffulli, S.; Nelson, A.; Pineau, J.; Skowron, A.; Song, D.; Storchan, V.; Zhang, D.; Ho, D. E.; Liang, P.; and Narayanan, A. 2024. On the Societal Impact of Open Foundation Models. arXiv:2403.07918.

Kay, J.; and King, M. 2020. *Radical Uncertainty: Decision-Making Beyond the Numbers*. New York, NY: W. W. Norton & Company, first edition edition. ISBN 978-1-324-00477-6.

Keating, P.; and Cambrosio, A. 2014. *Cancer on Trial: Oncology as a New Style of Practice*. Chicago, IL: University of Chicago Press. ISBN 978-0-226-14304-0.

Kinniment, M.; Koba Sato, L. J.; Du, H.; Goodrich, B.; Hasin, M.; Chan, L.; Miles, L. H.; Lin, T. R.; Wijk, H.; Burget, J.; Ho, A.; Barnes, E.; and Christiano, P. 2023. Evaluating Language-Model Agents on Realistic Autonomous Tasks. https://evals.alignment.org/language-model-pilot-report.

Knight, F. H. 1921. *Risk, Uncertainty and Profit*. Boston and New York: Houghton, Mifflin and Company.

Kolt, N.; Anderljung, M.; Barnhart, J.; Brass, A.; Esvelt, K.; Hadfield, G. K.; Heim, L.; Rodriguez, M.; Sandbrink, J. B.; and Woodside, T. 2024. Responsible Reporting for Frontier AI Development. arXiv:2404.02675.

Laux, J.; Wachter, S.; and Mittelstadt, B. 2024. Three pathways for standardisation and ethical disclosure by default under the European Union Artificial Intelligence Act. *Computer Law & Security Review*, 53: 105957.

Li, K.; Hopkins, A. K.; Bau, D.; Viégas, F.; Pfister, H.; and Wattenberg, M. 2024a. Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. arXiv:2210.13382.

Li, N.; Pan, A.; Gopal, A.; Yue, S.; Berrios, D.; Gatti, A.; Li, J. D.; Dombrowski, A.-K.; Goel, S.; Phan, L.; Mukobi, G.; Helm-Burger, N.; Lababidi, R.; Justen, L.; Liu, A. B.; Chen, M.; Barrass, I.; Zhang, O.; Zhu, X.; Tamirisa, R.; Bharathi, B.; Khoja, A.; Zhao, Z.; Herbert-Voss, A.; Breuer, C. B.; Zou, A.; Mazeika, M.; Wang, Z.; Oswal, P.; Liu, W.; Hunt, A. A.; Tienken-Harder, J.; Shih, K. Y.; Talley, K.; Guan, J.; Kaplan, R.; Steneker, I.; Campbell, D.; Jokubaitis, B.; Levinson, A.; Wang, J.; Qian, W.; Karmakar, K. K.; Basart, S.; Fitz, S.; Levine, M.; Kumaraguru, P.; Tupakula, U.; Varadharajan, V.; Shoshitaishvili, Y.; Ba, J.; Esvelt, K. M.; Wang, A.; and Hendrycks, D. 2024b. The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning. arXiv:2403.03218.

Libgober, B.; and Carpenter, D. 2024. Lawyers as Lobbyists: Regulatory Advocacy in American Finance. *Perspectives on Politics*, 1–20.

Lyon, A. R.; Dent, S.; Stanway, S.; Earl, H.; Brezden-Masley, C.; Cohen-Solal, A.; Tocchetti, C. G.; Moslehi, J. J.; Groarke, J. D.; Bergler-Klein, J.; Khoo, V.; Tan, L. L.; Anker, M. S.; von Haehling, S.; Maack, C.; Pudil, R.; Barac, A.; Thavendiranathan, P.; Ky, B.; Neilan, T. G.; Belenkov, Y.; Rosen, S. D.; Iakobishvili, Z.; Sverdlov, A. L.; Hajjar, L. A.; Macedo, A. V. S.; Manisty, C.; Ciardiello, F.; Farmakis, D.; de Boer, R. A.; Skouri, H.; Suter, T. M.; Cardinale, D.; Witteles, R. M.; Fradley, M. G.; Herrmann, J.; Cornell, R. F.; Wechelaker, A.; Mauro, M. J.; Milojkovic, D.; de Lavallade, H.; Ruschitzka, F.; Coats, A. J. S.; Seferovic, P. M.; Chioncel, O.; Thum, T.; Bauersachs, J.; Andres, M. S.; Wright, D. J.; López-Fernández, T.; Plummer, C.; and Lenihan, D. 2020. Baseline cardiovascular risk assessment in

cancer patients scheduled to receive cardiotoxic cancer therapies: a position statement and new risk assessment tools from the Cardio-Oncology Study Group of the Heart Failure Association of the European Society of Cardiology in collaboration with the International Cardio-Oncology Society. *European Journal of Heart Failure*, 22(11): 1945–1960.

Malgieri, G.; and Pasquale, F. 2024. Licensing high-risk artificial intelligence: Toward ex ante justification for a disruptive technology. *Computer Law & Security Review*, 52: 105899.

Marcus, G. 2023. Replies to Senate Queries. https://www.judiciary.senate.gov/imo/media/doc/2023-05-16_-_qfr_responses_-_marcus.pdf. Accessed: 2024-03-16.

Marks, H. M. 1997. *The Progress of Experiment: Science and Therapeutic Reform in the United States, 1900-1990*. Cambridge University Press. ISBN 978-0-521-78561-7. Google-Books-ID: j84gdplK7c0C.

Marks, S.; Rager, C.; Michaud, E. J.; Belinkov, Y.; Bau, D.; and Mueller, A. 2024. Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models. arXiv:2403.19647.

Matheny, J. 2023. A Model for Regulating AI. https://www.rand.org/pubs/commentary/2023/08/a-model-for-regulating-ai.html.

McClellan, A. 2022. Experimentation and Approval Mechanisms. *Econometrica*, 90(5): 2215–2247. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA17021.

McCraw, T. K. 1986. *Prophets of Regulation: Charles Francis Adams; Louis D. Brandeis; James M. Landis; Alfred E. Kahn*. Cambridge, Mass.: Belknap Press. ISBN 978-0-674-71608-7.

McGregor, S. 2021. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17): 15458–15463. Number: 17.

METR. 2024. Portable Evaluation Tasks via the METR Task Standard. https://metr.org/blog/2024-02-29-metr-task-standard/. Accessed: 2024-05-02.

Microsoft. 2023. Governing AI: A Blueprint for the Future. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW14Gtw.

Moore, T. J.; and Furberg, C. D. 2014. Development times, clinical testing, postmarket follow-up, and safety risks for the new drugs approved by the US food and drug administration: the class of 2008. *JAMA internal medicine*, 174(1): 90–95.

Naci, H.; Smalley, K. R.; and Kesselheim, A. S. 2017. Characteristics of Preapproval and Postapproval Studies for Drugs Granted Accelerated Approval by the US Food and Drug Administration. *JAMA*, 318(7): 626–636.

Naihin, S.; Atkinson, D.; Green, M.; Hamadi, M.; Swift, C.; Schonholtz, D.; Kalai, A. T.; and Bau, D. 2023. Testing Language Model Agents Safely in the Wild. arXiv:2311.10538.

Nevo, S.; Lahav, D.; Karpur, A.; Alstott, J.; and Matheny, J. 2023. Securing Artificial Intelligence Model Weights: Interim Report. Technical report, RAND Corporation.

Ngo, R.; Chan, L.; and Mindermann, S. 2024. The Alignment Problem from a Deep Learning Perspective.

Ojewale, V.; Steed, R.; Vecchione, B.; Birhane, A.; and Raji, I. D. 2024. Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling. arXiv:2402.17861.

OpenAI. 2023. Preparedness Framework (Beta). https://openai.com/preparedness. Accessed: 2024-05-02.

Oren, Y.; Meister, N.; Chatterji, N. S.; Ladhak, F.; and Hashimoto, T. 2023. Proving Test Set Contamination in Black-Box Language Models.

Ottaviani, M.; and Wickelgren, A. L. 2023. Approval regulation and learning, with application to timing of merger control. *The Journal of Law, Economics, and Organization*, ewac025.

Pal, K.; Bau, D.; and Miller, R. J. 2024. Model Lakes. arXiv:2403.02327.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, 1–22. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701320.

Perez, E.; Huang, S.; Song, F.; Cai, T.; Ring, R.; Aslanides, J.; Glaese, A.; McAleese, N.; and Irving, G. 2022. Red Teaming Language Models with Language Models. arXiv:2202.03286.

Pilz, K.; and Heim, L. 2023. Compute at Scale: A Broad Investigation into the Data Center Industry. arXiv:2311.02651.

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2023. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! arXiv:2310.03693.

Rando, J.; Paleka, D.; Lindner, D.; Heim, L.; and Tramèr, F. 2022. Red-Teaming the Stable Diffusion Safety Filter. arXiv:2210.04610.

Sastry, G.; Heim, L.; Belfield, H.; Anderljung, M.; Brundage, M.; Hazell, J.; O'Keefe, C.; Hadfield, G. K.; Ngo, R.; Pilz, K.; Gor, G.; Bluemke, E.; Shoker, S.; Egan, J.; Trager, R. F.; Avin, S.; Weller, A.; Bengio, Y.; and Coyle, D. 2024. Computing Power and the Governance of Artificial Intelligence. arXiv:2402.08797.

Schaeffer, R.; Miranda, B.; and Koyejo, S. 2023. Are Emergent Abilities of Large Language Models a Mirage? *Advances in Neural Information Processing Systems*, 36: 55565–55581.

Schuett, J. 2023. Defining the scope of AI regulations. *Law, Innovation and Technology*, 15(1): 60–82.

Schuett, J.; Anderljung, M.; Carlier, A.; Koessler, L.; and Garfinkel, B. 2024. From Principles to Rules: A Regulatory Approach for Frontier AI. arXiv:2407.07300.

Seger, E.; Dreksler, N.; Moulange, R.; Dardaman, E.; Schuett, J.; Wei, K.; Winter, C.; Arnold, M.; hEigeartaigh, S. O.; Korinek, A.; Anderljung, M.; Bucknall, B.; Chan, A.; Stafford, E.; Koessler, L.; Ovadya, A.; Garfinkel, B.; Bluemke, E.; Aird, M.; Levermore, P.; Hazell, J.; and Gupta,

A. 2023. Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives. arXiv:2311.09227.

Sevilla, J.; Heim, L.; Ho, A.; Besiroglu, T.; Hobbhahn, M.; and Villalobos, P. 2022. Compute Trends Across Three Eras of Machine Learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Sharkey, L.; Ghuidhir, C. N.; Braun, D.; Scheurer, J.; Balesni, M.; Bushnaq, L.; Stix, C.; and Hobbhahn, M. 2024. A Causal Framework for AI Regulation and Auditing. Publisher: Preprints.

Shavit, Y. 2023. What does it take to catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring. arXiv:2303.11341.

Sheshadri, A.; Ewart, A.; Guo, P.; Lynch, A.; Wu, C.; Hebbar, V.; Sleight, H.; Stickland, A. C.; Perez, E.; Hadfield-Menell, D.; and Casper, S. 2024. Targeted Latent Adversarial Training Improves Robustness to Persistent Harmful Behaviors in LLMs. arXiv:2407.15549.

Shevlane, T. 2024. Structured Access: An Emerging Paradigm for Safe AI Deployment. In Bullock, J. B.; Chen, Y.-C.; Himmelreich, J.; Hudson, V. M.; Korinek, A.; Young, M. M.; and Zhang, B., eds., *The Oxford Handbook of AI Governance*, 0. Oxford University Press. ISBN 978-0-19-757932-9.

Shevlane, T.; Farquhar, S.; Garfinkel, B.; Phuong, M.; Whittlestone, J.; Leung, J.; Kokotajlo, D.; Marchal, N.; Anderljung, M.; Kolt, N.; Ho, L.; Siddarth, D.; Avin, S.; Hawkins, W.; Kim, B.; Gabriel, I.; Bolina, V.; Clark, J.; Bengio, Y.; Christiano, P.; and Dafoe, A. 2023. Model evaluation for extreme risks. arXiv:2305.15324.

Shimbun, Y. 2023. Japan Govt to Establish AI Safety Institute in January. https://japannews.yomiuri.co.jp/politics/politics-government/20231221-157027/. Accessed: 2024-05-12.

Smith, G. 2024. Licensing Frontier AI Development: Legal Considerations and Best Practices. *The Lawfare Institute website (2024)*. Publisher: The Lawfare Institute.

Solaiman, I. 2023. The Gradient of Generative AI Release: Methods and Considerations. arXiv:2302.04844.

Srivastava, A.; Rastogi, A.; Rao, A.; Shoeb, A. A. M.; Abid, A.; Fisch, A.; Brown, A. R.; Santoro, A.; Gupta, A.; Garriga-Alonso, A.; Kluska, A.; Lewkowycz, A.; Agarwal, A.; Power, A.; Ray, A.; Warstadt, A.; Kocurek, A. W.; Safaya, A.; Tazarv, A.; and ... Wu, Z. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. arXiv:2206.04615.

Stein, M.; and Dunlop, C. 2023. Safe before sale. https://www.adalovelaceinstitute.org/report/safe-before-sale/. Accessed: 2024-03-12.

Stevenson, M. T. 2023. Cause, Effect, and the Structure of the Social World. SSRN:4445710.

Sunstein, C. R. 2009. Worst-Case Scenarios. In *Worst-Case Scenarios*. Harvard University Press. ISBN 978-0-674-03353-5.

Sunstein, C. R. 2023. Knightian Uncertainty. SSRN:4662711.

Taleb, N. N. 2014. *Antifragile: Things That Gain from Disorder*. New York: Random House Trade Paperbacks, reprint edition edition. ISBN 978-0-8129-7968-8.

Templeton, A.; Conerly, T.; Marcus, J.; Lindsey, J.; Bricken, T.; Chen, B.; Pearce, A.; Citro, C.; Ameisen, E.; Jones, A.; Cunningham, H.; Turner, N. L.; McDougall, C.; MacDiarmid, M.; Freeman, C. D.; Sumers, T. R.; Rees, E.; Batson, J.; Jermyn, A.; Carter, S.; Olah, C.; and Henighan, T. 2024. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread*.

The National Artificial Intelligence Advisory Committee. 2023. RECOMMENDATION: Improve Monitoring of Emerging Risks from AI through Adverse Event Reporting. https://ai.gov/wp-content/uploads/2023/12/Recommendation_Improve-Monitoring-of-Emerging-Risks-from-AI-through-Adverse-Event-Reporting.pdf. Accessed: 2024-03-21.

Thierer, A. 2023. Flexible, Pro-Innovation Governance Strategies for Artificial Intelligence. Technical Report 283, R Street Institute.

Thierer, A.; and Chilson, N. 2023. The Problem with AI Licensing & an "FDA for Algorithms". https://fedsoc.org/commentary/fedsoc-blog/the-problem-with-ai-licensing-an-fda-for-algorithms.

Trager, R.; Harack, B.; Reuel, A.; Carnegie, A.; Heim, L.; Ho, L.; Kreps, S.; Lall, R.; Larter, O.; hEigeartaigh, S. O.; Staffell, S.; and Villalobos, J. J. 2023. International Governance of Civilian AI: A Jurisdictional Certification Approach. arXiv:2308.15514.

Tutt, A. 2016. An FDA for Algorithms. *SSRN Electronic Journal*.

U.S. Department of Commerce. 2023. At the Direction of President Biden, Department of Commerce to Establish U.S. Artificial Intelligence Safety Institute to Lead Efforts on AI Safety. https://www.commerce.gov/news/press-releases/2023/11/direction-president-biden-department-commerce-establish-us-artificial. Accessed: 2024-05-02.

U.S. Department of Commerce. 2024. U.S. and UK Announce Partnership on Science of AI Safety. https://www.commerce.gov/news/press-releases/2024/04/us-and-uk-announce-partnership-science-ai-safety. Accessed: 2024-05-02.

Volden, C.; Ting, M. M.; and Carpenter, D. P. 2008. A Formal Model of Learning and Policy Diffusion. *The American Political Science Review*, 102(3): 319–332. Publisher: [American Political Science Association, Cambridge University Press].

Wallach, J. D.; Egilman, A. C.; Dhruva, S. S.; McCarthy, M. E.; Miller, J. E.; Woloshin, S.; Schwartz, L. M.; and Ross, J. S. 2018. Postmarket studies required by the US Food and Drug Administration for new drugs and biologics approved between 2009 and 2012: cross sectional analysis. *BMJ*, 361: k2031. Publisher: British Medical Journal Publishing Group Section: Research.

Wang, K. R.; Variengien, A.; Conmy, A.; Shlegeris, B.; and Steinhardt, J. 2022. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small.

Weidinger, L.; Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; Griffin, C.; Bariach, B.; Gabriel, I.; Rieser, V.; and Isaac, W. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. arXiv:2310.11986.

Weil, G. 2024. Tort Law as a Tool for Mitigating Catastrophic Risk from Artificial Intelligence. SSRN:4694006.

Wheeler, T. 2024. Licensing AI is not the answer–but it contains the answers. https://policycommons.net/artifacts/11371427/licensing-ai-is-not-the-answer-but-it-contains-the-answers/12260530/. Publisher: Brookings Institution.

Whittaker, M. 2021. The steep cost of capture. *Interactions*, 28(6): 50–55.

Yu, T.; Hsu, Y.-J.; Fain, K. M.; Boyd, C. M.; Holbrook, J. T.; and Puhan, M. A. 2015. Use of surrogate outcomes in US FDA drug approvals, 2003–2012: a survey. *BMJ Open*, 5(11): e007960. Publisher: British Medical Journal Publishing Group Section: Epidemiology.

Zwetsloot, R.; and Dafoe, A. 2019. Thinking About Risks From AI: Accidents, Misuse and Structure. https://www.lawfaremedia.org/article/thinking-about-risks-ai-accidents-misuse-and-structure.