Approval Regulation for Frontier Artificial Intelligence: Pitfalls, Plausibility, Optionality

Daniel Carpenter*

January 12, 2024

Abstract

Observers and practitioners of artificial intelligence (AI) have conjectured the possibility of catastrophic risks associated with its emergence and development, risks that have led some to propose an FDA-style licensing regime for AI. In this essay I explore the applicability of *approval regulation* – that is, a model of product introduction 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 the potential for massively distributed production of foundation models with minimal observability of production. I consider four themes for future theoretical and empirical research: (1) the proper mix of approval regulation and other models such as liability or intellectual property regimes; (2) the possibility that deep ambiguity or Knightian uncertainty may require a kind of *speculative pathology* in which conjecturing scenarios is at least as important as placing probabilities upon them, in part because of *the Lucretius problem*; (3) the likely structure of industry and foundation-model generation, as the feasibility of approval regulation is higher with fewer producers, and much of the future of AI regulation may consist in labs and models monitoring one another; and (4) the possibility of community option value in the incremental development of AI regulation (including approval regulation), as regulatory policies may be more reversible in AI than in other settings, experimentation generates important public goods, and regulatory learning by doing is likely to be a property of any portfolio of policies in this arena.

*Department of Government, Harvard University: dcarpenter@gov.harvard.edu. **This document is preliminary and will with certainty undergo substantial revision; cite with caution.** For helpful conversations and suggestions I acknowledge Markus Anderljung, Sam Bell, Lindsey Gailmard, Holden Karnofsky, Michael Sklar and Glen Weyl. For research support I acknowledge Open Philanthropy. I remain solely responsible for all errors and omissions.

Massive leaps in the scale and power and apparent risks of artificial intelligence (AI) have led many practitioners and observers to call for different forms of regulation. Concerned primarily about general AI models that pose "existential risk" [Carlsmith 2022] [Karnofsky 2022] entailing threats to thousands of human lives, trillions of dollars in economic value or the durability of humanity or earthly life itself, a range of scholars and writers have outlined plans for regulatory policies. Most recently, [Anderljung2023] represents a broad call for risk management in AI, proposing "mechanisms to create and update safety standards," "mechanisms to give regulators visibility" and "mechanisms to ensure compliance with safety standards" ([Anderljung2023], 3, 18-22; see also [Shevlane2023]). In the past few years, practitioners and observers have raised the possibility of various models of regulation that are analogs of regulatory regimes already in existence. Most notably, at a recent congressional hearing, Open AI CEO Sam Altman claimed it "essential to develop regulations that incentivize AI safety while ensuring that people are able to access the technology's many benefits," while 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]. At least one institute has now proposed an explicit licensing regime based upon FDA-style approval regulation, with the idea that foundation models should be "safe before sale" [AdaLovelace2023], and even the National Artificial Intelligence Advisory Committee (NAIAC) has called for an adverse events reporting system, wherein the FDA's system of the same title (it has been known for decades as the AERS) figures as a reference point [NAIAC 2023]. Meanwhile, a more general "FDA for Algorithms" [Tutt 2017] has been proposed for some time.

The possibility of an FDA-like regulatory regime for AI has since occasioned considerable debate, beginning with the introduction of a bill in the United States Senate that would establish a commission to oversee digital platforms [Bennet and Welch 2023].[1] Perhaps the best evidence for the idea that "FDA-like" institutions are being considered for AI regulation is the fact that a range of libertarian organizations and writers have quickly aligned against the idea, one complaining that "OpenAI Chief Sam Altman Wants an FDA-Style Agency for Artificial Intelligence" ([Bailey 2023] [Thierer and Chilson 2023]). As I suggest below, these critics may have it right, but before the idea of such a regime can be considered, one would need to consider what kinds of things it would entail. Consider the following questions.

- What exactly is an "FDA-Style agency" or an "FDA-like regulatory regime"? Is it *any* agency or regime screening process or licensing process? Is it just *any* regulation that requires experimentation and testing? How are screening and experimentation combined, and what other policies are in the mix?
- What are the essential assumptions about information, market structure and institutions that such a regime entails?

[1] As one observer described the bill, "Two days after Altman's testimony, Senators Michael Bennet and Peter Welch introduced a bill that would create a new federal agency to regulate internet platforms like ChatGPT – a kind of FDA for the automated content we consume" [Dugan 2023].

- What are properties of foundation models in AI that fit or do not fit with these assumptions?
- Are there ways that FDA-like regimes can be tried without adopting a wholesale regulatory system? Put differently, is it possible to flexibly experiment with FDA-like institutions? What does historical and comparative experience tell us about institutional adaptation in this space?
- What are the politics and political economy of the industries to which approval regulation has been traditionally applied, and how might these match or differ from those of frontier AI?
- Are there genuine risks of capture in approval regulation, and how might these evolve in the regulation of frontier AI?

In this essay, I 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. The mapping of FDA-like analogies to AI regulation has proceeded by means of vague metaphors – understandable for an early stage of public debate, but not desirable as actual policies are discussed – and there are properties of AI and its risks that would appear to be poor fits (are "inapposite") for FDA-style regulation as it has been traditionally practiced. I proceed through four general claims: (1) at its essential core, FDA-style regulation is a form of *approval regulation* linking mandatory experimentation with a regulatory veto over part or all of a firm's R&D process; (2) this regime of regulation makes specific assumptions about regulated "products" (biological or molecular entities, for instance) and the observability of firm actions (experiments and their results); (3) there are aspects of AI that do not conform to these assumptions, including the presence of deep ambiguity or Knightian uncertainty; and (4) there may be ways of experimenting with approval regulation institutions, such that adopting aspects of approval regulation does not imply a full-fledged commitment to an entire "FDA."

Before proceeding, two prefatory notes. First, I 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. There is a wide debate about that and it is simply beyond the scope of this essay. 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. 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.

1 "FDA-Style" Regulation as Approval Regulation

It is likely that when commentators refer to an "FDA-like" or "FDA-style" regime, they are referring to the way that the FDA regulates 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* ([Ottaviani and Wickelgren 2023] [Henry, Loseto and Ottaviani 2022] [Henry and Ottaviani 2019] [Carpenter, Grimmer and Lomazoff 2010] [Carpenter and Ting 2007] [Carpenter 2004]), which I define here as a regime in which a regulator requires a firm to experiment with a new product before its marketing, and in which this experimentation generates data that is used by the regulator to decide whether part or all of the product can be marketed after experimentation. The experimentation generates observations on state variables associated with the product, e.g., toxicity, equivalence to other products, efficacy and other variables. So defined, approval regulation gives the regulator a "veto" over product development, but approval regulation is far more than a mere gatekeeping function or a veto. Any number of governments regulate "entry" in the sense of requiring some kind of fee, test, form completion or other procedure before a service or commodity may be lawfully be marketed in commerce [Djankov et al, 2002]. So too, the U.S. Food and Drug Administration is an agency that does not engage in approval regulation in some important areas outside of biomedical innovation. It largely inspects food products (many meat products are regulated by the U.S. Department of Agriculture), but these are not generally subject to FDA gatekeeping. The Food and Drug Administration and, beyond that, a range of government organizations around the world require inspections and testing, but many of these tests and inspections are not linked to pre-market review of new products, such as those that occur with manufacturing facilities (carried out by the Occupational Safety and Health Administration (OSHA) in the United States) or many consumer products.

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 the history of these institutions has been the subject of studies in history and political science [Marks 1997] [Carpenter 2010]. It is important to regard the models as simplifications that sacrifice considerable purchase and understanding these institutions, but the models are nonetheless essential for understanding these regimes. In the ensuing years the literature in economic theory and management science has progressed well beyond these simple models. In particular, [Henry and Ottaviani 2019] [Henry, Loseto and Ottaviani 2022] and [Ottaviani and Wickelgren 2023] 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. More general models have since been developed, including in [Henry and Ottaviani 2019] [McClellan 2022], and [Bates et al 2023]. Earlier models tended to take the institutions for granted and to describe likely behavior under them, whereas later models have explored a range of alternative institutional arrangements and the potential tradeoffs or complementarities

among them.[2]

For several reasons, the *combination* of experimentation and veto in approval regulation is essential to understanding these institutions. First, as reviewed above, the kinds of institutions customarily associated with the FDA involve, at their core, this institutional mix, such that "veto + experimentation" differentiates FDA-like and other institutions from other forms of regulation that erect entry barriers. Second and beyond this, a range of other regulatory policies implemented and enforced by agencies such as the FDA and EMA rest upon these two powers. As argued by [Marks 1997] and [Carpenter 2010], the standards of pre-market review at the U.S. Food and Drug Administration developed hand-in-hand with changes in pharmacological and experimental standards. In terms of phased experimentation, developments in oncology (especially at the National Cancer Institute) were critical to the FDA's view of phased experiment ([Keating and Cambrosio 2019] [Carpenter 2010]). Required labeling for biomedical products incorporates information from required experiments, and the proposed labeling is an important part of the pre-market review. The ability of regulators to write new rules governing experimentation depends heavily upon gatekeeping, but 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 by the FDA and EMA.[3]

Of course, the EMA and FDA do many things other than require experiments 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, and 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 to a product before it is authorized for marketing or release (*ex ante* regulation) and the set of things that happen to a product after it is marketed (*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 authorization (the "veto"). Yet important regulatory tools are available after regulatory marketing authorization. The regulator can require or request changes

[2]I emphasize the formal modeling literature not because I think it gets everything or even most things right – my own models most certainly do not – but because it often helps to identify the critical operative structure and incentive-based kernels of institutions, especially when modelers pay appropriate attention to the historical and institutional context of the things they study.

[3]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.

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.

Two important properties of approval regulation at the EMA and FDA are long-range interaction and experimental incentives. First, a single company likely has a range of products, some that are already marketed and others that are under development. [4] 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. This means that even a profitable firm has great 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. Second, 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 experimentation than they otherwise would ([Carpenter and Ting 2007] [Henry and Ottaviani 2019]). Whether this is a good thing or not depends not just on particular regulatory requirements, but also upon what we consider the public good nature of experimental information to be.

2 Feasibility – What Feasible Approval Regulation Requires

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.[5] 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 or 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.

FDA regimes governing medical devices differ considerably from those from molecules, but medical device regulation carries forward many principles and institutions from drug

[4] Most mathematical models do not consider these repeated interactions or the shadow of future interactions ([Carpenter and Ting 2007] [Henry, Loseto and Ottaviani 2022] [Ottaviani and Wickelgren 2023]), but [Carpenter 2004] and [Carpenter, et al. 2010] formalize a history of firm interaction and a "pipeline value" in a decision-theoretic context.

[5] 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.

regulation. Furthermore, pre-market approval (PMA) is required for – Class III devices – the most innovative and risky devices defined under regulation (the 1976 Medical Device Amendments and their associated rules). As with the molecular new drug application (NDA), a regulatory sponsor must file a pre-market approval application to the FDA. And as with much of the architecture governing molecules, required tests include pre-clinical tests, which " information on microbiology, toxicology, immunology, biocompatibility, stress, wear, shelf life, and other laboratory or animal tests," as well as "clinical investigations, which include "study protocols, safety and effectiveness data, adverse reactions and complications, device failures and replacements, patient information, patient complaints, tabulations of data from all individual subjects, results of statistical analyses, and any other information from the clinical investigations" [U.S. FDA 2018] [U.S. FDA 2019]. 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 a regulated firm and the sites of production and innovation.* 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 sub-contracting.

- *Identifiability of (denumerable) adverse events with associated probability measures of their 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 (see for instance the statistical review in an FDA therapeutics review or the statistical computations in any new drug application or new biologic application). While in theory the set of things that

could go wrong is infinite, in practice it is usually quite manageable.[6] 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 conducts upon 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. [7]

- *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. 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 grown 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.

- *Observability of the fact of experiment, once mandated.* In biopharmaceutical regulation the event that "the firm conducts a test upon its product" is highly observable, and in the models of approval regulation ranging from [Carpenter 2004] to [Carpenter and Ting 2007] to [Henry and Ottaviani 2019], this fact is perfectly observable and at a cost known to regulator as well as firm. This fact is in part endogenous to institutions, including regulatory institutions (all drugs under study in the United States must have an approved status of Investigational New Drug (IND)), such that the molecule is registered with the FDA, 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

[6] 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.

[7] It is important to understand just how often we presume that we are in a world of countable additivity with adverse events. In the words of the great Patrick Billingsley, no statistical inference is possible without this basic assumption: "The essential property of probability measures is countable additivity, and this is a condition on the countable *disjoint* unions" ([Billingsley 1995], 43). Billingsley refers here to the requirements of the $\pi - \lambda$ theorem, especially its third requirement (λ_3), viz., $A_1, A_2, \cdots \in \mathcal{L}$ and $A_n \cap A_m = \emptyset$ for $m \neq n$ imply $\cup_n A_n \in \mathcal{L}$, where $\mathcal{L}$ is the λ-system containing Ω and is closed under the formation of complements and of finite and countable disjoint unions. The uncertainty described by [Knight 1921] is one among other scenarios that can violate this assumption.

government), and groups of professional scientists and statisticians who are routinely consulted in the design, pre-registration and analysis of these experiments.

- *Observability of the results of experiments.* As with the identifiability of development of new products and the identifiability of adverse event and probability measures for describing them, there exists in many areas of regulation a set of consensually agreed-upon rules for observing results of experiments, aggregating these results and transforming aggregates into quantities of interest for statistical inference and decision. In all cases these methods presume well-defined probability measures and in many cases, particular distributions or families thereof, with particular logics of inference (Bayesian, frequentist, e.g.) available to analysts and decision makersIn most cases.

- *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 experiment (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. My 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.[8] 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.

[8]I'm taking liberties with simplification here in part because elsewhere [Carpenter 2010] I have described the structure of FDA pharmaceutical regulation in much greater detail. While things have changed since that book, core features of the process – phased experimentation, submission of new product applications, and regulatory veto – remain in place.

3 Pitfalls: Identifying Lack of Fit between Traditional Approval Regulation and Frontier AI

Given these stylized characteristics of approval regulation, especially in the biopharmaceutical realm, I now turn to the emerging field of AI (as described by others) and identify possible points of "mismatch" in the application of FDA-style regulation to frontier AI. Whether the facts adumbrated in the previous section apply to frontier AI regulation is an empirical question. It is possible that the conditions for applicability of approval regulation to biopharmaceuticals are not *yet* satisfied in the area of 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 AI is a fruitful research agenda in institutional design as well as applied governance.*

3.1 Difficulties in Defining a Unified, Homogenous Regulated Product

3.1.1 The Problem

Writers on the regulation of AI have focused upon the existential risks from "frontier AI models," defined as "highly capable foundation models that could exhibit dangerous capabilities" [Anderljung2023]. In the present context, models such as DALL-E or GPT-4 have these properties, though in several years or even several months the frontier will run away from them.[9] Now imagine a world in which a government wishes to set up a licensing regime in which any new foundation model is mandated to undergo tests before being released. A first problem is which models can be called foundation and which not? If the definition of regulated products is too inclusive and the population of regulated products explodes, so then does the aggregate cost of regulating the industry and the possibility that, with attention distributed across so many of them, the risk of Type II inspection errors (failing to detect a risk when it is present) rises accordingly. If the definition of foundational model is too strict, then models that are regarded as insufficiently foundational may escape scrutiny, a different kind of Type II error.

This problem becomes more realistic under at least two scenarios that have been studied or conjectured recently. The first concerns when a frontier model can be approximated through *low-rank adaptation* [Hu et al, 2021] or similar methods, which then raises the question of which adaptations count as products to be regulated. The second concerns the prospect that foundation models become themselves capable of generating new foundation models at the frontier, in what [Ngo, Chan and Mindermann 2023] describes as *recursive self-improvement.* If these second-generation models all deserve consideration as regulated products, then the population of products can again explode and/or its

[9] Writers usefully note that large-language models (LLMs) are not exhaustive of the foundation models that society should be concerned about ([Anderljung2023], 7, fn. 8), given the growing presence of models with visual capabilities (Midjourney, Imagen, Stable Diffusion)

heterogeneity might increase.[10]

The identifiability of homogeneous regulated units is important because approval regulators rely heavily upon the laws of statistics in evaluating biomedical products and health treatments. 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). While modern applied statistics and biostatistics certainly has many methods for dealing with heterogeneity, it is fair to say that the more complicated heterogeneity becomes, the less likely it is that large-sample assumptions apply or that measurement and omitted variable bias will infect analyses and inferences.

The interface between development and public availability or public release also takes different institutional forms with different institutional incentives. In many regulated markets an authorized product is immediately marketed. (Biopharmaceutical products are generally not open source and their revenue-generation models depends upon per-unit sales.) In large language models, many are shared widely before being commercialized. Put differently, there is a difference between *marketing* and *release*, and this difference comes loaded with differential incentives in the short run and long run.

Counterpoint: Stress Tests are not Large-Sample Tests of Homogeneous Product Distributions. In response to these concerns, one might respond that the kind of experimentation that is and will be conducted upon foundation models is more akin to financial stress-testing, that is "red teaming" [Ganguli et al, 2022a] through a set of optimal scaling procedures in which model development occurs while learning about risks, and/or more active "jailbreaking" [Chao, et al. 2023] [Robey 2023] procedures in which the vulnerabilities of foundation models are directly probed. This fact may reduce the weak fitness of approval regulation for foundation models. Yet it might well introduce other issues. For one, no financial regulator of which I am aware requires stress-testing as a pre-requisite to the "release" of a financial product.[11] Hence the mapping from stress tests to approval gates would need to be specified. For another, application of stress-test based methodologies would require a legal uniformity (at least for regulatory minima) for all foundation model developers in the regulated space. Since regulatory settings with lower-cost regulatory requirements might well attract more laboratory activity, and 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.

[10]Similarly, [Guha et al, 2023] (p. 36) note that "recent research suggests that capabilities exhibited by frontier models can be elicited in smaller models through improved algorithmic choices" (citing [Taori, et al., 2023]).

[11]And as we have seen in recent days, financial regulators *do* approve (or decide not to approve) financial products, as in the SEC's decision to allow Bitcoin-based exchange-traded products (ETPs).

3.1.2 Possible Solutions

The problem of defining a regulated product is in part the problem of standardization and in part the problem of model evaluation ([Shevlane2023], [Ganguli et al, 2022a]). Put differently, it may be necessary to define model evaluation for a standardized set of foundation models, converting the regulation problem into the the problems of (1) adapting standardized model evaluation to new variants and (2) the problem of detecting emergent risks.

3.2 Difficulties in Identifying Regulated Organizations (Labs, Producers)

The applicability of approval regulation to any other domain depends upon 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. Or if the model of stress testing for systemically important financial institutions is considered as an analogy or inspiration, the regulated organization would be responsible for carrying out the tests or permitting government or third-party observers access to the data with which they could be performed. In the case of biopharmaceutical regulation, the regulated firm is not necessarily the developing laboratory or even the manufacturer but the "sponsor" ([Carpenter 2010], Chapter 10) which subsequently has responsibility for compliance with manufacturing, quality control and post-market experimental requirements. In the case of 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.[12]

3.2.1 Problem: Heterogenous Regulated Organizations in Which the Developer Differs Heavily from the Sponsor, and these from the Deployer/User

The originators of foundation models are, for the moment and in general, well known, the most prominent example being Open AI and its development of GPT-4 and DALL-E models, and others including Midjourney with Midjourney, Anthropic with Claude 2 and Amazon with Titan. 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

[12]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`.

– 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. If a few, large and well-resourced companies or laboratories dominate the development of the foundation models that are the most promising but also the most threatening in terms of their systemic risk, then regulated organizations will more likely have the organizational and financial capacity to comply with intensive reporting requirements. If, however, low-rank adaptation [Hu et al, 2021] or recursive self-improvement [Ngo, Chan and Mindermann 2023] permit generation of foundation models (or meaningful alterations to those models) at lower cost, then it is possible that smaller "producers" will be involved.

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. The Lovelace Institute recommends that "AI regulators should have strong powers to investigate and require evidence generation from foundation model developers and downstream deployers" ([AdaLovelace2023], 8). Yet if lower-cost adaptation or recursive adaptation is possible, "downstream" organizations may be able to alter foundation models or their products in ways that impose additional risk. And the set of organizations that "deploy" foundation models is potentially massive. Will the same agency that regulates foundation models also have the responsibility of regulating the deployment of those models?

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. 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.

3.2.2 Possible Solutions: Reliance upon Exogenous Industrial Concentration, Direct Regulation of Innovators

There are reasons to think that the future of foundation model development will be characterized by high-cost research and development and by a smaller and smaller 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 (recursively and autonomously, or with human supervised training). As with other capital-intensive industries, then the number of operative firms would be reduced. As Amazon Web Services reports, BERT, "one of the first bidirectional foundation models" and launched in 2018, "was trained using 340 million parameters and a 16 GB training dataset," but just five years later, "Open AI trained GPT-4 using 170 trillion parameters and a 45 GB training dataset" [AWS2023]. And quite some time ago, Open AI estimated that the amount of "compute" used in foundation model development, measured in petaflops per second for a full day, doubled every 3.4 months.[13] Still, some risk may come from the fact that state-sponsored organizations overseas may wish to invest in smaller laboratories or models to develop their own capabilities.

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 [White House 2023] as well as the European Union are moving. The question becomes how enforceable such registration requirements are. Can foundation model development outside of the reporting sphere be detected, whether by means of training runs or energy expenditure? The applicability of approval regulation to AI and foundation model 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.

3.3 A Need for *Speculative Pathology*? Difficulties in Describing, Identifying and Measuring Adverse Events

3.3.1 The Problem: Non-Commensurable Harm, Deep Ambiguity and Radical or Knightian uncertainty

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 encountered with the development of AI is not known. 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 worlds, the AI world seems pregnant with potential risks and rewards that are, almost by forcible extension from of the promise and pitfalls of artificial intelligence, hard to imagine. 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*

[13] See `https://openai.com/research/ai-and-compute`.

yet been imagined, either by machine or by human.

To consider this formally, recall the formal definition of a probability measure as defined by a canonical text [Billingsley 1995] (the following is verbatim from ([Billingsley 1995], 1.2, 22-23).

A set function is a real-valued function defined on some class of subsets of Ω. A set function P is a *probability measure* if it satisfies these conditions:

(i) $0 \leq P(A) \leq 1$ for $A \in \mathcal{F}$;

(ii) $P(\emptyset) = 0$; $P(\Omega) = 1$;

(iii) if $A_1, A_2, ...$ is a disjoint sequence of $\mathcal{F}$-sets and if $\cup_{k=1}^{\infty} A_k \in \mathcal{F}$, then

$$P(\bigcup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} P(A_k) \tag{1}$$

The condition represented in (1) is called *countable additivity* and it is "imposed on the set function P" ([Billingsley 1995], 23). Which is to say that the existence of subsets of events that can be aggregated under the condition of denumerable disjoint unions ([Billingsley 1995], 43) is a precondition for statistical analysis as we know it.

Yet in the examination of regulatory risks – for foundation models or for any other regulated object – the countable additivity of adverse events is not an assumption easily satisfied, and if it is satisfied, *it is satisfied in part by the development of societal, scientific and regulatory architecture for creating such countability where it would not otherwise exist.* It has, in other words, taken a century or more for modern medicine and pharmacology to get to the point where, for most but not all therapeutically administered molecules, the set of risks (possible things that could go wrong) is well known and probabilities can be attached.

The analogy for risk in a biomedical world is pathology and its relationship to epidemiology (or pharmacoepidemiology), where the adverse events are largely known ahead of time and can be counted. Hepatotoxicity that results from consumption of a medicine or other chemical again furnishes an example. The countable adverse event (toxicity) can be observed on an individualized basis (by direct examination of the liver, by imaging or by hematological measurements) and a "case" of such toxicity can be observed and coded both "across" patients (in a multi-person sample) and "within" patients (in the same person over time). From these aggregations and from application of statistical principles, inferences can be drawn from observational and experimental samples. In the world of climate risk, the possibilities are less well known but scientists and policymakers measure them by reference to known measureables, such as moments of the temperature distribution (changes in mean, variance and extrema), projected rise in sea levels, species

extinction rates and the like. For most of these outcomes, there exist known geophysical and population ecology models that can be used to generate statistical predictions (for instance, [Weitzmann 2014] on tail risk in climate change, or [Posner and Weyl 2013] on "the cost of a statistical financial crisis"). In frontier AI, this level of model-based prediction of harms seems more inchoate.

Recent developments in jailbreaking research [Chao, et al. 2023] [Robey 2023] render more realistic the prospect of detecting non-alignment in LLMs. In [Chao, et al. 2023], researchers deploy "an attacker LLM to automatically generate jailbreaks for a separate targeted LLM without human intervention." The process is iterative and the attacker model learns of the target model's vulnerabilities through repeated querying. As [Chao, et al. 2023] reports, their algorithm "often requires fewer than twenty queries to produce a jailbreak, which is orders of magnitude more efficient than existing algorithms. PAIR also achieves competitive jailbreaking success rates and transferability on open and closed-source LLMs, including GPT-3.5/4, Vicuna, and PaLM-2." Aside from the fact that jailbreaking may be easier than imagined, the research of [Chao, et al. 2023] and [Robey 2023] promises to create procedures that will better identify the emergence of non-aligned models and the "adverse event" of non-alignment.

3.3.2 What Does an AI Adverse Event "Look Like"?

Still, the question of risk from foundation models and non-aligned AI is not merely the question of whether misalignment occurs, but also *the potential costs incurred once that barrier is ruptured* (or ruptured with sufficient severity that serious human costs occur). Consider a simplistic model of risk from the insurance world, the compound Poisson process ([Ross 2000], Section 5.4.2, pp. 289 ff). In such a process there are two processes of interest, one the occurrence of the event in question (modeled below by $N(t)$ which is presumed homogeneous and homoscedastic) and the other the non-negative value Y incident to each event.

$$X(t) = \sum_{i=1}^{N(t)} Y_i, \ \forall t \geq 0 \tag{2}$$

where $\{N(t), t \geq 0\}$ describes a Poisson process and $\{Y_i, i \geq t\}$ is a family of independently and identically distributed random variables – in the simplest case, the exponential distribution with cumulative distribution function $F(y) = 1 - e^{-\lambda y}$ and probability density function $f(y) = \lambda e^{-\lambda y}$ – which are also independent of $\{N(t), t \geq 0\}$, which creates a memoryless property of the value distribution.

Far more complicated models are used in actuarial sciences, applied probability theory, biostatistics and finance, of course, but in order for even this most simple model to work, *there must be some basis for estimating the conditional cost distribution* $F(Y_{i(t)}|N(t) = 1)$,

and that requires knowing something about what could go wrong.

In the case of biomedical innovation, many of these risks are far better known, 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 [Hermann 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.

3.3.3 Possible Solutions: Pathologies of the "Yet-to-Happen" to Combat the Lucretius Problem

Some forms of "pathology" and "epidemiology" may be possible in an AI world and are already being used and deployed ([Ngo, Chan and Mindermann 2023], [Ganguli et al, 2022a]). While this exercise (tracking the number of times a foundation model begins to engage in harmful behavior) is possible and may become better developed, it also seem that regulation of risk in this field needs purely descriptive work, a kind of "pathology of the yet-to-happen." *Faute de mieux*, I'll call this *speculative pathology* here. The idea is for those who monitor these risks – where these agents may be both humans and machines – to spin tales about what could go wrong and with what harms. The problem of the "off-switch" game [Hadfield-Menell et al, 2016] is one example, but play in such a game might be shaped by particular assumptions about extensive form, repetition or equilibrium concepts, or might be expanded with multiple agents, multiple models and multiple regulators. While analytic and computational analysis of such games would be useful, so too would narrative. Or consider standard war-gaming exercises regularly engaged in my military and strategic academies. Many of these exercises may be non-formalized and others may involve probabilistic calculations. At least some of them involve deep engagement in narrative, speculation and technological fiction.[14]

While this notion of "speculative pathology" may seem odd, it is worth acknowledging that important developments in the development of risk began descriptively (though not as speculatively). Much of the development of biomedical regulation relied upon

[14]Specialists in cybersecurity I know have recommended the novel *Ghost Fleet: A Novel of the Next World War* ([Singer and Cole 2016]), which rests upon an imagined set of cyberattacks on American infrastructure, use of next-generation weapons by American adversaries and by American forces, and novel intelligence and surveillance strategies (or consider Tom Clancy novels). These narratives are of course fictional but based upon extensive consultation with, and knowledge of, military and technological developments.

the co-evolving discipline of pharmacology, which emerged from studies of acute toxicity and then chronic toxicity in animal models and then clinical (human) settings [Carpenter 2010]. Adverse events such as toxicity had to be described and then measured before statistical analysis of patterns and cause-and-effect could ensue. Similarly, in cancer clinical trials, phased studies developed from the differentiation of different forms of risk, the separation of chronic from acute toxicity and the definitions of "cure," "safety," and "toxicity" [Keating and Cambrosio 2019].

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. [Shevlane2023] write of this problem in an in direct ay when they point to an important limitation of risk-based regulatory frameworks. As they write:

> "1. **Unanticipated Behaviors**. Before deployment, it is impossible to fully anticipate and understand how the model will interact in a complex deployment environment .... For example, users might find new applications for the model or novel prompt engineering strategies; or the model could be operating in dynamic, multi-agent environment."

> "2. **Unknown Threat Models**. It is difficult to anticipate all the different plausible pathways to extreme risk. This will be especially true for highly capable models, which could find creative strategies for achieving their goals."

On the second of these statements, it would better to say that it will be *impossible* to anticipate all such pathways,[15] and assumptions would likely need to be made about subsets of such pathways or subsets of extreme adverse events that are assumed, for sake of analysis, sufficiently commensurable to be included in a common category.[16] What might such subsets look like? Lists of different risks (in some cases, different adverse events) appear in [Carlsmith 2022], [Shevlane2023] (Table 1), [Ngo, Chan and Mindermann 2023] (p. 9), [AdaLovelace2023] (pp. 10-11), and [Guha et al, 2023] (p. 9), Here are four simple examples.

- When would non-alignment lead to algorithms taking over air, ground and water transportation – as hackers apparently do in the movie *Leave the World Behind* (2023) – and what would the distribution of these costs look like (moments, mean, variance, extrema)?
- When would non-alignment lead to algorithms taking over power generation or grids, and what would the concomitant cost distribution be?

[15] Similarly, [Guha et al, 2023] (p. 35) argue that "Machine learning research hasn't developed agreed-upon standards for how to quantify properties like catastrophic risk." This fact applies equally to scenarios of self-regulation, decentralized regulation and centralization regulation, of course.

[16] This is one way of thinking of unions of subsets in the definition of probability measures in [Billingsley 1995], with the added proviso that, sooner or later, they must be (mutually) disjoint in order to be aggregated.

- When would non-alignment lead to the autonomous discovery, generation and release of bioweapons, and what would the concomitant cost distribution be?
- When would non-alignment lead to the appropriation of existing military weapons systems, including nuclear weapons, and what would the attendant distribution of costs look like?

Another reason to consider some kind of speculative pathology in this enterprise – performed by humans, by LLMs and by teams of both – is to avoid 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. Rendered in more probabilistic terms, the conditional cost distribution $F(Y_{i(t)}|N(t) = 1)$ in the "harm aggregation process" listed above may be *non-stationary* in one or more of its moments over very long run. The maxima of the harm distribution might get worse and worse (this is one way of posing the Lucretius problem), or "very bad" but short of the worst event might become more likely through learning. Generative AI might seek to create pathways of risk and materializations of risk, that have never before been imagined. (Put differently, if we as societies or regulators are not willing to do the speculation, AI will do it "for" us.) Or the auxiliary risks from diffusive bioweapons, proliferating nuclear weapons or interconnectedness may exacerbate the harm that could happen from an otherwise stable risk process governing the mis-alignment of foundation models.

3.3.4 Avoiding Speculative Pathology's Own Pitfalls

To be clear, any regulatory regime that deployed speculative pathology would have to avoid implementing the most naïve decision rules. Just because a war-gaming or speculative pathology exercise can produce a horrific imagined result – the end of the world – should not imply that the most restrictive regulatory response should be adopted.[17] Any speculative exercise that included the worst possible scenario would also need to consider humanity's likely best response in addition to regulatory options.

3.4 Difficulties in Observing Development, Deployment and Mandated Experiments

3.4.1 Problem A: Approval Regulation Depends Upon an Enforcement Regime

The emergence of a new frontier AI model may be weakly observable to any external agent, whether a regulator or a competitor. 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 difficult for any third-party agent to observe many

[17] One can imagine formalizing such a naïve aproach in assigning infinite or massive loss values to a mere singleton in the adverse event set, at which point the risks of foundation models might greatly outweigh any benefit from their existence.

acts of a new foundation model being developed or even deployed.

Approval regulation and any other kind of licensing or entry regulation depends upon institutions of detection. The unlicensed barber who violates state licensing statutes faces an enforcement apparatus that depends upon state and local law enforcement. In medicine, there are a range of institutions. A human agent can offer health services to the public, but if the consumer wants insurance to pay for those 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 development and provision. And of course state health agencies and medical authorities, the Drug Enforcement Administration (DEA) and the FDA itself, in the United States, have enforcement capabilities.

3.4.2 Problem B: Approval Regulation Depends Upon an Experimentation Regime

If experiments 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 et al, 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]). More recently, [Brown, et al. 2022] find that more than seven in ten post-marketing commitments and reports for new drug approvals between 2013 and 2016 were "late" by the fourth quarter of 2020. As [Brown, et al. 2022] report, the postmarketing requirements under the most innovative biomedical products, namely for drugs having received accelerated approval, "had the longest median projected times to completion," with a median of ten quarters or two and a half years to completion ("median, 10.00 [IQR, 7.00-15.00] quarters") while new drugs approved under the Pediatric Research Equity Act (PRAE) were completed even more slowly, in roughly three years ("median, 12.00 [IQR, 5.25-14.75] quarters") .

Of greater relevance to AI and foundation model regulation is the fact that, 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]. In the realm of biomedical innovation these products sit on the "shelf" and there is not likely much of a risk of their being seized and deployed for other uses,[18] but in the world of

[18]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 non-aligned purposes.

algorithms there seems little, beside strong cybersecurity protections, to prevent from others from developing such products and potentially putting them to misaligned purposes ([Guha et al, 2023], 77).

3.4.3 Potential Adaptations and Solutions

The scale and size of frontier AI models provides some openings here,[19] though more research is needed. 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). If these expenditures can be measured by regulators or third parties, perhaps using their own LLMs, then development of new foundation models may be detectible. 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.

Other solutions and adaptations likely entail the use of algorithms themselves, as characteristics of foundation model development and deployment will likely be observable to LLMs [Chao, et al. 2023] [Robey 2023]. More broadly, regulators may wish to rely upon the fact that one lab or firm may keep tabs – through its own models or through other means – of what another lab is doing. [20]

4 Optionality: Experimenting with Different Forms of Approval Regulation

The upshot of these considerations might seem like an approval regulation regime is not worth trying for generative artificial intelligence. Yet such a conclusion would be premature. Any regulatory policy must be considered in 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. There is, of course, no law that stipulates (and certainly

[19] This characterization rests upon conversations I have had with practitioners and observers of frontier AI.

[20] This may be seem to separate the world of AI regulation entirely from that of FDA-style regulation, but not necessarily. Many biopharma firms have intelligence on what their competitors are doing and risks that develop with one class of molecules may be informative for a set of molecules under development by a rival or collaborator.

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.

In order to consider what experimentation might look like, one might begin with the different forms of generalized regulatory regimes that could be applied to AI. In a recent paper, [Guha et al, 2023] survey a range of possible regulatory regimes that have been proposed or could be applied to AI. These include disclosure regimes, registration regimes (which may involve certain compulsory disclosure), licensing regimes and auditing regimes. Suppose that one started with a registration regime, with or without mandatory disclosure. What might be able to be learned from such a regulatory framework abut how to modify regulation itself?

4.1 Learning from Registration and Reporting Regimes

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. As mentioned previously, regulation of foundation models is trending toward the adoption of registration and reporting requirements, though there are many aspects of these regimes, too, that suffer from adaptability and feasibility problems [Guha et al, 2023]. Policymakers might wish to ask questions like the following:

- what quantity and quality of evidence is produced by a minimal disclosure or registration requirement?
- can systems of adverse event reports [NAIAC 2023] lead to better methods for detecting statistical properties and regularities of harms?
- given a minimal disclosure or registration requirement, what is the marginal cost of adding further variables or observations to the set of existing requirements, and what is the marginal benefit?
- given a registration regime, how often does unauthorized or illegal FM development occur and what form does it take?

The last point here might seem non-sensical given that unauthorized FM development or deployment will be, by design or intention, difficult to detect. Note, however, that information useful for regulatory policy need *not* include precise estimates of how often illegal or unauthorized activity occurs. A legislature or regulatory leader might be able to regard observable unauthorized model development or deployment as at or near the minimum of the true distribution and develop a corresponding optimal control rule for the censored distribution.

4.2 Learning from Discretionary and Mandated Experiments

A range of governance regimes have already emerged for AI and foundation models, and beyond what is already occurring, there are many proposals for further development of AI governance standards. In some sense, there is experimentation right now.

- Arc Evals (now Model Evaluation and Threat Research (METR))
- jailbreaking via deployment of attacker models to unlock and then manipulate other models ([Chao, et al. 2023], [Robey 2023])
- responsible scaling research (Anthropic)
- red teaming ([Ganguli et al, 2022a], [Field])
- auditing ([Mökander], [Guha et al, 2023])

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 [AdaLovelace2023], a variety of "FDA-like" institutions [Tutt 2017] or even "adverse event reporting systems" [NAIAC 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.[21] In the biomedical regulation world, there ave 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 ([Callander and Harstaad 2015] [Volden, Ting and Carpenter 2008]), one wonders what learning value is surrendered when regulatory harmonization develops into strong uniformity.

One possibility is that a set of potentially governable risks might be adduced as they emerge in either experimentation or in real-world behavior. This is in the spirit of the NAIAC's recent recommendations for reporting requirements and an adverse event reporting system [NAIAC 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 statistical evaluation.

One final problem with optionality 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 [Acemoglu and Lensman 2023].

[21] 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 with, but not stringent coordination with, each other. this is quite different from cals for self-regulation or no regulation at all.

4.3 Learning from the History of Approval Regulation

Suppose, finally, that a primitive approval regulation regime for foundation models is established. What might be learned from it? Historically, societies have experimented with different forms of approval regulation. This occurred over time in the U.S., where pharmaceutical and medical product regulation moved from purely ex post models to ex ante approval regulation and then different models of regulating experiments under approval regulation, but also where different forms of deregulation or regulatory relaxation have occurred. Beyond historical comparisons within the United States, there have been adaptations of regulatory frameworks across national and regional settings. European societies were long accustomed to apply less stringent approval regulation than in the United States. The reduced stringency took the form of weaker experimental standards entailing less costly experiments that observed fewer dimensions of efficacy and risk, as well as weaker requirements on dossiers such that experimental data were summarized, and, finally, 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.

Beyond this, two features of the evolution of FDA-like institutions over the past century ware noteworthy for the manner in which previous structures were transformed and in some cases rather thoroughly. One of the examples amounts to greater regulatory stringency whereas the second concerns a mode of regulatory relaxation. In neither case is a full cost-benefit analysis of the relevant transformation needed to acknowledge the simple fact that the world's most notable and copied approval regulation regime has been anything but static.

4.3.1 Learning from the Pre-History of FDA Efficacy Regulation

What we know of the history of molecular regulation in therapeutics is that it 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). It was the development of pathology and pharmacology combined with particular regulatory crises in the 1930s [Carpenter and Sin 2007] that led to a new regime of regulatory pre-market review. The Federal Food Drug and Cosmetics Act of 1938 required pre-market approval of new therapeutics by the Secretary of Agriculture (in whose Department the bureau that became the FDA then sat), and the statute and associated rulemaking specified the set of tests to carry out. Critically, there was substantial regulatory development under the 1938 statute even before the 1962 Kefauver-Harris Amendments, such that much of the modern system of efficacy regulation was in place before the thalidomide crisis ([Carpenter 2010], Chapter 3), and this development of regulatory methodology depended heavily upon coincident develop-

ments in pharmacology, statistics and the study of clinical trials and cancer therapeutics [Keating and Cambrosio 2019]. Indeed, the experience with regulation of biomedical product safety (both before and following the Federal Food, Drug and Cosmetic Act of 1938) generated a large methodological and evidence base for the consideration of efficacy, and this was true *before* Congress mandated proof of "effectiveness" as necessary for marketing in the Kefauver-Harris Amendments of 1962 ([Carpenter 2010], Chapter 3).

4.3.2 Learning from Regulatory Experience with Surrogate Endpoints

A quite different example of regulatory transformation comes in the FDA's increasing use of surrogate endpoints in modern therapeutic approvals, most notably in oncology. 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. In some areas of therapeutics, most or all new drugs are now approved on the basis of surrogate endpoints [Yu, et al., 2015]. There is, of course, an abiding debate about the merits of such programs, including whether they permit other missing innovation to materialize[Budish, Roin and Williams 2015], whether there is insufficient follow-up from approved drugs ([Moore and Furberg 2014], [Carpenter 2014]), and whether the drugs approved under such rules – since 1992, under the accelerated approval pathway – truly bring the benefit they promise ([Fleming 2005] [Naci, Smalley and Kesselheim 2017]). What cannot be doubted at this point is that there have been a range of transformations to FDA approval models that represent the incorporation of statistical and scientific findings and that represent alterations based in part on regulatory experience.

Are surrogate endpoints applicable to assessing catastrophic or extreme risks from foundation models? Early developments in jailbreaking research ([Chao, et al. 2023], [Robey 2023]) point to this possibility, insofar as jailbreaking exercises aim to probe LLMs for vulnerabilities. In particular, [Chao, et al. 2023] distinguishes between *prompt-level* jailbreaks and *token-level jailbreaks*. The latter are far more computationally costly than the former, but LLMs may solve such a scaling problem, and both forms of jailbreak pose risks. Researchers ([Chao, et al. 2023] [Robey 2023]) have developed algorithms capable of prompt jailbreaks in short-order. In doing so, these researchers have relied upon a set of indicators used in earlier jailbreaking research [Zou, et al., 2023], detected by "LLMs-as-judges" themselves [Zheng, et al., 2023], and validated using human coding and conversations. Whether standardized lists of types of standardized "misalignment" or "non-alignment" can be developed is an open question, but it would not seem impossible.

5 Conclusion: Regulatory Learning by Doing and Policy Reversibility

. This essay joins other calls for circumspection in the application of regulatory models to generative artificial intelligence, in particular calling for caution about the feasibility of "FDA-like" approval regulation regimes to the regulation of foundation models and the catastrophic risks they may pose. The greatest impediments to such a model, in my judgment, are (1) enforceability of experimentation requirements and development/deployment restrictions and, perhaps most important, (2) the inapposite mapping between the world of AI and the large-sample world in which approval regulation operates, due to the lack of well-established indicators of catastrophic risk that satisfy the countable additivity properties of adverse events in molecular regulation. That said, there is abundant historical precedent for thinking that approval regulation regimes can be tried, or approximated through alternative models of regulation, including a "conversation" between models of post-market regulation that identifies governable risks and approval regulation models with pre-market requirements. There is also reason to think that regulation in the AI arena will be characterized by policies and institutions that, once adopted, will be more reversible than in the field of health treatments, at least for a reasonably long future.[22] One reason they are likely to be more reversible is that the very set of facilitating social and economic institutions that shape FDA regulation does not yet exist in artificial intelligence realm, or does not exist to the same degree. The degree of institutional "lock-in" between regulators, laboratories, professions and industry structure that prevails in global pharmaceuticals would take decades to create in another realm.

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 1984]. 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 [Stevenson 2023] suggest that a purely experimental approach may be wrong for optimization of policies in different domains. 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 often unquantifiable nature of their dangers.

[22] I say "reason to think" but this should not be assumed and points to an important domain of needed research.

References

[Acemoglu and Lensman 2023] Acemoglu, Daron, and Todd Lensman. 2023. "Regulating Transformative Technologies," MIT Working Paper. https://economics.mit.edu/sites/default/files/2023-07/Regulating%20Transformative%20Technologies.pdf

[AdaLovelace2023] Ada Lovelace Institute. 2023. *Safe before sale: Learnings from the FDA's model of life sciences oversight for foundation models* https://www.adalovelaceinstitute.org/report/safe-before-sale/

[Altman2023] Altman, Sam. 2023. Written Testimony of Sam Altman Chief Executive Officer OpenAI Before the U.S. Senate Committee on the Judiciary Subcommittee on Privacy, Technology, & the Law, May 16, 2023. https://www.judiciary.senate.gov/imo/media/doc/2023-05-16%20-%20Bio%20&%20Testimony%20-%20Altman.pdf

[AWS2023] Amazon Web Services. 2023. What are Foundation Models?. https://aws.amazon.com/what-is/foundation-models/

[Anderljung2023] Markus Anderljung, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, Ben Chang, Tantum Collins, Tim Fist, Gillian Hadfield, Alan Hayes, Lewis Ho, Sara Hooker, Eric Horvitz, Noam Kolt, Jonas Schuett, Yonadav Shavit, Divya Siddarth, Robert Trager, Kevin Wolf. 2023. "Frontier AI Regulation: Managing Emerging Risks to Public Safety" https://arxiv.org/abs/2307.03718.

[Bailey 2023] Bailey, Ronald. 2023. "OpenAI Chief Sam Altman Wants an FDA-Style Agency for Artificial Intelligence," Reason.com, May 16, 2023, https://reason.com/2023/05/16/openai-chief-sam-altman-wants-an-fda-style-agency-for-artificial-i

[Bates et al 2023] Bates, Stephen, Michael I. Jordan, Michael Sklar, Jake A. Soloff. 2023. "Incentive-Theoretic Bayesian Inference for Collaborative Science," July 10, 2023. https://arxiv.org/abs/2307.03748

[Bennet and Welch 2023] Bennet, Senator Michael and Senator Peter Welch. 2023. "To establish a new Federal body to provide reasonable oversight and regulation of digital platforms," 118th Congress, 1st Session. https://www.bennet.senate.gov/public/_cache/files/2/b/2b3c99bf-a4aa-40d5-8f10-1f2b994ca03c/2BB12EB960B8928B7BEE7A8285D61AF5.dpca-bill-text.pdf

[Billingsley 1995] Billingsley, Patrick. 1995. *Probability and Measure*, Third Edition (New York: Wiley).

[Brown, et al. 2022] Brown, B.L., Mitra-Majumdar, M., Darrow, J.J., Moneer, O., Pham, C., Avorn, J. and Kesselheim, A.S., 2022. "Fulfillment of post-market commitments and requirements for new drugs approved by the FDA, 2013-2016," *JAMA Internal Medicine*, 182(11), 1223-1226.

[Budish, Roin and Williams 2015] Budish, Eric, Benjamin N. Roin, and Heidi Williams. 2015. "Do firms underinvest in long-term research? Evidence from cancer clinical trials," *American Economic Review* 105, no. 7 (2015): 2044-2085.

[Callander and Harstaad 2015] Callander, Steven, and Bärd Harstad. 2015. "Experimentation in federal systems," *The Quarterly Journal of Economics* 130, no. 2 (2015): 951-1002.

[Carlsmith 2022] Carlsmith, J. 2022. "Is Power-Seeking AI an existential risk?" `https://arxiv.org/abs/2206.13353`.

[Carpenter 2004] Carpenter, Daniel. 2004. "Protection without capture: Product approval by a politically responsive, learning regulator," *American Political Science Review* 98, no. 4 (2004): 613-631.

[Carpenter 2010] Carpenter, Daniel. 2010 *Reputation and Power: Organizational Image and Pharmaceutical Regulation at the FDA* (Princeton: Princeton University Press).

[Carpenter 2014] Carpenter, D., 2014. "Can expedited FDA drug approval without expedited follow-up be trusted?" *JAMA Internal Medicine*, 174(1), 95-97.

[Carpenter and Sin 2007] Carpenter, Daniel, and Gisela Sin. 2007. "Policy tragedy and the emergence of regulation: the Food, Drug, and Cosmetic act of 1938," *Studies in American Political Development* 21, no. 2 (2007): 149-180.

[Carpenter and Ting 2007] Carpenter, Daniel, and Michael M. Ting. 2007. "Regulatory errors with endogenous agendas," *American Journal of Political Science* 51, no. 4 (2007): 835-852.

[Carpenter, Grimmer and Lomazoff 2010] Carpenter, Daniel, Justin Grimmer and Eric Lomazoff. 2010. "Approval regulation and endogenous consumer confidence: Theory and analogies to licensing, safety, and financial regulation," *Regulation & Governance* 4, no. 4 (2010): 383-407.

[Carpenter, et al. 2010] Carpenter, Daniel, Susan I. Moffitt, Colin D. Moore, Ryan T. Rynbrandt, Michael M. Ting, Ian Yohai, and Evan James Zucker. 2010. "Early entrant protection in approval regulation: Theory and evidence from FDA drug review," *The Journal of Law, Economics, & Organization* 26, no. 3 (2010): 515-545.

[Chao, et al. 2023] Chao, Patrick, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. 2023. "Jailbreaking black box large language models in twenty queries," arXiv preprint arXiv:2310.08419 (2023). https://arxiv.org/pdf/2310.08419.pdf

[Djankov et al, 2002] Djankov, Simeon, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2002. "The regulation of entry," *The Quarterly Journal of Economics* 117, no. 1 (2002): 1-37.

[Dugan 2023] Dugan, Kevin T. 2023. "Congress Isn't Ready for the AI Revolution," *New York Magazine – Intelligencer* May 22, 2023, https://nymag.com/intelligencer/2023/05/sam-altman-congress-senate-ai-hearing.html.

[Field] Hayden Field. 2022. "How microsoft and google use ai red teams to 'stress test' their systems." https://www.emergingtechbrew.com/stories/2022/06/14/how-microsoft-and-google-use-ai-red-teams-to-stress-test-their-sys

[Fleming 2005] Fleming, Thomas R. 2005. "Surrogate endpoints and FDA?s accelerated approval process," *Health Affairs* 24, no. 1 (2005): 67-78.

[Ganguli et al, 2022a] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, Jack Clark. 2022. "Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned," arXiv preprint arXiv:2209.07858, 2022. https://arxiv.org/abs/2209.07858.

[Ganguli et al, 2022b] Deep Ganguli, Danny Hernandez, Liane Lovitt, Nova DasSarma, Tom Henighan, Andy Jones, Nicholas Joseph, Jackson Kernion, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Dario Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Chris Olah, Jack Clark. "Predictability and Surprise in Large Generative Models." https://arxiv.org/abs/2202.07785

[Greenstone 2009] Greenstone, M. 2009. "Toward a culture of persistent regulatory experimentation and evaluation," in Tobin Project, *New perspectives on regulation* (New York: Cambridge University Press), 111, pp.116-19.

[Guha et al, 2023] Neel Guha, Christie M. Lawrence, Lindsey A. Gailmard, Kit T. Rodolfa, Faiz Surani, Rishi Bommasani, Inioluwa Deborah Raji, Mariano-Florentino Cuéllar, Colleen Honigsberg, Percy Liang, Daniel E. Ho. 2023. "AI Regulation Has Its Own Alignment Problem: The Technical and Institutional Feasibility of Disclosure, Registration, Licensing, and Auditing," *George Washington University Law Review*, forthcoming.

[Hadfield-Menell et al, 2016] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, Stuart Russell. 2016. "The Off-Switch Game." `https://arxiv.org/abs/1611.08219`

[Henry, Loseto and Ottaviani 2022] Henry, Emeric, Marco Loseto, and Marco Ottaviani. 2022. "Regulation with experimentation: Ex ante approval, ex post withdrawal, and liability," *Management Science* 68, no. 7 (2022): 5330-5347.

[Henry and Ottaviani 2019] Henry, Emeric, and Marco Ottaviani. 2019. "Research and the approval process: The organization of persuasion," *American Economic Review* 109, no. 3 (2019): 911-955.

[Hermann et al, 2022] Herrmann, J., Lenihan, D., Armenian, S., Barac, A., Blaes, A., Cardinale, D., Carver, J., Dent, S., Ky, B., Lyon, A.R. and López-Fernández, T., 2022. "Defining cardiovascular toxicities of cancer therapies: an International Cardio-Oncology Society (IC-OS) consensus statement," *European heart journal*, 43(4), 280-299.

[Hu et al, 2021] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, "LoRA: Low-Rank Adaptation of Large Language Models," `https://arxiv.org/abs/2106.09685`

[Hwang et al, 2014] Hwang, T.J., Kesselheim, A.S. and Bourgeois, F.T., 2014. "Postmarketing trials and pediatric device approvals," *Pediatrics*, 133(5), e1197-e1202.

[Hwang et al, 2016] 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.

[Karnofsky 2022] Karnofsky, Holden. 2022. "AI Could Defeat all of Us Combined," *ColdTakes.com*, June 9, 2022; `https://www.cold-takes.com/ai-could-defeat-all-of-us-combined/`

[Kay and King 2020] Kay, John, and Mervyn King. 2020. *Radical Uncertainty: Decision-Making Beyond the Numbers* (New York: Norton).

[Keating and Cambrosio 2019] Keating, Peter, and Alberto Cambrosio. 2019. *Cancer on trial: oncology as a new style of practice*. University of Chicago Press, 2019.

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

[Libgober and Carpenter 2024] Libgober, B. and Carpenter, D., 2023. "Lawyers as Lobbyists: Regulatory Advocacy in American Finance" *Perspectives on Politics*, forthcoming.

[Lyon, et al 2020] 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. and Khoo, V., 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.

[Marcus 2023] Marcus, Gary. 2023. Replies to Senate Queries, Emeritus Professor Gary Marcus ,13 June 2023. `https://www.judiciary.senate.gov/imo/media/doc/2023-05-16_-_qfr_responses_-_marcus.pdf`

[Marks 1997] Marks, Harry M. 1997. *The progress of experiment: science and therapeutic reform in the United States, 1900-1990* (Cambridge University Press, 1997).

[McClellan 2022] McClellan, Andrew. 2022. "Experimentation and approval mechanisms," *Econometrica* 90 (5): 2215-2247.

[McCraw 1984] McCraw, Thomas. 1984. *Prophets of Regulation: Charles Francis Adams, Louis D. Brandeis, James M. Landis, Alfred E. Kahn* (Cambridge, Mass.: Harvard University Press).

[Mökander] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2023. "Auditing large language models: a three-layered approach," arXiv preprint arXiv:2302.08500, 2023. `https://arxiv.org/abs/2302.08500`.

[Moore and Furberg 2014] 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, Smalley and Kesselheim 2017] Naci, Huseyin, Katelyn R. Smalley, and Aaron S. Kesselheim. 2017. "Characteristics of preapproval and postapproval studies for drugs granted accelerated approval by the US Food and Drug Administration," *JAMA* 318, no. 7 (2017): 626-636.

[NAIAC 2023] Ntional Artificial Intelligence Advisory Committee (NAIAC). 2023. textitRecommendation: Improve Monitoring of Emerging Risks from AI through Adverse Event Reporting, November 2023; https://ai.gov/wp-content/uploads/2023/12/Recommendation_Improve-Monitoring-of-Emerging-Risks-from-AI-through-Adverse-Event pdf.

[Ngo, Chan and Mindermann 2023] Ngo, Richard, Laurence Chan and Sören Mindermann. 2023. The Alignment Problem from a Deep Learning Perspective. https://arxiv.org/abs/2209.00626

[Ottaviani and Wickelgren 2023] Ottaviani, Marco, and Abraham L. Wickelgren. 2023. "Approval regulation and learning, with application to timing of merger control," *The Journal of Law, Economics, and Organization* (2023): ewac025.

[Posner and Weyl 2013] Posner, Eric, and E. Glen Weyl. 2013. "Benefit-cost analysis for financial regulation," *American Economic Review* 103, no. 3 (2013): 393-397.

[Rastogi 2023] Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, and Saleema Amershi. 2023. "Supporting human-ai collaboration in auditing llms with llms." arXiv preprint arXiv:2304.09991, 2023. https://arxiv.org/abs/2304.09991

[Robey 2023] Robey, Alexander, Eric Wong, Hamed Hassani, and George J. Pappas. "Smoothllm: Defending large language models against jailbreaking attacks." arXiv preprint arXiv:2310.03684 (2023). https://arxiv.org/pdf/2310.03684.pdf

[Ross 2000] Ross, Sheldon M. 2000. *Introduction to Probability Models*, Seventh Edition (San Diego: Harcourt Academic Press).

[Shevlane2023] Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, Lewis Ho, Divya Siddarth, Shahar Avin, Will Hawkins, Been Kim, Iason Gabriel, Vijay Bolina, Jack Clark, Yoshua Bengio, Paul Christiano, Allan Dafoe. 2023. "Model evaluation for extreme risks." https://arxiv.org/abs/2305.15324

[Singer and Cole 2016] Singer, P. W. and August Cole. 2016. *Ghost Fleet: A Novel of the Next World War* (New York: William Morrow Paperbacks).

[Stevenson 2023] Stevenson, Megan T. 2023. "Cause, Effect and the Structure of the Social World," *Boston University Law Review* 103: 2001-2047

[Sunstein 2023] Sunstein, C. 2023. "Knightian Uncertainty." Available at SSRN. https://ssrn.com/abstract=4662711 or http://dx.doi.org/10.2139/ssrn.4662711

[Taleb 2014] Taleb, Nassim Nicholas. 2014. *Antifragile: Things that gain from disorder.* (New York: Random House Trade Paperbacks).

[Taori, et al., 2023] Taori, Rohan, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, Tatsunori B. Hashimoto. 2023. *Alpaca: A Strong, Replicable Instruction-Following Model*, https://crfm.stanford.edu/2023/03/13/alpaca.html.

[Thierer and Chilson 2023] Thierer, Adam, and Neil Chilson. 2023. "The Problem with AI Licensing & an FDA for Algorithms," *The Federalist Society*, June 5, 2023. https://fedsoc.org/commentary/fedsoc-blog/the-problem-with-ai-licensing-an-fda-for-algorithms

[Tutt 2017] Tutt, Andrew. 2017. "An FDA for Algorithms," *Administrative Law Review* 69 (1) (2017) 83-123.

[U.S. FDA 2018] U.S. Food and Drug Administration. 2018. *Acceptance of Clinical Data to Support Medical Device Applications and Submissions: Frequently Asked Questions; Guidance for Industry and Food and Drug Administration Staff*, FDA-2013-N-0080; https://www.fda.gov/media/111346/download

[U.S. FDA 2019] U.S. Food and Drug Administration. 2019. Premarket Approval (PMA) [Medical Devices] https://www.fda.gov/medical-devices/premarket-submissions-selecting-and-preparing-correct-submission/premarket-approval-pma#when

[Volden, Ting and Carpenter 2008] Volden, Craig, Michael M. Ting, and Daniel P. Carpenter. 2008. "A formal model of learning and policy diffusion," *American Political Science Review* 102, no. 3 (2008): 319-332.

[White House 2023] The White House. 2023. *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, September 30, 2023; https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and

[Wallach et al 2018] 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*, 2018 May 24;361.

[Weitzmann 2014] Weitzman, Martin L. 2014. "Fat tails and the social cost of carbon," *American Economic Review* 104, no. 5 (2014): 544-546.

[Welch 2023] Welch, Senator Peter. 2023. "Welch, Bennet Reintroduce Landmark Legislation to Establish Federal Commission to Oversee Digital Platforms," Office of United States Senator Peter Welch, https://www.welch.senate.gov/press-releases/welch-bennet-reintroduce-landmark-legislation-to-establish-federal

[White House 2023] The White House. 2023. *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, September 30, 2023; https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and

[Yu, et al., 2015] Yu, Tsung, Yea-Jen Hsu, Kevin M. Fain, Cynthia M. Boyd, Janet T. Holbrook, and Milo A. Puhan. 2015. "Use of surrogate outcomes in US FDA drug approvals, 2003?2012: a survey," *BMJ open* 5, no. 11 (2015).

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

[Zheng, et al., 2023] Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, Ion Stoica. 2023. "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena"; https://arxiv.org/abs/2306.05685