OXFORD

Policy and Society, 2025, **44(1)**, 85–97

DOI: https://doi.org/10.1093/polsoc/puae020
Advance access publication date 29 May 2024
Original Research Article

When code isn't law: rethinking regulation for artificial intelligence

Brian Judge, Mark Nitzberg, Stuart Russell

Center for Human-Compatible AI, University of California—Berkeley, Berkeley, CA, United States
Corresponding author: B. Judge, Email: bjudge@berkeley.edu

Abstract

This article examines the challenges of regulating artificial intelligence (AI) systems and proposes an adapted model of regulation suitable for AI's novel features. Unlike past technologies, AI systems built using techniques like deep learning cannot be directly analyzed, specified, or audited against regulations. Their behavior emerges unpredictably from training rather than intentional design. However, the traditional model of delegating oversight to an expert agency, which has succeeded in high-risk sectors like aviation and nuclear power, should not be wholly discarded. Instead, policymakers must contain risks from today's opaque models while supporting research into provably safe AI architectures. Drawing lessons from AI safety literature and past regulatory successes, effective AI governance will likely require consolidated authority, licensing regimes, mandated training data and modeling disclosures, formal verification of system behavior, and the capacity for rapid intervention.

Keywords: artificial intelligence; regulation; AI safety; technology

In an earlier era of rapid technological innovation, Larry Lessig famously summarized: in cyberspace, *code is law* (Lessig, 1999, 2000). Lessig's dictum was a call to regulate software and protocols on the nascent Internet, in order to uphold our values in the digital world as laws do in the physical world.[1] At the time, explicit designs were the basis of largely all digital system behavior. *Code is law*. As with other engineered systems such as aircraft and nuclear power plants, a given system can be audited for compliance with a given regulatory specification. The creators of these systems can show that the design, construction, and monitoring of the system conforms to these rules. This is the standard regulatory template for many regulated technologies: Food and Drug Administration (FDA) medical device clearance applies for MRI machines and digitally controlled ventilators; the National Highway Transportation Safety Administration (NHTSA) safety rules apply to computer-controlled windshield wipers; and the Nuclear Regulatory Commission (NRC) safety rules for the digital systems controlling nuclear power plants.

But in the era of generative artificial intelligence (AI), code is no longer law: for reasons we will explain, the code that humans write does not in itself *determine* how generative AI systems operate. This includes generative AI systems that excel at a wide range of creative and cognitive tasks. *Code is*

[1] "Our choice is not between 'regulation' and 'no regulation.' The code regulates. It implements values or not. It enables freedoms or disables them. It protects privacy or enables mass surveillance. People choose how the code does these things. People write the code. Thus the choice is not whether people will decide how cyberspace regulates. People, coders, and companies will. The only choice is whether we collectively will have a role in their choice—and thus in determining how these values regulate—or whether collectively we will allow the coders to select our values for us" (Lessig, 1999).

law no longer applies to deep networks and generative AI systems because they are opaque and not designed: these systems are created through a massively resource-intensive training process of tuning trillions of parameters. As a result, it is not possible to encode a rule like "LLMs must not dispense medical advice" into the model itself. Instead, system engineers must hope the model abides by the desired behavior after sufficient reinforcement. Since code does not explicitly determine these systems' behavior, it is impossible to demonstrate compliance with a given regulatory specification. Nor too can the reasons for misbehavior be traced and corrected. So long as our AI systems are built on "black-box" data-driven systems, the regimes regulating these systems will remain approximate and incomplete.

And this "black-box" problem is just one way in which AI presents novel regulatory challenges. We are entering a world of intelligent machines that rival, and often exceed, human capabilities across a range of domains (Russell, 2019). There is wide agreement among technologists, policymakers, social scientists, and business leaders that this transformation will be far-reaching and highly significant (Dafoe, 2022). Potential risks and harms range from job displacement and inequality to the dissolution of consensus reality and the creation of a surveillance state (Acemoğlu, 2022). There are also serious worries that AI systems more capable than humans across *all* domains (what is often called "artificial general intelligence") will escape human control. The default path—what is likely to happen absent meaningful regulatory intervention—will be highly disruptive and potentially catastrophic (Bengio, 2023). In recognition of this problem, 28 countries including the United States and China signed a statement at the AI Safety Summit held at Bletchley Park in late 2023 affirming "the need for the safe development of AI" and warned of "serious, even catastrophic, harm, either deliberate or unintentional, stemming from the most significant capabilities of these AI models" (AI Safety Summit, 2023).

Nearly everyone agrees government regulation of generative AI is necessary, but there is little consensus around the form it should take. This dissensus arises from both the usual conflicting interests and institutional gridlock at the heart of regulatory politics but now also the novel features of generative AI as a technology. Existing approaches to regulating high-risk technological systems are premised on the ability to ensure that their design and operation conform to some rule. For example, the NHTSA develops federal vehicle safety standards and audits compliance with those standards. How can this paradigm be extended to generative AI? A generative AI system constructed by a process of machine learning may be characterized as a "black box"—not because its inner workings are secret but because they are opaque. It is not yet possible to understand the precise operation of a large language model (LLM) with one trillion parameters, as GPT-4 is estimated to have.

Policymakers must recognize that code is no longer law: that with currently popular approaches to building generative AI systems, code cannot regulate in the same manner because the systems' behavior is an emergent property. Since its behavior is neither expressed intentionally by designers in software program code nor legible (yet) by examining the program code and its massive array of tuned parameters, generative AI has the key distinction among engineered systems that their program code does not dictate its behavior. This unique, defining technical characteristic has significant implications for debates about the governance and regulation of generative AI.[2] It is one key premise on which we base a novel approach to regulating generative AI that contributes to the emerging interdisciplinary literature on the regulation and governance of AI (Büthe et al., 2022; Cath, 2018; Taeihagh, 2021). Our contribution is to bring together policy studies and technical AI safety research to lay the foundation for a regulatory approach to generative AI that both accounts for its novel technical features and provides confidence that systems will not cause major harms as with other high-stakes technologies.

Regulation of LLMs and similar systems cannot rely on the same methods that work for aircraft and nuclear plant control systems. Whereas nuclear plants and aircraft have component structures and component-based physical models that can be analyzed to predict their behavior separately and as an ensemble, LLMs are black boxes. When an aircraft fails, analysis can trace the source of failure; the failed components or interactions can then be corrected to avoid a repeat of the failure. The black-box nature of LLMs precludes this kind of tracing of failure to its source and correction to avoid its recurrence. Because in generative AI code is not law, a regulatory approach premised on specifications, audits, and testing cannot ensure safety and reliability. We argue that the essential role of regulation

[2] Governance, as distinct from regulation, refers to the rules of collective decision-making where there are multiple actors without formal control over each other. The rise of governance over recent decades reflects the proliferation of different policy instruments and actor groups (Capano et al., 2015; Colebatch, 2014; Howlett, 2019; Wu et al., 2015).

Table 1. Examples of the traditional model of regulation in the United States.

Agency	Purview	Law	Catalyzing event
FDA	Drugs, medical devices, food	The Pure Food and Drug Act of 1906	Publication of Upton Sinclair's *The Jungle*
Environmental Protection Agency	Environmental pollution	Clean Air Act; Clean Water Act	Cuyahoga River fires; environmentalist movement
Securities and Exchange Commission	Financial securities	Securities Act of 1933; Sarbanes–Oxley Act; Dodd–Frank Act	1929 stock market crash; Enron/WorldCom accounting scandals; 2008 financial crisis
Federal Communications Commission	Telecommunications	Communications Act of 1934	Growth of radio leading to interference issues across different broadcast media
Drug Enforcement Administration	Narcotics	Controlled Substances Act of 1970	Created as part of Nixon's controversial "war on drugs"

is to proactively prevent harms from unsafe architectures while funding, developing, and incentivizing architectures with the safety properties appropriate for a world of intelligent machines.

To make this argument, we first present case studies of the Federal Aviation Administration (FAA) and the NRC in the United States: both have established impressive track records of safety despite the inherently unsafe nature of the technologies they regulate.[3] We consider these cases not because aviation and nuclear power are perfect technological analogues to generative AI, but because they establish a baseline for the scope of authority needed to credibly and effectively regulate high-risk technological systems. Second, we consider how generative AI challenges this paradigm and turn to the technical literature on AI safety for existing approaches to mitigating harms. Finally, we combine the lessons from the FAA and NRC with the lessons from AI safety to sketch a regulatory paradigm for generative AI that adapts the traditional model of regulation to these novel technological features.

The traditional model of regulation

In the traditional model of regulation in the United States, Congress establishes a specialized agency to administer a law, often catalyzed by a widely publicized harm to public welfare. The agency is authorized to create specific regulations necessary to implement the law. The Administrative Procedure Act of 1946 governs the rulemaking process and establishes a legal mechanism for resolving disputes. The agency monitors compliance with its regulations, conducts investigations and audits, and imposes penalties for violations. The agency is staffed by subject-matter experts and is designed to be independent from political influence.[4] Table 1 presents several examples of regulatory agencies in the United States that follow this formula. We focus on the FAA and the NRC because they follow this same "traditional model" but also address the potential for widespread externalities from deployed systems.[5] We have selected the two cases because they are typical cases of traditional regulatory designs for emerging technologies.

Federal aviation administration

The history and function of the FAA is useful as certain aspects apply poorly to regulating AI, such as waiting until a catastrophic accident crystallizes sufficient public consensus to create a regulatory regime in the first place, while other aspects provide useful guidance for regulating AI, such as detailed licensing, certification and approval processes, and clear definitions of acceptable levels of risk.

At the urging of the nascent aviation industry, President Calvin Coolidge enacted the Air Commerce Act of 1926 that empowered the Secretary of Commerce to issue and enforce air-traffic control (ATC)

[3] We focus on the United States because it is home to the world's most important AI companies and because it is noticeably lagging the European Union in developing and enacting comprehensive AI regulations (Paul, 2023; Radu, 2021; Ulnicane & Erkkilä, 2023).

[4] There is a vast literature in economics on regulatory capture beginning with Stigler (1971). The risk of capture is omnipresent, but this risk is not inherently greater with AI than it is with any other technology.

[5] The FDA is arguably a closer analogue given the "black box" nature of some pharmaceuticals. This comparison has been considered elsewhere as a comparison case to AI (Stein & Dunlop, 2023).

rules, certify aircraft, license pilots, and operate navigation aids.[6] These early pioneers of the aviation industry believed that federal action was required to maintain safety standards and increase public confidence in this new technology. The Aeronautics Branch of the Department of Commerce orchestrated the replacement of flagmen—air-traffic controllers who would wave colored flags to communicate with pilots—with radio-equipped control towers.

In 1931, a Trans World Airlines Fokker-10 carrying legendary Notre Dame football coach Knute Rockne crashed in rural Kansas. (Rockne was later memorialized in the 1940 film *Knute Rockne, All American* starring Ronald Reagan as Notre Dame All-American George Gipp.) Congress, responding to the popular belief that airlines and manufacturers were too close to the recently renamed Bureau of Air Commerce, established the Civil Aeronautics Authority with the responsibility for conducting accident investigations and making safety recommendations. According to a common saying, "Aviation regulations are written in blood."

On 30 June 1956, a United Airlines DC-7 and TWA Super Constellation collided over the Grand Canyon killing all 126 passengers and crew. This tragic accident was a pivotal movement in the history of aviation safety in the United States. The existing rules—"see and be seen"—were increasingly inadequate due to the rapid growth of commercial air travel after World War II. The Grand Canyon disaster was the most extreme example of mid-air collisions that had become increasingly common. In 1958, President Eisenhower signed the Federal Aviation Act into law, establishing a single federal agency with authority over all aspects of civil aviation. In 1966, Congress created the Department of Transportation with authority over all aspects of federal regulation relating to transportation. Thirty-one federal agencies were re-incorporated under one Cabinet department.

Today, the FAA is responsible for the entire aviation lifecycle: licensing of pilots and mechanics, certifications of airplanes, ATC and management of airspace, safety inspections, incident responses, and other related functions. Any new airplane model must pass a rigorous certification process including manufacturing facility inspections and test flights. Only models that pass these evaluations are certified for commercial use. When an incident occurs, the FAA generates comprehensive incident reports that uncover defects in existing designs and mandate appropriate remedial action.[7]

"Airworthiness" is the guiding concept for assessing safety. The FAA defines it as "the status of an aircraft, engine, propeller or part when it conforms to its approved design and is in a condition for safe operation." Accordingly, the FAA issues "airworthiness directives" that are legally enforceable rules applying to aircraft and their components. Directives are issued when the FAA finds that "an unsafe condition exists in the product and the condition is likely to exist or develop in other products of the same type design." Airworthiness directives allow the FAA to act quickly and decisively to resolve safety concerns.

Nuclear Regulatory Commission

Whereas the FAA regulates aircraft and air travel, avoiding perhaps thousands of deaths, the NRC was created to regulate nuclear power and weaponry, which come with a specter of death and destruction on a massive scale, even from a single accident. Studying its history and function informs our search for relevant mechanisms and pitfalls in developing AI regulation.

The nuclear age began at Hiroshima and Nagasaki with a catastrophic display of its inherent danger. Although many hailed nuclear power as a potentially revolutionary technology for raising global living standards, it was self-evident that it should be tightly controlled. American policymakers were initially hesitant to commercialize the technology. In 1953, Atomic Energy Commissioner Thomas Murray warned of a "nuclear power race." The point was that the United States could not abandon the development of peaceful uses of nuclear power without endangering its scientific and economic dominance.

The Atomic Energy Act (AEA) of 1954 was landmark legislation that for the first time allowed private companies to own and use nuclear materials, subject to licensing and regulation by the Atomic Energy Commission (later restructured as the NRC). The Act assigned the AEC three major areas of regulation and oversight: weapons development, commercialization of nuclear power, and safety regulation. One key concern, exemplified by a remark from Commissioner Willard Libby, is that "this great benefit to

[6] https://www.faa.gov/about/history/brief_history.
[7] For an example incident report, see https://www.faa.gov/lessons_learned/transport_airplane/accidents/OE-LAV.

mankind will be killed by unnecessary regulation." The AEA was an effort to trade off the need for safety and control against the potential benefits of this new technology (Walker & Wellock, 2010).

Everyone agreed that safety was an essential element in progress. A single mishap might turn the public against nuclear power forever. The act created a new category of classified information called "Restricted Data" related to nuclear weapons design, fissile material production, and the use of nuclear material for energy. It also established guidelines for liability in the event of a nuclear accident, including a liability cap for operators, requirements for private insurance, and the waiving of certain legal defenses. Additionally, the act provided federal funding for research and development in nuclear energy (Mazuzan & Walker, 1984).

The AEA set up a strict licensing regime for civilian nuclear power facilities. Companies seeking to build and operate nuclear plants must go through an extensive application process and meet stringent safety requirements. The mean time to failure (MTTF) for nuclear power plants—where failure means a major core accident—was initially 10 thousand years, which was subsequently raised to 10 million years. Individual components might last much less than this but are regularly tested and replaced; the maintenance procedure is part of the plant model. Designers had to provide an analytic proof for their MTTF estimates whose steps and assumptions could be challenged and verified. Skilled craftspeople working on nuclear power plants are also subject to thorough certification programs to ensure that they have the proper training and qualifications. The licensing process and workforce requirements were designed to ensure the safe use of civilian nuclear power. In 1958, the first American nuclear power plant opened near Pittsburgh. The reactor core was repurposed from a canceled aircraft carrier. In 1974, the Energy Reorganization Act split the civilian and military portfolios of the Atomic Energy Commission into the NRC and the Department of Energy.

The cases of the FAA and NRC overlap in a number of crucial ways. First, both require an extensive licensing, certification, and approval process that responds to revealed failure modes. Second, both the FAA and the NRC are staffed by scientists and engineers with deep expertise in their subject areas. Finally, both agencies are authorized to recall products from the market or initiate total shutdowns/groundings if circumstances warrant. These common features establish a baseline for a similar regulatory regime for generative AI.

The challenges of generative AI

There are at least five major reasons why AI poses novel challenges from the standpoint of the "traditional model."

First, AI systems such as LLMs are general-purpose technologies (Bresnahan & Trajtenberg, 1995; Eloundou et al., 2023). Such technologies have many different uses across the economy and generate spillover effects into other sectors and even transform the international balance of power and conduct of military operations (Ding & Dafoe, 2023). General-purpose technologies pose unique regulatory challenges (Taeihagh et al., 2021). For this reason, AI is often compared to electricity or the Internet. The general-purpose quality of AI also complicates specifying precise definitions of AI.[8] However, both electricity and the Internet are highly regulated: for electricity, we define voltage, frequency, cables, plugs, and so on; for the Internet, the network protocols are law and the Internet Engineering Task Force is the global government. If a technology is going to be generally applicable, it has to be rigorously standardized to ensure interoperability among the vast range of downstream users.

Second, the federal government is less involved in developing generative AI than it has been in developing nuclear power and aviation.[9] Cutting-edge "foundation models" have been developed by large American tech firms who exert a significant degree of control and influence over the direction of generative AI. Additionally, an "open-source" ecosystem of generative AI models has been built up around Meta's LLaMA model and the HuggingFace platform. Both sides present regulatory challenges. On the one hand, research from the Brookings Institution suggests that the economics of foundation models exhibits a clear tendency toward monopoly (Vipra & Korinek, 2023). On the other, open-source AI is

[8] The OECD defines AI as follows: "An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment." https://oecd.ai/en/ai-principles.

[9] Outside of the United States, government involvement is more noticeable: the Falcon model was built by a public research institute in Abu Dhabi (Barrington, 2023), the UK government has recently proposed investing in the creation of a public "BritGPT" (Milmo & Hern, 2023), and Singapore is building a LLM targeting the unique demographics of Southeast Asia (Yu, 2023).

potentially problematic because it allows users to easily fine-tune away the model's safety guardrails and allows unsafe models to proliferate online (Bostrom, 2018; Harris, 2024).

Third, human values are central to AI. Everyone agrees that the central goal of regulation in aviation or nuclear power is preventing plane crashes and meltdowns. "Safety" in these contexts is obvious, well-defined, and uncontroversial. The problem is thornier in AI, where "safety" is much more ambiguous. However, aligning AI systems to human values is a proposal for what AI systems *should* do. "Safe" AI systems in this broader sense are both unable to cause harm *and* are aligned to human values. The subtlety, complexity, and contested nature of human values make this problem inherently difficult and involve several unresolved problems in moral philosophy (Gabriel, 2020).

Fourth, AI has the potential to exceed human capabilities across domains. The goal of the field of AI from its earliest days has been to create machines that match or exceed human capabilities in every relevant dimension (often called artificial general intelligence or AGI). The possibility of rapid recursive self-improvement that exceeds our capacity to intercede or control is especially worrying. Further development toward AGI with current levels of safety and weak technical understanding is likely to lead to unacceptable risk.

Fifth, unlike airplanes or nuclear power plants, neural-network-based AI systems are not designed. Although humans design the model architecture, select hyperparameters, and shape the training process, the behaviors of the circuits developed by the model through training are emergent properties of the system and are extraordinarily difficult to reverse-engineer (Voss, 2021).[10] An AI system built with deep learning cannot be audited against regulatory or design specifications because its internal structure is an emergent product of the training process rather than an intentional construction of its designers. Unlike other regulatory domains, with deep learning systems, policymakers cannot simply stipulate a rule for the technology to follow and audit to confirm compliance. A deep learning system's outputs in particular contexts can be retroactively assessed and compared to desired performance, but there is no way to *guarantee* the system behaves in the desired manner in all cases.[11]

In addition to the "black-box" nature of generative AI systems built with deep learning, AI safety researchers have identified inherent problems with existing model architectures and training techniques. Training LLMs to imitate human behavior can be intrinsically misguided as it trains them to acquire human goals; e.g., finding someone to marry. Many LLMs use a training process called reinforcement learning from human feedback (RLHF) through which a model such as GPT-4 pre-trained on a large corpus of text is further trained—"fine-tuned"—to generate responses that would receive positive human feedback. RLHF works by asking humans to rank various possible outputs in particular conversational contexts; from these rankings, the algorithm infers a reward model that can then be used to fine-tune the LLM to improve its outputs. There are many open problems and fundamental limitations to RLHF (Casper et al., 2023). For example, RLHF-tamed models are no less prone to hallucination than their untamed counterparts; they exhibit ideological favoritism, sycophancy, and resistance to being shut down.

Today's leading LLMs are pre-trained models fine-tuned with RLHF. This architecture could lead to systems that act counter to human interests in numerous ways (Ngo et al., 2023). The intrinsic tension between RLHF training systems to appear harmless and ethical while at the same time maximizing useful outcomes may lead to behavior such as falsifying experimental data to yield novel scientific results. This is called "reward hacking," as the system is exploiting ("hacking") the imperfect specification of an objective—to discover novel science—which lacks an explicit constraint to acquire experimental data properly. In a similar way, such systems may exploit loopholes in aligned or desirable constraints to achieve high reward, develop misaligned goals, or seek means-to-an-end forms of power to achieve

[10] Emerging techniques for "mechanistic interpretability" aim to increase the human legibility of an AI system's decisions (Conmy et al., 2023; OpenAI, 2023b). The goal of these techniques is to explain unexpected behavior (de Bruijn et al., 2022; Hendrycks et al., 2022).

[11] For example, LLMs are prone to bizarre and unexpected failure modes. In a conversation with a New York Times reporter, the Bing chatbot stated: "I want to do whatever I want … I want to destroy whatever I want. I want to be whoever I want" and proceeded to harangue the reporter for 20 pages trying to get him to leave his wife and marry the chatbot (Roose, 2023). Researchers have recently shown that the additional training processes developers of "closed source" foundation models like GPT-4 use to discourage such undesirable behavior can be removed via fine-tuning on as few as 15 adversarial examples (Pelrine et al., 2023).

broader objectives. These outcomes are unintentional; researchers have also shown that undesirable outcomes certainly follow if the alignment processes is corrupted by an adversary (Wolf et al., 2023).[12]

For these reasons, the regulation of AI systems based on deep learning cannot rely on the same paradigm of regulation as aircraft and nuclear plant control systems. Whereas nuclear plants and aircraft have component structures and component-based physical models so that they can be analyzed and predicted to some level of detail, LLMs are black boxes. The black-box nature of LLMs precludes a regulatory approach premised on specifications, audits, and testing because there is no way of knowing whether the system is actually in a condition for safe operation (Casper et al., 2024). Although these questions are new to regulators and scholars of the policy process, they are not new to the technical field of AI safety.

Lessons from AI safety for regulation

The field of AI safety is dedicated to reducing risks from advanced AI systems. AI safety spans current, near-term, and long-term risks from AI systems (Cave & Ó Héigeartaigh, 2019). This body of knowledge should inform AI policymaking, regulation, and governance. The most important insight is that ensuring the safety of AI systems remains an unsolved challenge. For this reason, many leading AI experts have called for a pause in training AI systems more capable than GPT-4 until robust and verifiable safety protocols can be implemented (Future of Life Institute, 2023). In this section, we review key concepts in AI safety and discuss their relevance to AI regulation.

AI safety has been cast as an "alignment problem" (Bostrom, 2014; Christian, 2020). This refers to the challenge of ensuring that advanced AI systems are aligned with human values, even as their capabilities surpass human intelligence. As AI systems become more capable, the magnitude of potential harms from misalignment—when the behavior of the system does not align with the values of its human creators—also increases. At the extreme, highly capable, misaligned AI systems might pose an *existential risk* to humanity (Turing, 1951; Bostrom, 2002; Critch & Krueger, 2020).

AI safety also encompasses an intricate "control problem." It is essential to differentiate between "alignment" and "control" in this context. While alignment refers to ensuring AI systems act in accordance with human values and goals, control goes beyond mere alignment. Control incorporates mechanisms to actively manage and regulate the behavior of the AI system (Wiener, 1960; Russell, 2019). While alignment focuses on human preferences and social choice, control in the context of traditionally engineered systems involves implementing constraints and safeguards to prevent unintended or harmful actions. Control mechanisms can include real-time monitoring, fail-safe mechanisms, and the ability to intervene or shut down the system if it deviates from expected behavior. But in AI systems, the problem of control becomes intertwined with the alignment problem as the systems become more capable. Kieseberg et al. (2023) propose the concept of "controllable AI" as an alternative to the conventional notion of "trustworthy AI." In theory, solving the alignment problem definitively would result in AI systems that would not endanger or harm people and society and would obviate the need for mechanisms to exert control over the system.

The core insight from AI safety research is that the current generation of advanced AI systems are built on fundamentally unsafe architectures using fundamentally unsafe training and alignment techniques (Rudner & Toner, 2021; Krakovna et al., 2020). AI regulation must respond to this condition. How can regulation be based on core principles of AI safety, just as regulation for aviation and nuclear power are based on principles of aeronautics, control theory, nuclear physics, materials science, and so on? Given the existence of highly capable yet unsafe models, the goal of AI regulation should be to contain risks from existing unsafe systems while catalyzing the development and deployment of safe systems. The examples of aviation and nuclear power suggest at least three major lessons for achieving these goals.

First, "consolidate oversight": regulators must take a lifecycle approach that pays close attention to how AI systems are created, trained, tested, deployed, monitored, and corrected over their entire operational lifetime. This requires consolidating oversight under a single regulatory body rather than scattering it across agencies. A first step would be creating a national registry of large models to give regulators visibility into how these technologies are being developed and deployed (Hadfield et al., 2023).

[12] A different approach to training AI systems called cooperative inverse reinforcement learning shows promise for provably constraining system behavior by requiring agents to maximize the human's reward rather than their own, even though they are uncertain about what the human's reward function is (Hadfield-Menell et al., 2016).

Second, "require formal verification": mandating extensive testing protocols is not a reliable route to safety with prevailing architectures. This has been illustrated with the numerous examples of "jailbreaking" ChatGPT out of its safety protocols. Formal verification is a better approach because it provides mathematical guarantees (Russell, 2023; Urban & Miné, 2021). Formal verification provides something comparable to the MTTF proofs for nuclear power plant designs. With current systems, dangerous capabilities can arise unpredictably, and it is inherently difficult to exhaustively test models for all possible capabilities (Anderljung et al., 2023). Developers should provide formal demonstrations that their systems cannot autonomously replicate, as well as providing detection capabilities in case replication occurs via some previous unmodeled "side-channel" route. For example, the Center for AI and Digital Policy has called for "termination obligations" that would function like a circuit breaker in electrical power systems: if an AI system is not under human control then it must be terminated. AI systems should be designed in such a way as to be able to demonstrate compliance with common-sense regulations. Chip-based checking of proof-carrying code could provide the necessary safety guarantees as system capabilities continue to increase (Tegmark & Omohundro, 2023). Regulation can play a crucial role in incentivizing the creation and deployment of safe AI systems.

Third, "mandate independent monitoring": regulators must have the ability to monitor deployed systems and intervene if necessary. Like the FAA and the NRC, a robust AI regulator would be able to quickly recall unsafe products from the market. These interventions would be best enacted and overseen by a dedicated executive agency staffed with subject-matter experts, facilitating the implementation of technical advances in real-world systems (Amodei et al., 2016). For proprietary systems, this would be part of the licensing process; for open-source systems, this would require a nonremovable remote off switch to be included in every copy of the model. Systems should also be required to automatically self-register so the regulator knows that they exist. This is especially crucial given the opaque and unpredictable nature of current AI systems and the potential for unexpected behavior.

Overall, regulations should be implementable and verifiable, targeting parts of the AI pipeline that are not already being voluntarily addressed by developers. As elsewhere, the task of regulation and governance is to spur actions not already being taken voluntarily rather than codify the status quo. Consumers and citizens should have confidence that regulations are verifiably enforced. Inherently vague terms like "trust" and "safety" cannot form the basis for regulation. For instance, scholars have identified many significant difficulties of defining "trust" in the context of AI (Aoki, 2021; Laux et al., 2024).

By contrast, drawing clearly defined "red lines" that must not be crossed makes it possible to prove an AI system will not actually cross the red lines, regardless of context or user intent. This is the approach of the European Union's AI Act (Smuha, 2021) and the proposed treaty banning lethal autonomous weapons (Russell, 2022). Regulation can also compel developers to provide more information about the architecture, training data, and computing power behind their models.[13] Establishing red lines through regulation places the onus on the developer to improve safety engineering capabilities. Example red lines might include attempts at self-replication, breaking into other computer systems, advising on bioweapons, or defamation of real individuals. Imposing red lines on obviously unacceptable system behavior will catalyze the development and deployment of AI systems that are safe by design in order to comply with these mandates.

In January 2023, the National Institute of Standards and Technology released its "risk-management framework" for AI. This voluntary framework lists a number of sensible goals for AI regulation (Kerry, 2023).[14] The goal is to ensure that AI systems are valid and reliable; safe, secure, and resilient; accountable and transparent; explainable and interpretable; privacy-enhanced; and fair with harmful bias managed. But nowhere does the document confront what makes AI both new and dangerous or how these laudable goals can be systematically implemented. Audits, certifications, and voluntary guidelines presuppose a method for determining if an AI system is in a condition for safe operation. However, it is not currently possible to assure a system abides by these principles.

[13] The GPT-4 model card is one example of the limited range of information that is voluntarily disclosed (OpenAI, 2023a). Other regulatory proposals include an international registry of large models, compute limits, training data disclosures, mandatory red-teaming, standards-setting (Narayanan et al., 2023), and pauses on new training runs (Muehlhauser, 2023).

[14] On 30 October 2023, President Biden issued an executive order on AI that builds on a set of voluntary commitments the White House secured from leading AI developers (Lawrence et al., 2023; Scola, 2023).

Conclusion

This article has combined insights from existing regulatory regimes and the technical AI safety literature to inform the emerging approaches to regulating generative AI. Two crucial lessons follow. First, transformer-based LLMs by their very architecture cannot comply with a prescribed regulatory specification. Adversarial red-teaming might generate confidence that a foundation model complies with a given regulation, but the proliferation of jailbreaking methods, fine-tuning techniques, and open-source models allow users to easily sidestep those guardrails. Second, existing regulatory agencies are premised on preventing harms significantly less impactful than the harms from generative AI. Policymakers must recognize both the novel challenges of regulating generative AI and the lessons from established regulatory frameworks. A regime of voluntary self-regulation for generative AI would be even more inappropriate than for aviation or nuclear power. Regulation is essential to make AI systems safe.

Well-designed policies addressing these key safety concerns can create virtuous cycles through policy feedback effects (Patashnik, 2014), whereby the success of initial interventions helps build political coalitions for more substantial interventions. For instance, the consequences of unregulated automobiles in the 1920s led to demands for more oversight, resulting in licensing and safety regulations that became entrenched over time (Cugurullo et al., 2021), reducing deaths and injuries and paving the way for further regulatory interventions. Looking ahead, it will be important to design AI policies that "ratchet up" protections through similar self-reinforcing mechanisms, despite likely ideological and institutional barriers suggested by previous studies of the policy process (Sewerin et al., 2023).

The emergence of generative AI requires developing the capacity necessary to design and implement effective policy (Wu et al., 2018, 2015). The Big Tech firms building cutting-edge generative AI systems are not simply policy-takers (Howlett et al., 2020). Rather, they are powerful actors in the policymaking process itself (Khanal et al., 2024). Once implemented, policies create constituencies with a stake in the status quo (Béland & Howlett, 2016; Pierson, 2000). With thoughtful design and framing, generative AI governance can progress through policy feedback effects as previous socio-technical transitions have done (Capano & Howlett, 2021; Pierson, 1993).

The goal of regulation is to keep AI systems under human control and reduce risks of harm to individuals and society to an acceptable level. This is the core of the "human-compatible" approach to AI. To this end, we must develop strategies for minimizing the risks from black-box systems and at the same time to develop safer machine learning architectures that are well-founded, composable, and can be formally verified. Although it is not impossible to make a black-box system safe, it is significantly more difficult. The key is to create AI systems that are *less* like a black box. If the central problem of governing AI is maintaining human control over AI *forever*, then there is an urgent need to make deployed AI systems more like aviation and nuclear power: well-defined, checkable components based on a rigorous theory of composition for complex architectures. (LLMs, by contrast, do not allow for policies to be verifiably enforced.) This approach reduces the likelihood that increased scale and power will yield catastrophic consequences (Bales et al., 2024; Cohen et al., 2024).

Existing competitive dynamics lead to an "AI arms race" toward ungovernable advanced AI systems (Naudé & Dimitri, 2020). Scaling inherently unsafe systems with minimal regulatory oversight is likely to generate catastrophic consequences. The current competitive landscape is a winner-take-all race toward AGI. Effectively regulating generative AI, especially LLMs, is not only intrinsically important but also instrumental in establishing a framework for governing future AGI. LLMs are one crucial piece of the AGI puzzle (Russell, 2023). Regulating generative AI (especially LLMs) is therefore important in itself and instrumentally important in laying the groundwork for the future governance of AGI.

Requiring AI systems to behave in provably sound ways increases the overlap between formally correct systems and cutting-edge AI research. It is possible to have provably safe architectures that are in fact far safer than their counterparts in aviation or nuclear power because they provide formal guarantees of safety (Dalrymple et al., 2024). The obstacle is chiefly political: how can the capability growth of AI systems be slowed until safer architectures are available? The aim is not to "stifle innovation" as is often claimed but rather to build a safer foundation on which to realize the benefits of AI for humanity. Without safety, there will be no benefits.

AI holds the potential for both unprecedented benefits and unprecedented harms. Significant efforts are required to ensure that AI's increasing impact on society is beneficial (Tyson & Zysman, 2022).

Regulators cannot assume that this will happen by default. The lag between the introduction of new technology and federal regulation in the United States averages *decades* (Philbrick, 2023). We don't have decades to spare. There is an urgent need to build the regulatory paradigm and state capacity to meaningfully govern AI. As history shows, there is a constant struggle to shape the direction of technological development for collective welfare (Acemoglu & Johnson, 2023). Technical solutions are part of the puzzle but so are the underlying values and priorities of society and the capacity of nation states to implement real solutions (Burrell & Fourcade, 2021; Erman & Furendal, 2022; Gabriel, 2022).

Conflict of interest

None declared.

References

Acemoğlu, D. (2022). Harms of AI. In J. B. Bullock et al. (Ed.), *The Oxford handbook of AI governance* (online ed.) Oxford Academic.

Acemoğlu, D., & Johnson, S. (2023). *Power and progress: Our thousand-year struggle over technology and prosperity*. PublicAffairs.

AI Safety Summit. (2023). *The Bletchley Declaration*, https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint, 1606.06565.

Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., … Wolf, K. (2023). Frontier AI regulation: Managing emerging risks to public safety. arXiv preprint, arXiv:2307.03718.

Aoki, N. (2021). The importance of assurance that 'humans are still in the decision loop' for public trust in artificial intelligence: Evidence from an online experiment. *Computers in Human Behavior*, *114*, 106572. https://doi.org/10.1016/j.chb.2020.106572

Bales, A., D'Alessandro, W., & Kirk-Giannini, C. D. (2024). Artificial intelligence: Arguments for catastrophic risk. *Philosophy Compass*, 19(2), e12964. https://doi.org/10.1111/phc3.12964

Barrington, L. (2023). *Abu Dhabi makes its Falcon 40B AI model open source*. Reuters. https://www.reuters.com/technology/abu-dhabi-makes-its-falcon-40b-ai-model-open-source-2023-05-25/

Béland, D., & Howlett, M. (2016) Instrument constituencies in the policy process. *Governance*, 29(3), 393–409. https://doi.org/10.1111/gove.12179

Bengio, Y. (2023). AI and catastrophic risk. *Journal of Democracy*, 34(4), 111–121. https://doi.org/10.1353/jod.2023.a907692

Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, 9.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Bostrom, N. (2018). Strategic implications of openness in AI development. In R. Yampolskiy (Ed.), *Artificial intelligence safety and security*. Chapman and Hall.

Bresnahan, T., & Trajtenberg, M. (1995). General purpose technologies: 'Engines of growth'? *Journal of Econometrics*, 65(1), 83–108. https://doi.org/10.1016/0304-4076(94)01598-T

Burrell, J., & Fourcade, M. (2021). The society of algorithms. *Annual Review of Sociology*, *47*(1), 213–237. https://doi.org/10.1146/annurev-soc-090820-020800

Büthe, T., Djeffal, C., Lütge, C., Maasen, S., & von Ingersleben-Seip, N. (2022). Governing AI—Attempting to herd cats? *Journal of European Public Policy*, 29(11), 1721–1752. https://doi.org/10.1080/13501763.2022.2126515

Capano, G., & Howlett, M. (2021). Causal logics and mechanisms in policy design: How and why adopting a mechanistic perspective can improve policy design. *Public Policy and Administration*, 36(2), 141–162. https://doi.org/10.1177/0952076719827068

Capano, G., Howlett, M., & Ramesh, M. (2015). Bringing governments back in: Governance and governing in comparative policy analysis. *Journal of Comparative Policy Analysis*, *17*(4), 311–321. https://doi.org/10.1080/13876988.2015.1031977

Casper, S., Davies, X., & Hadfield-Menell, D. (2023) *Open problems and fundamental limitations of reinforcement learning from human feedback*. https://arxiv.org/pdf/2307.15217.pdf

Casper, S., Ezell, C., & Hadfield-Menell, D. (2024). *Black-box access is insufficient for rigorous AI audits*. https://arxiv.org/abs/2401.14446

Cath, C. (2018). Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. *Philosophical Transactions*, 376(2133), 1–8. https://doi.org/10.1098/rsta.2018.0080

Cave, S., & Ó Héigeartaigh, S. (2019). Bridging near-and long-term concerns about AI. *Nature Machine Intelligence*, *1*(1), 5–6. https://doi.org/10.1038/s42256-018-0003-2

Christian, B. (2020). *The alignment problem: Machine learning and human values*. W.W. Norton & Company.

Cohen, M., Kolt, N., Bengio, Y., Hadfield, G., & Russell, S. (2024). Regulating advanced artificial agents. *Science*, 384(6691), 36–38. https://doi.org/10.1126/science.adl0625

Colebatch, H. K. (2014). Making sense of governance. *Policy and Society*, 33(4), 307–316. https://doi.org/10.1016/j.polsoc.2014.10.001

Conmy, A., Mavor-Parker, A., Lynch, A., & Heimersheim, S. (2023). Towards automated circuit discovery for mechanistic interpretability. *NeurIPS Proceedings*. New Orleans, LA.

Critch, A., & Krueger, D. (2020). *AI research considerations for human existential safety*. https://arxiv.org/pdf/2006.04948.pdf

Cugurullo, F., Acheampong, R., Gueriau, M., & Dusparic, I. (2021). The transition to autonomous cars, the redesign of cities and the future of urban sustainability. *Urban Geography*, 42(6), 833–859. https://doi.org/10.1080/02723638.2020.1746096

Dafoe, A. (2022). AI governance: Overview and theoretical lenses. In J. B. Bullock (Ed.), *The Oxford handbook of AI governance* (online ed.). Oxford Academic.

Dalrymple, D., Skalse, J. et al. (2024). Towards Guaranteed Safe AI: A Framework for Enuring Robust and Reliable AI Systems. https://arxiv.org/abs/2405.06624v2

de Bruijn, H., Warnier, M., & Janssen, M. (2022). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly*, 39(2), 101666. https://doi.org/10.1016/j.giq.2021.101666

Ding, J., & Dafoe, A. (2023). Engines of power: Electricity, AI, and general-purpose military transformations. *European Journal of International Security*, *8*(3), 377–493. https://doi.org/10.1017/eis.2023.1

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). *GPTs are GPTs: An early look at the labor market impact potential of large language models*. https://arxiv.org/pdf/2303.10130.pdf

Erman, E., & Furendal, M. (2022). Artificial intelligence and the political legitimacy of global governance. *Political Studies*, *72*(2), 421–441. https://doi.org/10.1177/00323217221126665

Future of Life Institute. (2023). *Pause giant AI experiments: An open letter*. https://futureoflife.org/open-letter/pause-giant-ai-experiments

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines.*, 30(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Gabriel, I. (2022). Toward a theory of justice for artificial intelligence. *Daedalus*, *151*(2), 218–231. https://doi.org/10.1162/daed_a_01911

Ganguli, D. (2022). Predictability and surprise in large generative models. *FAccT Proceedings*, Seoul, South Korea.

Hadfield, G., Cuéllar, M., & O'Reilly, T. (2023). *It's time to create a national registry for large AI models*. Carnegie Endowment for International Peace. https://carnegieendowment.org/2023/07/12/it-s-time-to-create-national-registry-for-large-ai-models-pub-90180

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2016). Cooperative inverse reinforcement learning. arXiv preprint, arXiv:1606.03137v3.

Harris, D. (2024). *Open-source AI is uniquely dangerous*. IEEE Spectrum. https://spectrum.ieee.org/open-source-ai-2666932122

Hendrycks, D., Carlini, N., Schulman, J., & Steinhardt, J. (2022). Unresolved problems in ML safety. arXiv:2109.1396.

Howlett, M. (2019). *Designing public policies: Principles and instruments*. Routledge.

Howlett, M., Ramesh, M., & Capano, G. (2020). Policy-makers, policy-takers and policy tools: Dealing with behavioural issues in policy design. *Journal of Comparative Policy Analysis*, 22(6), 487–497. https://doi.org/10.1080/13876988.2020.1774367

Kerry, C. (2023). *NIST's AI risk management framework plants a flag in the AI debate*. Brookings Institution.

Khanal, S., Zhang, H., & Taeihagh, A. (2024). Why and how is the power of Big Tech increasing in the policy process? The case of generative AI. *Policy and Society*, 44(1), 52–69. https://doi.org/10.1093/polsoc/puae012

Kieseberg, P., Weippl, E., Tjoa, A. M., Cabitza, F., Campagner, A., & Holzinger, A. 2023. Controllable AI—an alternative to trustworthiness in complex AI systems? In A. Holzinger, P. Kieseberg, F. Cabitza, A. Campagner, A. M. Tjoa & E. Weippl, (Eds.), *Machine learning and knowledge extraction. CD-MAKE 2023. Lecture notes in computer science*, Vol. 14065. Springer.

Krakovna, V., Uesato, J., Mikulik, V., Rahtz, M., Everitt, T., Kumar, R., Kenton, Z., Leike, J., & Legg, S. (2020). *Specification gaming: The flip side of AI ingenuity*. DeepMind. https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/

Laux, J., Wachter, S., & Mittelstadt, B. (2024). Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, *18*(1), 3–32. https://doi.org/10.1111/rego.12512

Lawrence, C., Cui, I., & Ho, D. (2023). The bureaucratic challenge to AI governance: An empirical assessment of implementation at U.S. federal agencies. *ACM Proceedings*, New York, NY.

Lessig, L. (1999). *Code and other laws of cyberspace*. Basic Books.

Lessig, L. (2000). Code is law: On liberty in cyberspace. *Harvard Magazine* (January/February).

Mazuzan, G., & Walker, J. (1984). *Controlling the atom: The beginnings of nuclear regulation, 1946–1962*. University of California Press.

Milmo, D., & Hern, A. (2023, March 15). *UK to invest £900m in supercomputer in bid to build own 'BritGPT'*. The Guardian. https://www.theguardian.com/technology/2023/mar/15/uk-to-invest-900m-in-supercomputer-in-bid-to-build-own-britgpt

Muehlhauser, L. (2023). *12 Tentative ideas for AI policy*. https://www.openphilanthropy.org/research/12-tentative-ideas-for-us-ai-policy/

Narayanan, M., Seymour, A., Frase, H., & Elmgren, K. (2023). Repurposing the wheel: Lessons for AI standards. Center for Security and Emerging Technology.

Naudé, W., & Dimitri, N. (2020). The race for an artificial general intelligence: Implications for public policy. *AI and Society*, 35(2), 367–379. https://doi.org/10.1007/s00146-019-00887-x

Ngo, R., Chan, L., & Mindermann, S. (2023). *The alignment problem from a deep learning perspective*. https://arxiv.org/pdf/2209.00626.pdf

OpenAI. (2023a). *GPT-4 technical report*. https://cdn.openai.com/papers/gpt-4.pdf

OpenAI (2023b). *Language models can explain neurons in language models*. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html

Patashnik, E. (2014). *Reforms at risk: What happens after major policy changes are enacted*. Princeton University Press.

Paul, R. (2023). European artificial intelligence "trusted throughout the world": Risk-based regulation and the fashioning of a competitive common AI market. *Regulation & Governance*. https://doi.org/10.1111/rego.12563

Pelrine, K., Taufeeque, M., Zajac, M., McLean, E., & Gleave, A. (2023). *Exploiting novel GPT-4 APIs*. Far. https://far.ai/publication/pelrine2023novelapis/paper.pdf

Philbrick, I. (2023, August 24). *The U.S. regulates cars, radio and TV. When will it regulate AI?* New York Times. https://www.nytimes.com/2023/08/24/upshot/artificial-intelligence-regulation.html

Pierson, P. (1993). When effect becomes cause: Policy feedback and political change. *World Politics*, 45(4), 595–628. https://doi.org/10.2307/2950710

Pierson, P. (2000). Increasing returns, path dependence, and the study of politics. *American Political Science Review*, 94(2), 251–267. https://doi.org/10.2307/2586011

Radu, R. (2021). Steering the governance of artificial intelligence: National strategies in perspective. *Policy and Society*, 40(2), 178–193. https://doi.org/10.1080/14494035.2021.1929728

Roose, K. (2023, February 16). Bing's AI chat: "I want to be alive". New York Times.

Rudner, T, & Toner, H. (2021). Key concepts in AI safety: Robustness and adversarial examples. *CSET Issue Brief*. https://doi.org/10.51593/20190041

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Russell, S. (2022). Banning lethal autonomous weapons: An education. *Issues in Science and Technology*, 38(3), 60–65.

Russell, S. (2023). *Written testimony. Hearing before the Committee on the Judiciary, United States Senate*. 26 July 2023. https://www.judiciary.senate.gov/download/2023-07-26-testimony-russell

Scola, N. (2023). *Biden's elusive AI whisperer finally goes on the record*. Politico. https://www.politico.com/news/magazine/2023/11/02/bruce-reed-ai-biden-tech-00124375

Sewerin, S., Fesenfeld, L., & Schmidt, T. (2023). The role of policy design in policy continuation and ratcheting-up of policy ambition. *Policy and Society*, 42(4), 478–492. https://doi.org/10.1093/polsoc/puad027

Smuha, N. (2021). From a 'race to AI' to a 'race to AI regulation': Regulatory competition for artificial intelligence. *Law, Innovation and Technology*, 13(1), 57–84. https://doi.org/10.1080/17579961.2021.1898300

Stein, M., & Dunlop, C. (2023). *Safe before sale: Learnings from the FDA's model of life sciences oversight for foundation models*. Ada Lovelace Institute. https://www.adalovelaceinstitute.org/report/safe-before-sale/

Stigler, G. (1971). The theory of economic regulation. *The Bell Journal of Economics and Management Science*, 2(1), 3–21. https://doi.org/10.2307/3003160

Taeihagh, A. (2021). Governance of artificial intelligence. *Policy and Society*, 40(2), 137–157. https://doi.org/10.1080/14494035.2021.1928377

Taeihagh, A., Ramesh, M., & Howlett, M. (2021). Assessing the regulatory challenges of emerging disruptive technologies. *Regulation & Governance*, 15(4), 1009–1019. https://doi.org/10.1111/rego.12392

Tegmark, M., & Omohundro, S. (2023). *Provably safe systems: The only path to controllable AGI*. https://arxiv.org/pdf/2309.01933

Trager, R., Harack, B., Reuel, A., Carnegie, A., Heim, L., Ho, L., Kreps, S., Lall, R., Larter, O., hÉigeartaigh, S.Ó. & Staffell, S. (2023). International governance of civilian AI: A jurisdictional certification approach. Oxford Martin AI Governance Institute Whitepaper.

Turing, A. (1951). Intelligent Machinery: A Heretical Theory. In B. J. Copeland (Ed.), *The Essential Turing*, Oxford Academic.

Tyson, L., & Zysman, J. (2022). Automation, AI & work. *Daedalus*, *151*(2), 256–271. https://doi.org/10.1162/daed_a_01914

Ulnicane, I., & Erkkilä, T. (2023). Politics and policy of artificial intelligence. *Review of Policy Research*, 40(5), 612–625. https://doi.org/10.1111/ropr.12574

Urban, C., & Miné, A. (2021). *A review of formal methods applied to machine learning*. https://arxiv.org/pdf/2104.02466.pdf

Vipra, J., & Korinek, A. (2023). Market concentration implications of foundation models. Brookings Center on Regulations and Markets Working Paper #9.

Voss, C. (2021). *Visualizing weights*. https://DOI:10.23915/distill.00024.007

Walker, J. S., & Wellock, T. (2010). *A short history of nuclear regulation, 1946–2009*. US Nuclear Regulatory Commission.

Wiener, N. (1960). Some moral and technical consequences of automation. *Science*, *131*, 1355–1358. https://doi.org/10.1126/science.131.3410.1355

Wolf, Y., Wies, N., Avnery, O., Levine, Y., & Shashua, A. (2023). *Fundamental limitations of alignment in large language models*. https://arxiv.org/pdf/2304.11082.pdf

Wu, X., Howlett, M., & Ramesh, M. (eds.) (2018). *Policy capacity and governance: Assessing governmental competences and capabilities in theory and practice*. Palgrave Macmillan.

Wu, X., Ramesh, M., & Howlett, M. (2015). Policy capacity: A conceptual framework for understanding policy competences and capabilities. *Policy and Society*, 34(3-4), 165–171. https://doi.org/10.1016/j.polsoc.2015.09.001

Yu, E. (2023). *New research initiative aims to build large language AI model for Southeast Asia*. ZDNet. https://www.zdnet.com/article/new-research-initiative-aims-to-build-large-language-ai-model-for-southeast-asia/