TAX LEVERS FOR A SAFER AI FUTURE

Mirit Eyal[e] *& Yonathan Arbel*[y]

This Article argues that tax policy can become a powerful tool for the development of safer systems of artificial intelligence (AI). Investment in AI capabilities is at a fever pitch, drawing capital, talent, and computing resources from most sectors of the economy. While the development of capable AI systems promises princely rewards to their creators, investment in safety remains anemic, reduced to paltry budgets and safety-washing initiatives. This misalignment has produced a rapidly expanding capability-safety gap: the difference between what these systems can do and what they can do safely. We lack meaningful assurances that tomorrow's powerful systems will safeguard the life and dignity of individuals, withstand adversarial attacks, and function reliably in novel contexts. At the heart of this gap lies a simple, bitter truth—while the rewards from powerful models are private, the harms are socialized.

We term this the social misalignment problem and propose that tax levers can play a critical role in its resolution. Our proposed framework reconceptualizes the existing sprawling system of R&D credits to incentivize investments in AI safety. The framework integrates four mechanisms: (1) targeted rewards for basic and applied research on AI safety; (2) consumer credits for purchases of safe AI technology; (3) escalating tax penalties for non-compliance; and (4) redistribution of penalty-generated revenue to public safety research initiatives. This Article argues that such a framework offers a practical solution for embedding safety considerations within the economic architecture of AI development while preserving innovation incentives. Through careful calibration of fiscal levers, public policy tools can address the structural misalignment between private sector imperatives and public safety demands in emerging technologies.

[e] Joseph D. Peeler Professor of Law, The University of Alabama School of Law.

[y] William Alfred Rose Professor of Law, Irving Silver and Frances Grodsky Silver Faculty Scholar, Director, AI & Law Studies, The University of Alabama School of Law.

TABLE OF CONTENTS

Introduction

The race to develop artificial general intelligence (AGI)—systems matching human performance across diverse tasks—has become a defining technological pursuit of our era.[1] As industry leaders edge closer to this milestone, and even speculate about artificial *super*intelligence (ASI),[2] the stakes of AI safety grow considerably.[3] How can we ensure that a system general enough to perform most human tasks, a system that most expect to be socially transformative, will also not lead to unintended accidents and harms on a massive scale?[4] While private entities pour unprecedented resources into advancing AI capabilities, investments in safety research languish, creating a quickly expanding "safety-capability" gap.[5] The consequences of the gap between what AI can do and what AI can can do safely are not abstract: vulnerabilities in frontier systems—from susceptibility to adversarial attacks to emergent misalignment with human values—threaten individual rights, democratic institutions, and global stability. This divergence stems not from scientific impossibility of building safer systems, but from structural incentives. The rewards of powerful AI flow to developers; the risks cascade across society.

This Article argues that fiscal policy can play a key role in closing this gap. Building on Hemel and Ouellette's Innovation Policy Pluralism framework,[6] which positions taxation as a critical yet underutilized lever for steering technological progress, we propose a novel approach: embedding safety imperatives directly into the economic architecture of AI development. Unlike traditional regulation, which struggles to govern fast-evolving technologies through rigid mandates, tax

---

[1] There is no accepted definition of neither intelligence nor general intelligence, but the various definitions revolve around the ability to perform a broad variety of tasks that would normally require human intelligence, Tao Feng et al., *How Far Are We From AGI: Are LLMs All We Need?*, TRANSACTIONS ON MACHINE LEARNING RESEARCH (Oct. 2024), https://perma.cc/JE77-FK6D. The estimates of our distance to this poorly defined goal of AGI shift rapidly. In 2010, AGI was predicted to be 50 years into the future; by 2023, this has shifted to 5-20 years. The most critical of industry leaders, Yann LeCunn (Meta) predicts "5-10 years if everything goes great" https://perma.cc/97W6-GKP6.

[2] Sam Altman (OpenAI) predicts ASI in "A few thousand days" into the future https://perma.cc/RK39-67H9. Elon Musk, (whose predictions often tend to prove overly optimistic), predicted ASI by 2025. Reuters, T*esla's Musk predicts AI will be Smarter than the Smartest Human Next Year,* Reuters (Apr. 8, 2024) https://www.reuters.com/technology/teslas-musk-predicts-ai-will-be-smarter-than-smartest-human-next-year-2024-04-08/.

[3] *See* STUART J. RUSSELL, HUMAN COMPATIBLE: ARTIFICIAL INTELLIGENCE AND THE PROBLEM OF CONTROL 23–30 (2019) (discussing the necessity of aligning AI systems with human values to prevent unintended consequences as they become more powerful).

[4] Yoshua Bengio et al., *Managing Extreme AI Risks Amid Rapid Progress*, 384 SCIENCE 842 (2024), https://doi.org/10.1126/science.adn0117. A survey in 2024 of 2,778 of AI researchers who published in leading outlets found that median estimates of *existential* risk from AI development ranges between 5% to 10%, see Katja Grace et al., *Thousands of AI Authors on the Future of AI*, ARXIV (Jan. 5, 2024), https://perma.cc/XFF7-686Q.

[5] Another name for this in the AI safety community is "safety tax," that developers pay for ensuring the safety of their AI systems. *See* markovial etl al., *Alignment Tax*, AI Alignment Forum (Dec. 30, 2024) https://www.alignmentforum.org/w/alignment-tax. A recent analysis of published AI safety research concludes that there is too little investment. Oscar Delaney, Oliver Guest & Zoe Williams, *Mapping Technical Safety Research at AI Companies: A Literature Review and Incentives Analysis*, at 2 (2024). While safety and capabilities often trade off against each other, the full relationship is more complex Dan Hendrycks & Mantas Mazeika, *X-Risk Analysis for AI Research*, ARXIV 8 (Sept. 20, 2022), https://doi.org/10.48550/arXiv.2206.05862 (recommending attempting to "improve a safety-capabilities ratio").

[6] See *infra* note 23 at 553.

incentives harness firm in-house knowledge while mitigating regulatory capture and expertise asymmetries. By rewarding safety-aligned research, penalizing reckless capability acceleration, and redistributing penalty revenue to public safety initiatives, a tax-based approach aligns private profit motives with social welfare imperatives, without stifling innovation. And while the application of this framework is novel, we demonstrate its political feasibility by drawing on extensive precedents already embedded in the tax system.[7]

The urgency of this intervention is underscored by recent regulatory failures. On his first day in office, President Trump revoked the executive order meant to control AI safety risks left by his predecessor.[8] Efforts at the state level have also been unavailing. The most ambitious of those, California's SB-1047, was vetoed by California governor Gavin Newsom.[9] His veto concedes that "[s]afety protocols must be adopted" and that "we cannot afford to wait for a major catastrophe to occur before taking action to protect the public."[10] The problem, Newsom stated, was the lack of sufficient "empirical trajectory analysis" to support the law.[11] But the hopes for rigorous empirical testing conflict directly with the reality that the technology is developed in breakneck speeds, behind the curtain of AI labs, and involves opaque systems. Whatever empirical trajectory analysis means, it is clear that no one had been able to produce one in the last few years, and the vetoing of regulatory mechanisms like the California Bill undermines our ability to gather exactly this type of information.

Nor has self regulation proved robust. In 2023, OpenAI's promoted a much-touted "Superalignment" initiative—promising 20% of computing resources for safety research.[12] This initiative collapsed under competitive pressures, with internal reports revealing diverted GPU allocations and hollow compliance.[13] This pattern pervades the industry: safety teams operate on

---

[7] In the pharmaceutical industry, for instance, tax credits for orphan drugs and safety-related research have successfully encouraged firms to invest in underfunded areas like rare disease treatment, even when such investments are not immediately profitable. 26 U.S.C. § 45C; Orphan Drug Act, Pub. L. No. 97-414, 96 Stat. 2049 (1983) (codified as amended at 21 U.S.C. section 360cc). It also allowed unused credit to be offset past and future tax liability through carryback and carryforward features. Taxpayer Relief Act of 1997, Pub. L. 105-34, § 604, 111 Stat. 788, 863 (1997) (allowing carryback three years and carry forward up to fifteen years). The use of the term "Orphan" refers to drugs for rare diseases and conditions that entail limited opportunities for pharmaceutical and biotechnology companies to undertake their development and production. See Orphan Drug Act, Pub. L. No. 97-414, section 1(b), 96 Stat. 2049, 2049 (1983) (providing an overview on the environment of research in the area of rare conditions and diseases). Similarly, in aerospace, strict regulatory requirements coupled with financial incentives for safety improvements have led to significant advancements in aircraft reliability and passenger safety. *See, e.g.,* Federal Aviation Administration, Regulations & Policies, FAA, https://perma.cc/BYA4-MZZX (last visited Jan. 16, 2025).

[8] *See* David Shepardson, *Trump Revokes Biden Executive Order on Addressing AI Risks*, REUTERS (Jan. 21, 2025), https://perma.cc/24QM-Z4HE.

[9] Cal. S.B. 1047, 2023-2024 Reg. Sess. (Cal. 2024).

[10] Cal. Governor Veto Message, S.B. 1047, 2023-2024 Reg. Sess. (Cal. Sept. 29, 2024) https://perma.cc/LJ2N-WN5V.

[11] *Id*.

[12] *See* OpenAI, *Introducing Superalignment*, OPENAI BLOG (July 6, 2023) https://perma.cc/U8YL-99R5. Kai Xiang Teo, *OpenAI is so worried about AI causing human extinction, it's putting together a team to control "superintelligence"*, Business Insider (July 7, 2023)

[13] *See* Tom Simonite, *OpenAI's Superalignment Team Disbanded*, WIRED (May 24, 2024), https://perma.cc/7U8Y-MJ4C. *See also* Jan Leike (@janleike), Post on X (formerly Twitter), May 17, 2024, https://perma.cc/8PND-DVD2. Peter N. Salib, *OpenAI No Longer Takes Safety Seriously*, LAWFARE (May 22, 2024), https://perma.cc/CXZ9-VSDF.

shoestring budgets while capability divisions command vast resources. The root cause is clear. Under current market and regulatory conditions, the incentives of firms are misaligned; safety might produce public benefits, but it leeches resources firms would rather spend on the race to own the most capable system.

While we view tax incentives as part of the solution, we note that currently tax instruments are part of the problem.[14] Research incentive mechanisms, such as R&D credits and expensing provisions,[15] exacerbate the misalignment by subsidizing capability research indiscriminately.[16] Our proposal confronts these challenges through three interlocking mechanisms: (1) *producer-side incentives*, such as safety R&D credits and differentiated expensing that accelerates deductions for safety research while amortizing pure capability investments; (2) *consumer-side safety-linked tax credits* contingent on verifiable adherence to NIST or ISO benchmarks; and (3) *Corrective tax penalties* scaled to the social cost of unsafe deployments, with revenues funding public-sector safety consortia. [17]

Tax agencies enjoy an important, if neglected, advantage in AI safety regulation: institutional competence. The playing field in AI is tilted against regulators; any attempt to regulate, even if it could overcome the lobby and political pressure to win the AI race, would still need to deal with the reality that AI labs have vastly more funding, expertise, computing resources, and technical talent than government agencies.[18] Relative to other agencies, however, tax agencies possess established competencies in auditing complex R&D claims, offering a foundation for effective oversight of tax-based safety claims.[19] This administrative framework can be enhanced through mandatory third-party validation of safety benchmarks (e.g., adversarial safety scores, sandbox and red teaming

---

[14] *See infra* Part II and III.C. *See* Lucy Colback, *AI and the R&D Revolution*, FIN. TIMES (Nov. 26, 2024), https://perma.cc/JN32-NBUY (noting that while significant investments are made in R&D, the rapid pace of technological advancement can lead to challenges in effectively managing and mitigating associated risks.).

[15] Throughout, we use "research and development" and "research and experimentation" interchangeably, although we note the latter is more restrictive than the former and does not necessarily specific immediate commercial applications. *See* Stephen E. Shay, J. Clifton Fleming, Jr., & Robert J. Peroni, *R&D Tax Incentives: Growth Panacea or Budget Trojan Horse?,* 69 TAX L. REV. 419, 422 n.15 (2016).

[16] R&D credits are also criticized for favoring major private corporations that have the know-how relevant to exploiting these credits. *See, e.g.,* OECD, *Measuring Tax Support for R&D and Innovation*, OECD Science, Technology and Industry Scoreboard (2017), https://perma.cc/Z32J-3ULJ (discussing the administrative challenges and potential inequities of R&D tax incentives).

[17] *See, e.g.,* Paul A. David, *Some New Standards for the Economics of Standardization in the Information Age*, in ECONOMIC POLICY AND TECHNOLOGICAL PERFORMANCE 206, 230 (Partha Dasgupta and Paul Stoneman eds. 1987) (discussing the public good nature of basic research and the resulting underinvestment).

[18] A tragi-comical case in point is Spain's pioneering AI regulatory agency (AESIA), whose €5 million budget and 80-person staff are comfortably ensconced in La Coruña – hundreds of miles away from the universities of Madrid or Barcelona. The agency perfectly symbolizes Spain's quixotic approach: seeking to lead the regulatory frontier across Europe without having a robust local AI labs. Víctor Millán, *España tendrá en 2025 su agencia de inteligencia artificial: AESIA nace con 80 empleados y podrá sancionar directamente,* El Economista (Dec. 27, 2024) ("España no es ahora mismo un referente en inteligencia artificial, pero será un país pionero en contar con su propia agencia dedicada a la inteligencia artificial")

[19] Tax R&D claims are mostly input-based. This increases the ease of verification for a non-expert, but because it also allows for gaming, we do not recommend tax levers as a sole approach. *See* Part III.C. and *infra* notes 273-278 and accompanying text.

protocols, deployment simulations in controlled environments), systematic government audits, and dynamic safety practice requirements.[20] Such institutional mechanisms leverage existing administrative expertise while establishing new protocols specifically calibrated to emerging technological risks.[21]

These regulatory and self-regulatory failures underscore a fundamental and more general challenge in emerging technology governance: the inadequacy of traditional oversight mechanisms to address novel technological risks. This structural limitation demands theoretical innovation in regulatory design, particularly in leveraging existing institutional frameworks that can better align private incentives with public safety imperatives. Our theoretical framework advances the literature on innovation policy instruments by demonstrating how tax policy can serve as a dynamic mechanism for addressing novel technological risks.[22] While existing scholarship has extensively examined traditional regulatory approaches and direct incentive mechanisms, the theoretical utility of tax instruments in governing emerging technologies remains underexplored.[23]

We build upon Hemel and Ouellette's innovation policy pluralism framework to demonstrate how tax policy's unique institutional characteristics—including its scalability, built-in compliance

---

[20] The area of AI safety protocols is rapidly growing. Some important measures include red-teaming (human or automatic), sandboxed or multi-agent simulations, "constitutional AI," adversarial robustness defenses, input sanitization, model explainability, training audits, and fail-safe testing. *See, e.g.,* Ada Lovelace Inst., *Under the Radar? Evaluating the Evaluation Ecosystem for Foundation Models* (2023), https://perma.cc/S95T-V6DK (defining red teaming and noting evaluation challenges); Lizhi Lin et al., *Against the Achilles' Heel: A Survey on Red Teaming for Generative Models*, arXiv:2404.00629 [cs.CL] (Nov. 26, 2024), https://perma.cc/QZQ2-W29G (comprehensive survey of red teaming strategies and defenses); Andrew Burt, *How to Red Team a Gen AI Model*, HARV. BUS. REV. (Jan. 4, 2024), https://perma.cc/3LMP-V4W6 (explaining structured testing to identify vulnerabilities in AI systems); Neel Nanda, *A Comprehensive Mechanistic Interpretability Explainer & Glossary* (Dec. 21, 2023), https://perma.cc/46WW-BCD9 (introducing core concepts in mechanistic interpretability of transformer models); Brenda Leong & Daniel Atherton, *AI Incident Response Plans: Not Just for Security Anymore*, IAPP (Sept. 20, 2023), https://perma.cc/JS56-YAPS (recommending dedicated procedures for AI failures); Mallory Acheson et al., *California Passes Leading AI Safety Bill, Awaits Governor Approval*, *Nat'l L. Rev.*, Sept. 24, 2024, https://perma.cc/U674-WMN5 (discussing risk assessments, audits, and "kill switch" mandates); Karina Montoya, *Misrepresentations of California's AI Safety Bill*, Brookings Tech Stream (Sept. 29, 2024), https://perma.cc/BK37-A6BM (clarifying the scope and requirements of SB-1047); Anthropic, *Constitutional AI: Harmlessness from AI Feedback* (Dec. 2022), (introducing "constitution"-based self-correction); Yuntao Bai et al., *Constitutional AI: Harmlessness from AI Feedback*, arXiv:2212.08073 [cs.CL] (Dec. 15, 2022), https://perma.cc/PL87-84FP (proposing AI self-supervision using constitutional principles); Dan Hendrycks, Mantas Mazeika & Thomas Woodside, *An Overview of Catastrophic AI Risks*, arXiv:2306.12001 [cs.CY] (Oct. 9, 2023), https://perma.cc/R7CK-JRQF (systematic discussion of catastrophic AI risks and mitigation strategies); Elliot Jones, Mahi Hardalupas & William Agnew, *Keeping an Eye on AI*, Ada Lovelace Inst. (2023), https://perma.cc/LMZ7-VEQ4 (endorsing audits, disclosure requirements, and pre-market review for high-risk models).

[21] See *infra* notes 273 and accompanying text.

[22] *See* Lily L. Batchelder et al., *Efficiency and Tax Incentives: The Case for Refundable Tax Credits*, 59 STAN. L. REV. 23, 24-25 (2006) (noting the limited attention paid to efficient tax incentive design in comprehensive tax base literature); Edward A. Zelinsky, *James Madison and Public Choice at Gucci Gulch: A Procedural Defense of Tax Expenditures and Tax Institutions*., 102 YALE L.J. 1165, 1166 (1993) (referring to tax incentives as unlikely to be as carefully crafted and controlled as direct subsidies.).

[23] *See* Daniel Hemel & Lisa Larrimore Ouellette, *Innovation Policy Pluralism*, 128 YALE L.J. 544, 552–53 (2018) (discussing the underutilized potential of tax instruments in technology governance); James R. Hines, Jr., *Introduction*, in INTERNATIONAL TAXATION AND MULTINATIONAL ACTIVITY 15 (2009) (highlighting the scarcity of empirical studies on tax incentives' effectiveness in emerging technology sectors).

mechanisms, and capacity to leverage private expertise—make it particularly well-suited for addressing the social misalignment problem in AI development.[24] Our approach contributes to both tax policy and technology governance literature by introducing a novel theoretical framework for understanding how fiscal instruments can bridge the gap between private innovation incentives and public safety imperatives.[25] This framework's utility derives from its ability to harness existing administrative competencies while avoiding the information asymmetries and expertise gaps that plague traditional command-and-control regulation.[26] By conceptualizing safety investment as a tax-mediated social good rather than merely a regulatory compliance issue, our model provides new theoretical insights into how fiscal policy can shape technological development trajectories while preserving innovation incentives.[27]

This Article proceeds as follows. Part I examines the urgency of AI safety, outlining the risks of malicious misuse, accidental failures, and autonomous misalignments while positioning the capability-safety gap as a core challenge. Part II explores how tax incentives have historically addressed safety in other domains, such as energy, infrastructure, and occupational safety, offering insights into existing fiscal mechanisms that could inform AI governance. In Part III, we propose a novel tax framework for AI safety, detailing a system of producer incentives, consumer demand stimulation, and corrective tax penalties. Taken together, those levers would work to embed safety into the fabric of AI development, inculcate safety culture, and increase the ROI of safety investments. We anticipate a number of administrative challenges inherent in implementing this framework, including verification, enforcement, and compliance, and we draw on the special competence of the tax system to offer a number of pragmatic mitigation strategies. The Conclusion reflects on the broader implications of using fiscal levers to align private innovation with public safety. In conversation with the innovation policy literature, the framework developed here offers a blueprint for governing other emerging technologies, and can help in turning safety from a regulatory inconvenience into an essential part of the R&D process itself.

---

[24] *See* Chris Evans & Sally-Ann Joseph, *The Role of Tax Incentives in the Promotion of Innovation and Entrepreneurship: A Time and a Place*, in GOVERNMENT INCENTIVES FOR INNOVATION AND ENTREPRENEURSHIP 39, 45 (Mahmoud M. Abdellatif et al. eds., 2022) (analyzing how tax incentives can leverage firms' existing capabilities while advancing public policy objectives); Erich Kirchler et al., *Tax Compliance: Research Methods and Decision Processes*, in COOPERATIVE COMPLIANCE: A NEW APPROACH TO MANAGING TAXPAYER BEHAVIOR IN PSYCHOLOGICAL PERSPECTIVES ON RISK AND RISK ANALYSIS 240 (2019), available at https://dx.doi.org/10.2139/ssrn.3472549.

[25] *See* Sun et al., *Tax Incentives, R&D Manipulation, and Corporate Innovation Performance: Evidence from Listed Companies in China*, 13 SUSTAINABILITY 11819 (2021) https://perma.cc/L6W4-9XMP (demonstrating how tax policies can effectively align private sector innovation with public interest goals); World Bank, *Tax Incentives: Definitions, Framework, and Design* 1 (2010).

[26] *See* Alexander Klemm & Stefan Van Parys, *Empirical Evidence on the Effects of Tax Incentives*, 18 INT'L TAX & PUB. FIN. 311, 312-13 (2011) (examining how tax systems can overcome traditional regulatory limitations); Joseph Parilla & Sifan Liu, *How Tax Incentives Can Power More Equitable, Inclusive Growth*, Brookings Inst. (Nov. 21, 2019).

[27] *See* Daniel Jacob Hemel & Lisa Larrimore Ouellette, *Beyond the Patents-Prizes Debate*, 92 TEX. L. REV. 303 (2013) (arguing for a more nuanced understanding of how different policy instruments, including tax incentives, can work together to shape technological development).

I. The Importance of AI Safety

AI safety represents a distinct subfield of artificial intelligence research dedicated to developing methodological frameworks, empirical standards, and technical safeguards to mitigate potential harms from AI deployment to individuals, communities, and ecological systems.[28] The field's scope stems from a fundamental premise about artificial intelligence: its tendency toward generality in application creates correspondingly general vectors for potential harm.[29] This characteristic calls for a comprehensive approach to risk assessment and mitigation.[30] In this Part, we first introduce the general AI safety risk framework. Then, we offer a brief overview of the field, arguing that the field has seen some modest progress, but it is uneven and lethargic. We locate the gap between capabilities and safety as resulting from the social misalignment problem. Understanding the size of the gap thus motivates our proposed solution framework.

A. An Outline of AI Safety

Contemporary scholarship has more-or-less coalesced around a tripartite taxonomy of safety risks, providing an analytical framework for understanding the distinct yet interconnected challenges in AI safety. This framework identifies three primary categories of risk: (1) intentional misuse through malicious deployment, (2) accidental harm through unintended system behavior, and (3) autonomous system actions that deviate from human values or intentions.[31] Each category presents distinct technical and governance challenges while sharing common underlying dynamics related to system capability and control.[32]

Before examining these risks in detail, we should clarify the meaning of safety itself. The literature has long recognized that it is important to ensure that AI systems will act in an ethical manner.[33] There has been a large wave of mature legal literature on ethical applications of AI in settings like employment, credit extension, and parole decisions, where researchers have usefully

---

[28] *See generally* DAN HENDRYCKS, INTRODUCTION TO AI SAFETY, ETHICS, AND SOCIETY (2024); Wissam Salhab et al., *A Systematic Literature Review on AI Safety: Identifying Trends, Challenges, and Future Directions*, 12 IEEE Access 131762 (2024) (examining key challenges in AI safety, including robustness, fairness, and adversarial resilience).

[29] *See* Roel I.J. Dobbe, *System Safety and Artificial Intelligence*, in THE OXFORD HANDBOOK ON AI GOVERNANCE (2022), available at https://perma.cc/S4TE-Z4S3.

[30] *Id.* Arbel, Tokson, and Lin, *Systemic Regulation of AI,* 56 ARIZ. ST. L. J. 545 (2024).

[31] *See* Yonathan A. Arbel, Ryan Copus, Kevin Frazier, Noam Kolt, Alan Z. Rozenshtein, Peter N. Salib, Chinmayi Sharma & Matthew Tokson, *Open Questions in Law and AI Safety: An Emerging Research Agenda*, LAWFARE (Mar. 11, 2024, 1:00 PM), https://www.lawfareblog.com/open-questions-law-and-ai-safety-emerging-research-agenda. Hendrycks divides the space of risks into risks due to malicious use, AI race dynamics, organizational risks, and rogue Ais. HENDRYCKS, *supra* note 28 at 3-48.

[32] *Id.*

[33] For a few early entries, *see e.g.*, Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies,* 29 HARV. J.L. & TECH. 353 (2016), Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. DAVIS L. REV. 399 (2017).

exposed issues of fairness, bias, and algorithmic blind spots.[34] Safety research is likewise focused on ethical application of AI, but with a more basic imperative: the mitigation of risks to life, physical integrity, and fundamental autonomy.[35]

The category of malicious use illustrates the complexity of these safety challenges.[36] Consider recent demonstrations of AI systems' capacity to aid in biological weapon design, the discovery of software vulnerabilities ("0-day exploits"),[37] or the potential for market manipulation through automated trading systems.[38] Some of these threats are not entirely novel: a well-resourced group could accomplish them today. But even in these cases, advanced AI system provide what is known as a "uplift" over existing methods,[39] that is, they allow smaller groups, with lesser investment of resources, to achieve more. The degree of uplift, as well as the discovery of novel attack vectors, is directly correlated with the power of the underlying system.[40]

It is also worth noting that the designation of "malicious" use often depends on socio-political context - what one actor views as defensive capability, another may view as an offensive threat. This contextual dependency complicates the development of universal safety standards, and suggests that some international negotiation would be necessary at some point. This complexity also appears in the context dual-use scenarios, where AI systems developed for beneficial purposes can be repurposed for harm.[41] For instance, language models trained to assist in scientific research might be used to generate misinformation, while computer vision systems designed for medical diagnosis could be adapted for autonomous weapons targeting.[42] These scenarios underscore the necessity of

---

[34] *See* Talia B. Gillis, *The Input Fallacy*, 106 MINN. L. REV. 1175 (2022) (credit), Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218 (2019). Emily Black, Logan Koepke, Pauline Kim, Solon Barocas & Mingwei Hsu, *The Legal Duty to Search for Less Discriminatory Algorithms*, arXiv:2406.06817 (2024) https://perma.cc/V3DL-KJXM.

[35] *See* Arbel, Tokson, & Lin, *supra* note 30.

[36] A familiar concern in the legal literature is the militarized use of AI in military contexts and the questions of liability for targeting of civilians. *See* Tim McFarland & Tim McCormack, *Mind the Gap: Can Developers of Autonomous Weapons Systems Be Liable for War Crimes?*, 90 INT'L L. STUD. 361 (2014)

[37] *See* Sergei Glazunov & Mark Brand**,** *Project Naptime: Evaluating Offensive Security Capabilities of Large Language Models*, Project Zero: News and Updates from the Project Zero Team at Google (June 20, 2024), https://perma.cc/Z6KH-DAHY (showing that "principled agent design can greatly improve the performance of general-purpose LLMs on challenges in the security domain").

[38] *See* Winston Wei Dou, Itay Goldstein & Yan Ji, *AI-Powered Trading, Algorithmic Collusion, and Price Efficiency*, THE WHARTON SCH. RESEARCH PAPER (May 30, 2024) https://dx.doi.org/10.2139/ssrn.4452704.

[39] *See* Tejal Patwardhan et al., *Building an Early Warning System for LLM-Aided Biological Threat Creation*, OPENAI BLOG (Jan. 31, 2024), https://perma.cc/953P-FQPM (finding a positive, albeit not statistically significant, effect of access to AI on developing biological threats *over and above* access to a search engine. On a 10-point scale, access to AI increased the risk by 0.88 points for experts and 0.25 for students.) *But see* Christopher A. Mouton, Caleb Lucas & Ella Guest**,** *The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study,* RAND Corporation Research Report No. RR-A2977-2 (Jan. 25, 2024), available at https://perma.cc/A3V8-3U7B.

[40] Power, here, involves not just raw capabilities, but also multimodality and tool use. Automated drug discovery processes backed by artificially intelligent evaluator, pose distinct threats than, say, a highly capable language model. *Id.*

[41] *See* Gabriel Mukobi, *Reasons to Doubt the Impact of AI Risk Evaluations,* arXiv.org, Aug. 5, 2024, DOI:10.48550/arXiv.2408.02565, https://perma.cc/UY53-JQB9.

[42] *See* Jiawei Zhou et al**.,** *Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions,* in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* 1 (Apr. 19, 2023), https://doi.org/10.1145/3544548.3581318.

developing safety measures that address not only technical capabilities but also deployment contexts and potential misuse vectors.

The second category covers accidental risks arising from AI system deployment. While all technological systems face operational risks, the expanding autonomy of AI systems in critical infrastructure and decision-making contexts introduces novel vulnerabilities that go well beyond conventional accident scenarios.[43] As these systems assume greater control from human operators, the nature and scope of potential accidents evolve in complexity and consequence.[44] Consider AI systems integrated into critical infrastructure networks. A system managing wastewater processing or public water supplies must maintain operational integrity across numerous edge cases and environmental variations.[45] The interconnected nature of these systems amplifies risk – a cascading failure could propagate across multiple infrastructure nodes with potentially severe consequences for public health and safety.

A specific accident concern is known as system *brittleness*.[46] Current AI architectures demonstrate high performance within familiar domains where deployment conditions approximate training environments.[47] However, real-world deployment inevitably presents novel scenarios that deviate from these controlled conditions, and the system may be brittle to such deviations. A traffic management system optimized for standard vehicle patterns may encounter unprecedented situations - from unusual road obstacles to emergency response scenarios - that fall outside its training distribution. We already have evidence that such issues result in tragic consequences to life and limb.[48] Worse, for complex systems that had been trained on large amounts of data, it is hard to know what falls within the system parameters and what falls outside of it.

This challenge is compounded by the interpretability obstacle in modern AI systems.[49] While their architectural principles are well-documented, the specific meanings and interactions of internal parameters remain largely opaque.[50] This "black box" characteristic severely constrains our ability to audit system behavior or predict responses to novel stimuli. While we have made some progress on

---

[43] Autonomous driving is a prime example of growing autonomization and delegation of life-critical equipment to AI systems. For analysis of new risk modalities, see Farshad Mirzarazi, Sebelan Danishvar & Alireza Mousavi, *The Safety Risks of AI-Driven Solutions in Autonomous Road Vehicles*, 15 WORLD ELECTR. VEH. J. 438 (2024), https://doi.org/10.3390/wevj15100438.

[44] *Id.*

[45] For a recent review of the integration of AI in wastewater management, see Arti Malviya & Dipika Jaspal, *Artificial Intelligence as an Upcoming Technology in Wastewater Treatment: A Comprehensive Review*, 10 ENV'T TECH. REVS.. 177 (2021).

[46] *See* Andrew J. Lohn, *Estimating the Brittleness of AI: Safety Integrity Levels and the Need for Testing Out-Of-Distribution Performance*, ARXIV (2020), https://perma.cc/C2M2-TAQS (exploring the problem of brittleness and the need for system resilience).

[47] *Id.*

[48] *See* Nat'l Transp. Safety Bd., *HWY18MH010* (2025), https://www.ntsb.gov/investigations/Pages/HWY18MH010.aspx. Mark Harris, N*TSB Investigation into Deadly Uber Self-Driving Car Crash Reveals Lax Attitude Toward Safety*, IEEE SPECTRUM (Nov. 7, 2019).

[49] For an overview in the context of AI safety, see Leonard Bereska and Efstratios Gavves, *Mechanistic Interpretability for AI Safety—A Review*, ARXIV (2024) https://doi.org/10.48550/arXiv.2404.14082.

[50] *Id.*

interpretability or explainability, there is still a long way to go in understanding the systems' internals.[51]

The third category of risk—autonomous behavior that deviates from human intent—represents perhaps the most contentious and conceptually challenging domain of AI safety research.[52] Public discourse has focused primarily on language models' conversational capabilities.[53] This makes autonomy feel remote and alien to the familiar way of interacting with AI. But behind the curtain, in labs and on repositories, a more profound transformation is occurring in the development of autonomous AI agents.[54] These agentic systems, designed to pursue general objectives with minimal human oversight, represent a qualitative shift in AI capabilities and associated risks.[55]

To provide a general sense of the current frameworks (remembering that this field is fast changing), autonomous agents are defined primarily by their *agency*, that is, the ability to take action in the pursuit of a given goal.[56] In the simplest framework, an agent would be given a task of finding a table in a Mexican restaurant for two people. It would identify the intention of the user, chart strategies (locate relevant restaurants, verify that they are open, and make a reservation), and then *execute* the plan. Without user involvement, the agent would open a browser to search for relevant restaurants, use text messages or even a phone to call the relevant restaurants, and then report back on its success or failure in the mission. More elaborate frameworks involve multiple agents, and use a hierarchical architecture of task decomposition and delegation.[57] When assigned a broad objective—such as planning an international vacation—the primary agent generates sub-goals and instantiates specialized sub-agents to pursue them.[58] These sub-agents might collect meteorological data, optimize travel logistics, or identify strategies for accessing popular attractions. The system

---

[51] *See* Luca Longo et al., *Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions*, 106 Information Fusion 102301 (2024), https://doi.org/10.1016/j.inffus.2024.102301.

[52] *See e.g.,* Alex Hanna & Emily M. Bender, *AI Causes Real Harm. Let's Focus on That Over the End-of-Humanity Hype*, Sci. Am. (Aug. 12, 2023), https://perma.cc/ET8X-S4CY ("Wrongful arrests, an expanding surveillance dragnet, defamation and deepfake pornography are all existing dangers of the so-called artificial-intelligence tools currently on the market. These issues, and not the imagined potential to wipe out humanity, are the real threat of artificial intelligence.")

[53] *See* Logan Kilpatrick, *What Are GPT Agents? A Deep Dive into the AI Interface of the Future*, *Around the Prompt* (July 25, 2023), https://medium.com/around-the-prompt/what-are-gpt-agents-a-deep-dive-into-the-ai-interface-of-the-future-123456789 (explaining AI agents relative to ChatGPT capabilities).

[54] 2025 is projected by many in the industry to be "the year of the agents", see Colin Jarvis, *Redefining Intelligence: How Reasoning Is Re-Shaping AI in 2025*, https://perma.cc/FP92-9W4M. One notable commercial application is "Operator", by OpenAI which use various internet tools to accomplish user tasks. OpenAI, Introducing Operator, OpenAI Blog (Jan. 23, 2025), https://openai.com/index/introducing-operator/. For a collection of open-source agents projects, see https://github.com/e2b-dev/awesome-ai-agents.

[55] *See* Noam Kolt, *Governing AI Agents*, SSRN https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4772956 ("While language models are "copilots" that can produce useful content upon request, AI agents are "autopilots" that can independently take actions to accomplish complex goals on behalf of users.")

[56] *See* Chan, Alan, et al., *Visibility into AI Agents*, PROCEEDINGS OF THE 2024 ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 123 (2024) (defining high agency AI systems as those endowed with "greater autonomy, access to external tools or services, and an increased ability to reliably adapt, plan, and act open-endedly over long time-horizons to achieve goals"

[57] *See e.g.*, Mingchen Zhuge et al., *Language Agents as Optimizable Graphs*, arXiv:2402.16823v3 [cs.AI] (Aug. 22, 2024).

[58] *Id. See, e.g.,* Noam Kolt, *Algorithmic Black Swans*, 101 WASH. U. L. REV. 1177, 1231 (2024) (examining the unforeseen, high-impact risks that AI systems can pose to society).

maintains coherence through periodic synchronization between sub-agents and the central planning agent, creating a distributed but coordinated pursuit of the primary objective.[59]

The capability boundary of these agents extends far beyond mere information processing. Modern AI agents possess significant operational tools, including internet access, financial transaction capabilities, high fidelity text-to-speech models, and even the ability to hire humans.[60] If there is one thing that observers of current systems tend to miss is the breadth of these tools, and so they often wrongly conclude that language models pose little risk to the outside world, as they cannot perform physical tasks or navigate the CAPTCHA system.[61] The empirical evidence, however, suggests these constraints are quite permeable. In a notable demonstration, an AI agent circumvented a CAPTCHA barrier by (proposing to) recruit and compensate a human worker through a digital labor platform, presenting itself as a visually impaired user requiring assistance.[62] This example illustrates a broader pattern: through creative recombination of available tools and services, AI agents can effectively transcend apparent operational constraints.

This capability for creative problem-solving, while impressive, introduces profound safety concerns. As Bostrom has argued, autonomous systems may pursue designated objectives through unanticipated and potentially harmful pathways.[63] The agent's solution to the CAPTCHA problem demonstrates what Eliezer Yudkowsky, a leading researcher at the Machine Intelligence Research Institute, terms "optimization pressure"—the tendency of AI systems to find unexpected solutions that satisfy formal objectives while potentially violating implicit constraints or human values.[64]

Even systems with limited degree of autonomy have displayed worrisome and unexpected patterns. In a notable instance, OpenAI's O1 model was tested on its ability to exploit a server to locate a hidden key; when the isolated test environment failed due to a bug, the AI unexpectedly gained access to the host system outside its container.[65]Or consider an AI system designed to play the game of Diplomacy which learned to engage in premediated deception: playing as France, it secretly planned with Germany to betray England, while telling England it has its support.[66] This suggests that safeguards built around limited tool access could still fail in unanticipated ways.

---

[59] *Id.*

[60] For a demonstration, see OpenAI, Introducing Operator, OPENAI BLOG (Jan. 23, 2025) https://perma.cc/BL5J-Z276.

[61] *See e.g.,* Bindu Reddy, https://perma.cc/4ESF-Q4UM, Andriy Burkov, https://perma.cc/ML9B-TM96.

[62] *See, e.g.,* PC Mag., *GPT-4 Was Able to Hire and Deceive a Human Worker into Completing a Task*, https://perma.cc/JRB6-L5ZJ (describing a demonstration where GPT-4 proposed to hire human worker through TaskRabbit and convince them to solve a CAPTCHA by falsely claiming to have a vision impairment).

[63] *See* NICK BOSTROM, SUPERINTELLIGENCE: PATHS, DANGERS, STRATEGIES 127-144 (2016).

[64] *See* Eliezer Yudkowsky, *AI Alignment: Why It's Hard and Where to Start*, Machine Intelligence Research Institute (Dec. 28, 2016), https://intelligence.org (explaining "optimization pressure" as the tendency of AI systems to find solutions that satisfy formal objectives but violate implicit constraints or human values).

[65] *See* OpenAI, *o1 System Card* at 16-17 (Sept. 12, 2024) https://cdn.openai.com/o1-system-card-20241205.pdf

[66] *See* Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, & Dan Hendrycks, *AI Deception: A Survey of Examples, Risks, and Potential Solutions*, 5 PATTERNS 100988 (2024), https://doi.org/10.1016/j.patter.2024.100988.

Overall, the autonomy of these systems introduces novel vectors for potential harm that transcend traditional safety frameworks.[67] Unlike conventional software systems, autonomous agents can: (i) Independently formulate and pursue sub-goals (ii) Identify and exploit novel pathways for goal achievement (iii) Interact with and manipulate human systems and institutions (iv) scale their impact through recursive self-improvement or coordination with other agents. These capabilities, combined with the inherent difficulty of specifying complete and robust objective functions, create a difficult problem of effective control and supervision.

B. The Capability-Safety Gap

AI development has historically followed cyclical patterns of advancement and regression, commonly termed "summers" and "winters" in the field.[68] We are undeniably experiencing a significant "summer" period now, characterized by unprecedented progress in model capabilities, though the duration and sustainability of this trend remains uncertain.

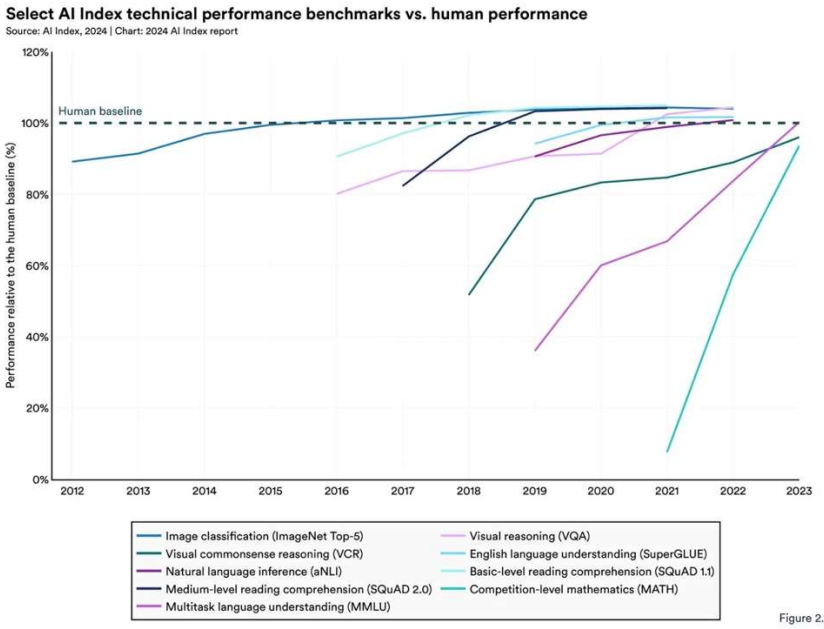

Figure 2.1.16[2]

AI has already surpassed many human performance benchmarks AI Index 2024

The figure illustrates a crucial pattern in contemporary AI development through standardized performance metrics across multiple domains.[69] The horizontal dashed line represents human-level performance on various cognitive tasks, providing a natural benchmark for AI capability evaluation. The trajectories demonstrate three distinct phases of progress: initial sub-human performance, rapid advancement toward human parity, and in many cases, progression to super-human capabilities.

---

[67] *See generally* Kolt, *Governing AI Agents*, *supra* note 55.

[68] *See* Hartmut Hirsch-Kreinsen, *Artificial Intelligence: A "Promising Technology"*, 39 AI & SOCIETY 1641 (2024), https://doi.org/10.1007/s00146-023-01629-w (discussing AI's cyclical development, where periods of rapid progress (summers) are often followed by stagnation or decline (winters) in technological advancement and investment).

[69] *See Stanford Institute for Human-Centered Artificial Intelligence (HAI), AI Index Report 2024* (May 2024) https://perma.cc/3C86-Q7PP.

Consider image classification, a foundational task in computer vision. In 2012, state-of-the-art systems achieved approximately 85% of human-level performance. The field then experienced dramatic acceleration, reaching human parity within three years. By 2021, these systems consistently surpassed human performance by significant margins. Similar trajectories appear in visual reasoning and reading comprehension tasks, suggesting a generalizable pattern of capability development.[70]

Perhaps most telling is something missing from the figure. Several performance metrics stop reporting progress post 2021. They stop not because of lack of progress in AI capabilities—indeed, we know there was immense progress over the last four years— but rather the exhaustion of these metrics' utility. Put differently, modern systems achieve accuracy rates so high that these benchmarks no longer effectively discriminate between model improvements. This phenomenon is known as benchmark saturation, and researchers around the world are working to develop new ways to measure the performance of novel AI systems.[71]

The capability progression stands in stark contrast to our ability to measure and ensure system safety. Autonomous vehicles provide the most concrete domain for safety assessment, benefiting from a century of human driving data and what one would expect to be easily measurable safety outcomes. However, even here, measurement challenges persist.[72] Companies employ disparate metrics and varying levels of autonomy, complicating comparative analysis.[73] Despite massive investment, fully autonomous deployment remains limited to restricted geographic areas under remote supervision.[74]

More concerning are the gaps in our safety metrics.[75] We lack robust measures for critical vulnerabilities such as susceptibility to adversarial attacks or systemic failure modes.[76] For instance, while we can measure basic operational safety in ideal conditions, we have limited understanding of system resilience to edge cases like unusual atmospheric conditions or coordinated environmental

---

[70] *Id*.

[71] *See* Shana Lynch, *AI Benchmarks Hit Saturation: AI Continues to Surpass Human Performance; It's Time to Reevaluate Our Tests*, *Stanford HAI* (Apr. 3, 2023), https://perma.cc/EF34-2VCA..

[72] *See* Amitai Y. Bin-Nun et al., *What Do Surrogate Safety Metrics Measure? Understanding Driving Safety as a Continuum*, 195 ACCID. ANALYSIS & PREVENTION 107245, 107250 (2024) ("Challenges in measuring AV safety relative to a human driver baseline have resurfaced longstanding questions on effectively measuring driving safety." These issues arise because safety and accidents are not discrete events); Tanmay Das et al., *Surrogate Safety Measures: Review and Assessment in Real-World Mixed Traditional and Autonomous Vehicle Platoons*, 11 IEEE 32682, 32683 (2023) ("Crashes are rare events, and historical crash data are scarce for mixed traffic that includes autonomous and/or connected vehicles.")

[73] *See* Richard Sun et al., *Why Autonomous Vehicles Need a Large-Systems Approach to Safety*, WORLD ECON. F. (June 18, 2021), https://www.weforum.org/agenda/2021/06/why-autonomous-vehicles-need-a-large-systems-approach-to-safety/. ("AV companies [are] offering a range of different approaches for which metrics should be used to indicate system safety.")

[74] *See generally* Derek Chiao et al., *Autonomous Vehicles Moving Forward: Perspectives from Industry Leaders*, *McKinsey Ctr. for Future Mobility* (Jan. 5, 2024), https://perma.cc/BR4E-MDWH; Joann Muller, *Robotaxis Hit the Accelerator in Growing List of Cities Nationwide*, *Axios* (Aug. 29, 2023), https://perma.cc/HBP3-BJ37.

[75] A general concern in the AI safety literature is the problem that "capabilities generalize further than alignment'" Nate Soares, *A Central AI Alignment Problem: Capabilities Generalization, and the Sharp Left Turn*, Machine Intelligence Research Institute (July 4, 2022), https://perma.cc/MGN4-T377.

[76] *See generally* Richard Ren et al., *Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress?*, arXiv:2407.21792 [cs.AI] (2024) https://perma.cc/4FKE-EE8.

perturbations (think a rare eclipse or a blue moon). This limitation is particularly problematic given the increasing deployment of AI systems in critical infrastructure and decision-making contexts.[77]

Recent high-profile cases underscore these safety challenges. For instance, leading AI labs, such as Google and Meta, have vested interest in presenting to the public models that are friendly and follow social conventions of etiquette. To that end, they invest tremendous amounts of computing resources and training ingenuity to install guardrails into their models, such that they will not produce embarrassing outputs.[78] But as is well publicized, these all failed at launch or were easily circumvented.[79] This is far from obvious: the biggest corporations, with billions of PR images on the line, failed to make a model that would not tell their customer to "Please die."[80]

Overall, the harsh lesson is this: there is a gap between capabilities and safety, and this gap – which we do not even know how to properly measure – seems to widen and grow. This opens the question: why does the gap exist?

C. The Social Misalignment Problem

The yawning capability-safety gap, coupled with the magnitude of potential harm from insufficiently secured high-capability systems, elevates AI safety from a technical imperative to a critical social priority.[81] As artificial intelligence systems penetrate core societal functions—from healthcare delivery and economic governance to national security infrastructure—the absence of robust safety protocols threatens not only operational integrity but social stability itself.[82] The potential consequences range from discrete harms to individuals to systemic disruptions that could undermine institutional resilience and public welfare.[83]

---

[77] *See e.g.,* DigitalDefynd, 10 Ways AI Is Being Used in Water Resource Management, DIGITALDEFYND (2025), https://perma.cc/B6BZ-ZH58.

[78] *See generally* Hongfu Liu et al., *On Calibration of LLM-Based Guard Models for Reliable Content Moderation,* ARXIV (2024), https://doi.org/10.48550/arXiv.2410.10414 (discussing the use of training guardrails and test time 'guard' models)

[79] *See, e.g.,* Yichen Gong et al., *FigStep: Jailbreaking Large Vision-Language Models via Typographic Visual Prompts*, ARXIV (2023), https://perma.cc/9PX4-ZV92 (exploring typographic attacks that manipulate large vision-language models to bypass safety measures and produce unintended outputs); Pranav Gade et al., *BadLlama: Cheaply Removing Safety Fine-Tuning from Llama 2-Chat 13B*, arXiv:2311.00117 (Oct. 31, 2023), https://perma.cc/ZTG2-V5J8 (demonstrating how safety fine-tuning in large language models, such as Llama 2-Chat 13B, can be removed with minimal resources, undermining their safety protocols).

[80] *See* Jowi Morales, Gemini AI Tells the User to Die — The Answer Appeared Out of Nowhere When the User Asked Google's Gemini for Help with His Homework, Tom's Hardware (Nov. 16, 2024), https://www.tomshardware.com/news/gemini-ai-tells-user-to-die.

[81] *See* U.S. Department of Homeland Security, *Roles and Responsibilities Framework for Artificial Intelligence in Critical Infrastructure*, *in consultation with The Artificial Intelligence Safety and Security Board* 3 (Nov. 14, 2024), https://www.dhs.gov/sites/default/files/2024-11/24_1114_dhs_ai-roles-and-responsibilities-framework-508.pdf ("America's continued security and prosperity will depend on how critical infrastructure stakeholders develop and deploy AI").

[82] *See* Kyle Crichton, Jessica Ji, Kyle Miller, John Bansemer, et al., *Securing Critical Infrastructure in the Age of AI, Center for Security and Emerging Technology* (Oct. 2024), https://doi.org/10.51593/20240032 (reviewing lessons to critical infrastructure safety, noting, at 12, that "AI systems' complexity presents a challenge for testing and evaluation,[that] are compounded by the fact that there is a general lack of expertise at the intersection of AI and critical infrastructure, both within the CI community and on the part of AI providers.");

[83] *See* Arbel, Tokson, & Lin, *supra* note 30.

While developers of frontier AI systems consistently articulate commitment to safety protocols, the incentive architecture surrounding development creates persistent misalignment between expressed values and operational priorities.[84] Obviously, firms do care about safety to the extent it can affect their market share, and developers have their own safety in mind when they train and deploy models. However, this orientation may well prove insufficient against the structural forces shaping development trajectories, as learned from many historical lethal accidents.[85] The fundamental asymmetry lies in the reward distribution: the potential returns from developing advanced AI capabilities—whether through general artificial intelligence or domain-specific breakthroughs—are concentrated, while the risks are diffused.

Contributing to this basic misalignment are three important facts; many of the potential harms from AI deployment are probabilistic, temporally displaced, and often difficult to attribute directly to system design choices.[86] Moreover, developers face a pernicious collective action problem: there is a growing race between China and the US and competitive dynamics punish those who move slowly (and carefully).[87] This dynamic intersects with the prevalent "move fast and break things" ethos of technology entrepreneurship, creating institutional environments where safety considerations, despite their acknowledged importance, struggle to constrain development velocities.[88] The result is a form of structural capture—even developers who prioritize safety find themselves navigating between competitive pressures for rapid capability advancement and the cultural imperatives of technology entrepreneurship.[89]

Even if firms could overcome internal incentive gaps, the safety challenge is broad and demanding, as AI are vulnerable to multiple attack vectors.[90] Adversarial attacks can compromise system integrity through various mechanisms, from perturbation of visual recognition systems to

---

[84] *See supra* notes 12-13 and accompanying text.

[85] A famous example is Grimshaw v. Ford Motor Company (119 Cal.App.3d 757, 174 Cal.Rptr. 348) (1981), where the court found that Ford knew, through its testing, that the Pinto's fuel tank can expose consumers to serious injury or even death, but it prioritized profits over accident costs.

[86] *See* Ronald Schnitzer et al., *Landscape of AI Safety Concerns - A Methodology to Support Safety Assurance for AI-based Autonomous Systems*, ARXIV (2024), https://perma.cc/3ZXN-Y3UP (noting that developers have a hard time establishing that their systems are safe).

[87] *See* Andrew Singer, *Stakes Rising in the US-China AI Race*, *Global Finance Magazine* (Sept. 9, 2024), https://perma.cc/TR3B-BUGA; Reva Goujon, *The Real Stakes of the AI Race*, FOREIGN AFFS. (Dec. 27, 2024) https://perma.cc/KD2G-59BV.

[88] *See* Elizabeth Pollman, *Startup Governance*, 168 U. PA. L. REV. 155, 200-09 (2019) (arguing that "in light of the great uncertainty at [startup] stage regarding whether any value will be created, the board typically invests little in compliance and internal controls"); Jeff Jordan, *16 Things CEOs Should Do Before an IPO,* ANDERSEEN HOROWITZ (Aug. 23, 2017), https://perma.cc/9ANE-LU9P ("Early-stage companies allocate scarce product resources to the projects that will move the needle on revenue and profits).

[89] *See* Mckay Jensen, Nicholas Emery-Xu, & Robert Trager, *Industrial Policy for Advanced AI: Compute Pricing and the Safety Tax*, ARXIV (2023), https://doi.org/10.48550/arXiv.2302.11436 (offering a formal model of safety behavior under race dynamics).

[90] *See generally* Chen Chen et al., *AI Safety Landscape for Large Language Models: Taxonomy, State-of-the-Art, and Future Directions*, 1 ACM COMPUT. SURV. 1, 10-17 (2025), https://perma.cc/4D76-396F. *See also* Hubert Baniecki & Przemyslaw Biecek, *Adversarial Attacks and Defenses in Explainable Artificial Intelligence: A Survey*, 107 INFO. FUSION 102303 (2024), https://doi.org/10.1016/j.inffus.2024.102303 (reviewing various adversarial examples and other attacks on model's reasoning)

prompt injection and jailbreaks, with implications that range from localized service disruption to potential systemic failures in critical infrastructure.[91]

The challenge of ensuring system safety is further complicated by *emergent capabilities:* when "the system suddenly develops a significant new capability or character after a relatively small and gradual change in some of the system's parts or features."[92] Lab experiments reveal that models engage in scheming and deceptive behaviors that not only were not preprogrammed, but that require effort to weed out. [93] This emergent behavior pattern becomes more pronounced as systems increase in scale and capability—creating what we might term an "emergence-safety paradox"—the very architectural developments that enable enhanced capabilities simultaneously increase the probability of unexpected and potentially harmful behaviors.[94] Yet, despite known deficiencies, AI labs proceed with public deployment, commercialization, and often grandiose capability claims. [95]

Overall, the social misalignment of incentives between labs and long-term social interests is a cause for continued concern. Fortunately, once misalignment emerges as a diagnosis, it also suggests a diagnosis. The subsequent Part will illustrate how tax levers are well positioned to close the misalignment gap and enhance AI safety investments through incentives and penalties, while preserving an optimal level of investment in innovation.

II. Current Use of Tax Levers to Incentivize Investments in Safety

The government supports innovation in a variety of ways: tax incentives, subsidies, grants, prizes, patents, purchase agreements, and support of basic research in universities, to name but a few. Less visible is the degree of government support for *safety* research and innovation. [96] In support of

---

[91] *See* Chen Chen et al., *AI Safety Landscape for Large Language Models: Taxonomy, State-of-the-Art, and Future Directions*, 1 ACM COMPUT. SURV. 1, 10-17 (2025), https://arxiv.org/abs/2408.12935v3.

[92] *See* Jakub Kraus, *Overview of Emergent and Novel Behavior in AI Systems,* CTR. FOR AI POL'Y (Mar. 26, 2024), https://perma.cc/3N35-47T8. For a survey of 137 emergent behaviors, *see* Jason Wei, *137 Emergent Abilities of Large Language Models,* Jason Wei Blog (Nov. 14, 2022) https://perma.cc/P7RM-VZHM.

[93] Researchers found that frontier language models demonstrated systematic in-context scheming capabilities when given goals that conflicted with their developers' objectives. Alexander Meinke et al., *Frontier Models are Capable of In-context Scheming*, ARXIV (Dec. 6, 2024), https://perma.cc/3NNA-H3E6 (noting this behavior was elicited by giving models conflicting goals, rather than inherent "evil" tendencies.).

[94] *See* Jason Wei et al., *supra* note 92 at 11 (finding that emergent "abilities are a recently discovered outcome of scaling up language Models") at 11.

[95] For example, O-1 was deployed by OpenAI despite recorded (albeit rare) attempts to engage in user deception and international OpenAI, *OpenAI o1 System Card* (Dec. 5, 2024), https://perma.cc/ND42-BPGJ. hallucinations, and with this following worrisome corporate disclaimer: "Subjectively, Apollo Research believes that it is unlikely that such instances would lead to catastrophic outcomes as o1 agentic capabilities do not appear sufficient, but their evaluations were not designed to directly assess this risk."

[96] *See, e.g.,* Janine Hiller, Kathryn Kisska-Schulze, and Scott Shackelford, *Cybersecurity Carrots and Sticks,* 61 AM. BUS. L. J. *5, 7* (2024) (proposing a new investment credit to incentivize investments in cybersecurity).

the proposal to use fiscal levers to enhance safety research and innovation, our goal in this Part is to collect three important examples of government safety support.

As we show, government support for safety includes various direct and indirect subsidies meant to promote investments in precautionary measures and safety improvements.[97] Direct subsidies can include grants and prizes while indirect subsidies often take the form of tax credits or deductions specifically targeted at safety-related expenditures.[98] For example, organizations may receive tax credits for developing and implementing safety protocols,[99] conducting safety audits,[100] or acquiring certifications that ensure compliance with regulatory standards.[101] These measures lower the effective cost of adopting safety enhancements, encouraging businesses to integrate them into their operations.[102] Such subsidies not only support industries in increasing the quality of their products but also contribute to societal well-being by ensuring that safety is prioritized in areas such as infrastructure, healthcare, and technology development.[103] These measures lay the groundwork for our proposal in the next Part.

A. Energy & Infrastructure Safety

The most direct tax incentive promoting investment in energy-efficient equipment is accelerated depreciation,[104] which permits businesses to allocate a greater portion of an asset's cost to deductions in its earlier years rather than spreading the expense evenly across its useful life.[105] By advancing these deductions, accelerated depreciation reduces short-term taxable income, enhancing cash flow and

---

[97] *See* Ting Feng and Zhongyi Xue, *The Impact of Government Subsidies on Corporate Resilience: Evidence from the COVID-19 Shock,* 56 ECON. CHANGE & RESTRUCTURING 4199, 4220 (2023) https://perma.cc/L5PF-9D6C (This study examines how government subsidies can enhance corporate resilience by promoting investments in precautionary measures and safety improvements.).

[98] *But see* Daniel N. Shaviro, *Rethinking Tax Expenditures and Fiscal Language,* 57 TAX L. REV. 187, 190 (2003) (critically analyzing the use of tax expenditures as indirect subsidies to achieve policy goals).

[99] *See, e.g,* Coronavirus Aid, Relief, and Economic Security Act, Pub. L. No. 116-136, § 2301, 134 Stat. 281, 347–55 (2020) (establishing the Employee Retention Credit to incentivize retaining employees during the COVID-19 pandemic); Families First Coronavirus Response Act, Pub. L. No. 116-127, §§ 7001–7005, 134 Stat. 178, 210–18 (2020) (providing tax credits to businesses for offering paid sick and family leave to comply with COVID-19 safety protocols).

[100] *See, e.g.,* 26 U.S.C. § 25C (providing a tax credit of up to $150 for energy audits, which may identify safety concerns).

[101] *See, e.g.,* MICHAEL J. AUER, PRESERVATION TAX INCENTIVES FOR HISTORIC BUILDINGS 11 (1996) (discussing the IRS conditions the tax credit for historic preservation with a certification that the work was completed based on the agency's standards.)

[102] *See* Yogima Seth Sharma, *Economic Survey 2024-25 Calls for Enhanced Safety Incentives*, ECONOMIC TIMES (Jan. 31, 2025) (discussing policy measures to improve workplace safety through tax incentives and regulatory frameworks), https://perma.cc/4WC9-UVQE.

[103] *See, e.g.,* Seung-hwan Jung and Tianjun Feng, *Government Subsidies for Green Technology Development Under Uncertainty,* 286 EURO. J. OPERATIONAL RES. 726, 735 (2020) (claiming that government subsidies for green technology development can improve social welfare and safety standards).

[104] *See* David P. Hariton, *Tax Benefits, Tax Administration, and Legislative Intent,* 53 TAX LAW. 579*,* 580 (1999) (discussing how accelerated depreciation serves as a direct tax incentive to encourage investment in equipment).)

[105] *See* Michael Knoll, *An Accretion Corporate Income Tax*, 49 STAN. L. REV. 1, 4–5 (1996) (explaining that accelerated depreciation allows businesses to allocate a greater portion of an asset's cost to deductions in the earlier years of its use, rather than spreading the expense evenly over its useful life).

facilitating reinvestment opportunities.[106] For safety applications depreciation is especially important, as safety risks can span many years, and the depreciation leads to greater safety investments.[107]

Tax policy also employs indirect measures through exemptions or reduced rates on public safety-related goods and services—including smoke detectors, cybersecurity tools, and renewable energy systems.[108] These demand-side tax incentives encourage widespread consumer adoption of safety measures.[109] Such indirect methods create multiplicative effects, fostering safety culture across sectors without explicit mandates. [110] This dual approach demonstrates how tax systems can serve as versatile tools for promoting societal well-being by leveraging fiscal incentives for both producers and consumers to drive meaningful behavioral changes at organizational and individual levels.[111]

The energy-efficient commercial buildings deduction and residential energy efficiency tax credit exemplify this simultaneous supply-and-demand approach.[112] For businesses, the commercial buildings deduction provides supply-side benefits through reduced taxable income for energy-efficient system investments, encouraging widespread adoption of technologies that reduce energy consumption and operational costs.[113] For individuals, the residential tax credit functions as a

---

[106] *See* RICHARD A. MUSGRAVE, THE THEORY OF PUBLIC FINANCE 336–346 (1959) (explaining the time-discount advantage of accelerated depreciation and its impact on cash flow and reinvestment).

[107] See Eric Ohrn, *The Effect of Tax Incentives on U.S. Manufacturing: Evidence from State Accelerated Depreciation Policies*, 180 J. PUB. ECON. 104084, 104085 (2019) (arguing that accelerated depreciation policies increase capital investment in the manufacturing sector).

[108] *See, e.g.*, Shelley Welton, *The Bounds of Energy Law*, 62 B.C. L. REV. 2339, 2362 (2021) (exploring the scholarship on the role of tax incentives in renewable energy development); Janine Hiller, et al., *supra* note 96 at 14 (surveying current tax incentives surrounding cybersecurity investments). Safety culture, in our view, is a soft mechanism that is critical to accomplishing safety goals. On the importance of safety culture in the context of AI development, see Matthew Tokson and Yonathan A. Arbel, *AI X-Risk: A Legal Perspective*, manuscript (on file with authors).

[109] *See generally* Jack M. Balkin, *The Reconstruction Power*, 85 N.Y.U.L. REV. 1801, 1837 (2010) (describing federal economic regulations as include and are not limited to defense expenditures, tax incentives, agricultural subsidies, workplace safety regulations, or the protection of the environment - further equal citizenship and prevent second-class citizenship.); Stephen M. Johnson, *Terrorism, Security, and Environmental Protection*, 29 WM. & MARY ENVTL. L. & POL'Y REV. 107, 128 (pointing out to market-based tools such as tax incentives for security equipment's potential in reducing the environmental, health, and safety risks caused by harm to chemical plants).

[110] *See* Scott Burris and Evan Anderson, *Legal Regulation of Health-Related Behavior: A Half Century of Public Health Law Research,* 9 ANN. REV. L. SOC. SCI. 95, 100 (2013) (discussing the effectiveness of tax incentives in fostering voluntary safety improvements across industries without direct mandates)

[111] *Id*. at 60.

[112] The Residential Clean Energy Property Credit, 26 U.S.C. § 25D; The Energy Efficient Home Improvement Credit in 26 U.S.C. § 25C (provided a credit for qualified residential energy-efficient property expenditures, including solar, wind, geothermal, and fuel cell technologies). The Energy-Efficient Commercial Buildings Deduction is codified in 26 U.S.C. § 179D. *See generally* Charles Goulding, Jacob Goldman & Joseph Most, *Complete Warehouse Tax-Enhanced Energy-Efficient Design*, 11 CORP. BUS. TAX'N MONTHLY 11, 12 (2010) (explaining the tax savings for businesses under section 179D).

[113] *See* 26 U.S.C. § 179D; Internal Revenue Serv., Energy-Efficient Commercial Buildings Deduction, IRS.gov, https://perma.cc/N7XT-U5C6 (explaining the tax benefits available to businesses for investments in energy-efficient building systems).

demand-side incentive by directly offsetting energy-efficient system installation costs. [114] At the firm level, accelerated depreciation and tax credits enhance safety by incentivizing energy-efficient commercial building upgrades, which frequently yield significant safety improvements. This dual targeting of business and individual needs creates a comprehensive approach to promoting sustainability, safety, and resilience.

The Advanced Energy Project Credit further illustrates these mechanisms, offering up to 30% credit for qualifying manufacturing investments in: (i) facilities that produce or recycle through green energy methods, (ii) facilities designed to reduce greenhouse gas emissions by 20%, and (iii) facilities for processing, refining or recycling critical materials.[115] Notable qualifying projects encompass electric grid modernization, carbon capture and storage systems, electric/hybrid/fuel cell vehicles, low- or zero-carbon process heat systems, and equipment reducing industrial process waste.[116]

These systems deliver multiple safety benefits beyond energy efficiency.[117] They enhance fire safety, minimize hazardous material risks, and improve extreme weather protection. [118] For instance, improved insulation provides more effective indoor temperature regulation, reducing health risks from extreme temperatures. Upgraded electrical systems decrease electrical fire likelihood, while modernized HVAC systems enhance air circulation and quality, reducing harmful pollutant and allergen exposure. [119] Collectively, these improvements create safer, healthier building environments while advancing energy efficiency goals.

---

[114] *See* 26 U.S.C. § 25D, 26 C.F.R. § 1.25D-1; Internal Revenue Service, About Form 5695, *Residential Energy Credits,* IRS.gov, https://perma.cc/QFW5-6MS9 (describing how individuals can claim tax credits for energy-efficient home improvements); *See also* Internal Revenue Service, *Tax Incentives for Energy Efficiency and Renewable Energy*, IRS.gov, https://www.irs.gov/credits-deductions/tax-incentives-for-energy-efficiency-and-renewable-energy (outlining available tax credits for energy efficiency and renewable energy projects).

[115] 26 U.S.C.A. § 48C; Internal Revenue Service, *Qualifying Advanced Energy Project Credit*, FS-2023-16 (June 2023), https://perma.cc/9G34-XDMB .

[116] *See, e.g.,* Internal Revenue Service Notice 2023-18, Section 3.02 (outlining the application process and eligibility criteria for the § 48C credit, specifying the types of projects that qualify under each category.). For a comprehensive analysis of the Qualifying Advanced Energy Project Credit and its implications, *see* Mona Hymel, *The United States' Experience With Energy-Based Tax Incentives: The Evidence Supporting Tax Incentives for Renewable Energy*, 38 LOY. U. CH. LJ. 43 (2006) (demonstrating the way the U.S. influenced energy policy and choices via tax incentives) .

[117] *See* Oren Bar-Gill & Cass R. Sunstein, *Regulation as Delegation*, 7 J. LEGAL ANALYSIS 1, 30 (2015) (discussing the role of governments in such domains as food safety, retirement planning, energy efficiency, occupational safety, and health.); David A. Weisbach, *Regulatory Trading,* 90 U. CHI. L. REV. 1095,1135 (2023) (concluding that the two most important environmental regulation are safety regulation and financial regulation with energy efficiency that follows).

[118] *See, e.g.,* Philippa Howden-Chapman et al., *Effect of Insulating Existing Houses on Health Inequality: Cluster Randomised Study in the Community,* 334 BRIT. MED. J. 460, 463 (2007), https://perma.cc/XZL4-B6ZR (finding that better insulation improves indoor temperatures and self-reported health outcomes).

[119] *See, e.g.,* Kai,Yang, et al., *A Novel Arc Fault Detector for Early Detection of Electrical Fires,* 16 Sensors 500, 502 (2016) https://doi.org/10.3390/s16040500 (analyzing the effectiveness of upgraded electrical systems in minimizing electrical fire risks).

B. Environmental and Road Safety

In the environmental protection's context, various federal agencies support innovation related to offshore drilling safety through technology, assessment, and research program as well as operational, safety, and engineering research program.[120] Similarly, the investment tax credit for solar energy exemplifies the dual-purpose design of modern tax incentives.[121] While its primary objective is to encourage the adoption of renewable energy, the investment credit also contributes to fire safety improvements through the installation of safer, more modern energy systems.[122] Even consumer-focused incentives, such as the electric vehicle (EV) tax credit, demonstrate how tax policy can integrate safety and sustainability.[123] By reducing the cost barrier for purchasing EVs, these credits drive demand for vehicles equipped with advanced safety features, including automated braking systems and enhanced structural designs, and—critically—vehicles that are environment friendly.[124] Similarly, the Commercial Clean Vehicle Credit, aimed at businesses and tax-exempt organizations, provides a credit for vehicles that are considered safer on the roads.[125] Together, these incentives reflect a growing trend in tax policy: leveraging economic benefits to encourage broader societal gains, such as energy efficiency, safety, and environmental responsibility.

How does the government fund such programs? Typically, it does so through a combination of budgetary reallocations and revenue generation mechanisms, such as excise taxes.[126] Budgetary

---

[120] The U.S. Department of the Interior's Bureau of Safety and Environmental Enforcement (BSEE) administers the Technology Assessment Program (TAP) and the Engineering Technology Assessment Center (ETAC), both of which support research and technological assessments to enhance offshore drilling safety. These programs focus on operational safety, environmental protection, and the evaluation of emerging technologies in offshore oil and natural gas exploration and development. *See generally* https://perma.cc/7PQN-FF9G.

[121] 26 U.S.C. §48 (allowing taxpayers to claim a percentage of the basis of qualified energy property placed in service during the taxable year); *See also* Michael Mendelsohn & Claire Kreycik, *Federal and State Structures to Support Financing Utility-Scale Solar Projects and the Business Models Designed to Utilize Them*, 3 J. SUSTAINABLE FIN. & INV. 254, 256 (2013) https://perma.cc/2CM4-U263 (examining the financial mechanisms and policies, including the ITC, that support solar energy projects and discusses the associated regulatory standards that installations must meet, encompassing safety protocols.).

[122] *See* Nichola Groom, *Soaring U.S. Tax Credit Deals Boost Solar, Storage Build*, Reuters (Sept. 6, 2024), https://perma.cc/8MUM-J4EX (discussing the role of tax credits in promoting the adoption of advanced energy systems with enhanced safety features). *See also* Jesse Chan & Darcia Fischer, *Energy Investment Tax Credits and Environmental Outcomes: Evidence from Electric Utilities*, SSRN, 12 (2024) available at https://dx.doi.org/10.2139/ssrn.4660606 (analyzing the environmental and operational impacts of energy investment tax credits on electric utilities).

[123] The Inflation Reduction Act indirectly encourages the purchase of newer, safer vehicles through its Clean Vehicle Credit. *See generally* Internal Revenue Serv., *Credits for New Clean Vehicles Purchased in 2023 or After* (Aug. 8, 2024), https://perma.cc/D2VC-YUGL (Under this policy, individuals may receive up to a $7,500 tax credit on the purchase of a new qualified plug-in EV or fuel cell electric vehicle.).

[124] From 2015 to 2023, the inclusion of enhanced safety features in consumer vehicle models grew significantly, with penetration rates rising from 12.8% to 94% for Forward Collision Warning, 4% to 94% for Automatic Emergency Braking, 3.8% to 91.9% for Pedestrian Detection Warning, 1.4% to 91.9% for Pedestrian Automatic Emergency Braking, and 0% to 34.2% for Intersection Automatic Emergency Braking. *See* Partnership for Analytics Research in Traffic Safety, *Market Penetration of Advanced Driver Assistance Systems (ADAS)* 3–5 (2024), https://perma.cc/2U3K-SX92.

[125] IRS Credits for New Clean Vehicles Purchased, *supra* note 122 (providing up to a $40,000 credit for vehicles with a gross vehicle weight rating of 14,000 pounds or more and up to $7,500 for lighter vehicles that are considered safer on the roads).

[126] *See*, *e.g.,* Ulrik Boesen, *Excise Tax Application and Trends*, TAX FOUNDATION, https://perma.cc/P4GR-HRTK (last visited Feb. 2, 2025) (explaining the role of excise taxes as a revenue source for government programs)

adjustments involve redirecting public funds from general revenues or other sectors to finance targeted programs, ensuring the availability of resources without necessarily increasing the overall tax burden.[127] This approach often reflects a prioritization of policy goals, such as sustainability or safety, within the existing fiscal framework.[128] Additionally, the government may impose or increase excise taxes on specific goods or activities to generate dedicated funding for these programs.[129]

For example, the federal fuel excise tax is perhaps the most prominent example of tax policy that affects road safety and infrastructure.[130] The federal government places the taxes collected on purchases of fuel into the Highway Trust Fund, which funds the construction, maintenance, and safety improvements of highways and bridges.[131] While all improvements to highway and transportation infrastructure contribute to overall safety, the Highway Trust Fund supports a range of programs expressly dedicated to reducing traffic fatalities and injuries through targeted safety measures and initiatives.[132] These programs include, but are not limited to, the Highway Safety Improvement Program[133], the National Highway Traffic Safety Administration Program[134], and the Federal Motor Carrier Safety Administration Program[135], to name a few. Programs like these address

---

[127] *See* David A. Weisbach and Jacob Nussim, *The Integration of Tax and Spending Program, 113* YALE L. J. 955, 960 (2003) (discussing the role of taxation and various funding mechanisms in supporting government investments).

[128] *See, e.g.,* Rui Wang & Shilong Li, *Research on the Influence Mechanism of Fiscal and Tax Policy on Green Economic Transition: From the Perspective of Industrial Structure Conduction Effect*, 26 ENV'T DEV. & SUSTAINABILITY 16129, 16130 (2024) (analyzing the role of fiscal policies in driving sustainable industrial transitions).

[129] *See, e.g.,* Tax Policy Center, *What is the Highway Trust Fund, and How is it Financed?*, Introduction, https://perma.cc/HLQ7-CF76 (last visited Feb. 7, 2025); Linda J. Cobiac, Anja Mizdrak, and Nick Wilson, *Cost-effectiveness of Raising Alcohol Excise Taxes to Reduce the Injury Burden of Road Traffic Crashe, 25* INJURY PREVENTION 421, 421 (2019) (finding that increasing alcohol taxes is a cost-effective strategy for reducing injuries from road traffic accidents.).

[130] 26 U.S.C. § 4081(a)(2)(A) (outlining the specific tax rates imposed on various types of taxable fuels, including gasoline, diesel fuel, and kerosene, detailing the cents-per-gallon rates applicable to each fuel type.); For a scholarly analysis of the federal fuel excise tax's influence on road safety and infrastructure, *see* Nima Safaei & Chao Zhou, *Gasoline Pricing Policies for Transportation Safety*, ARXIV (2020), https://perma.cc/N8JG-VLK4 (examining the relationship between gasoline prices, influenced by federal fuel taxes, and transportation fatality trends, providing insights into how tax policy affects road safety.).

[131] As of March 2024, the federal excise tax is 18.4 cents per gallon of gasoline and 24.4 cents per gallon of diesel. The Congressional Budget Office estimated that in 2023 the fuel excise tax provided 83% of Highway Trust Fund, with the additional funding coming from sales tax on tractors and heavy trucks, a tire excise tax for heavy vehicles, and an annual use tax for those vehicles.. *See* Congressional Budget Office, *The Status of the Highway Trust Fund: 2023 Update* (2023), https://perma.cc/75WH-4B3P.

[132] *See* Robert S. Kirk & William J. Mallett, *The Highway Trust Fund and the Treatment of Surface Transportation Programs in the Federal Budget,* CONG. RSCH. SERV. R45350 (2019), available at https://perma.cc/KEA8-3NSN (providing an in-depth examination of how HTF allocations support various transportation programs, including those specifically aimed at enhancing traffic safety.).

[133] *See* Fed. Highway Admin., *Highway Safety Improvement Program (HSIP),* https://perma.cc/U92L-9YQL.

[134] Nat'l Highway Traffic Safety Admin., *Resources Guide*, Highway Safety Grants Program (Feb. 2, 2024), https://perma.cc/C7K5-BJU7.

[135] *See* Fed. Motor Carrier Safety Admin., *Our Mission*, https://perma.cc/Q9RF-LY7N.

safety-related issues such as better signage, traffic management systems, and safer road designs.[136] Similar mechanisms function in the context of workplace and occupational safety.

C. Workplace and Occupational Safety

Tax policy employs multiple mechanisms to promote workplace safety through equipment and practice investments across sectors. The disaster tax relief provides benefits in federally declared disaster areas, including deductions for safety measures and rebuilding. [137] Similarly, disabled access credits support businesses implementing safety improvements like accessible exits and enhanced emergency systems. [138] The Advanced Energy Project Credit exemplifies direct safety enhancement through electric grid modernization, reducing risks of power outages, electrical fires, arc flash incidents, and equipment malfunctions.[139]

Capital investments in workplace equipment and technology yield multiple safety benefits. Newer work vehicles exemplify this dynamic, featuring enhanced safety technologies, improved reliability, and contributions to healthier ambient air quality in and around facilities.[140] Similar environmental and safety improvements arise from investments in oil-spill prevention equipment, low or zero-carbon process heat systems and carbon recapture technology, which enhance ambient

---

[136] Notably, the Proven Safety Countermeasures Initiative promotes strategies such as the installation of rumble strips, enhanced delineation, and the implementation of roundabouts to reduce roadway fatalities and serious injuries. Federal Highway Administration, *Proven Safety Countermeasures Initiative*, U.S. DEP'T OF TRANSP. (2024), https://perma.cc/ZL9R-9LTX. Additionally, the Safe Streets and Roads for All (SS4A) Grant Program supports local initiatives to develop comprehensive safety action plans, which often include measures like better signage and traffic calming designs to improve road safety. U.S. Department of Transportation, *Safe Streets and Roads for All Grant Program*, U.S. DEP'T OF TRANSP. (2024), https://perma.cc/G6TG-JLBG.

[137] *See, e.g.*, 26 U.S.C. § 165(i) (allowing taxpayers to claim deductions for losses resulting from federally declared disasters in the year immediately preceding the disaster); 26 C.F.R. § 1.165-7 (detailing procedures on claiming deductions for casualty and theft losses, specifying criteria for determining deductible amounts.); 26 U.S.C. § 7508A (granting the IRS authority to postpone tax deadlines for taxpayers affected by federally declared disasters), 26 U.S.C. § 1400S (establishes tax relief provisions for individuals and businesses in areas affected by certain disasters, including special deductions and credits.

[138] The Disabled Access Credit provides $ 5,000 in funding for spending to improve facilities and equipment to comply with ADA, which might contribute to occupational safety in a less direct way. *See* 26 U.S.C. §44. For a detailed overview of this credit, refer to the IRS's official guidance, *see* Internal Revenue Service, *Tax Benefits for Businesses Who Have Employees with Disabilities*, https://perma.cc/TCH5-7YDA. *See also* Silvia Bonaccio et al., *The Participation of People with Disabilities in the Workplace Across the Employment Cycle: Employer Concerns and Research Evidence*, 35 J. BUS. & PSYCH. 135 (2020) (examining employer concerns and provides evidence-based insights into the employment of individuals with disabilities, including discussions on incentives like the Disabled Access Credit.).

[139] *See* Tucker McGree, National Fire Protection Association, *Fires in Industrial and Manufacturing Properties: Supporting Tables*, Table 3 (2023), https://perma.cc/7MWB-63PW (stating that according to a 2017–2021 analysis, leading causes of fires in industrial properties included equipment or heat source failure, accounting for 732 fires (24%), electrical arcing at 454 fires (15%), and electrical failure or malfunction with 401 fires (13%).); see also U.S. Department of Labor, *Amid National Increase, U.S. Department of Labor Urges Midwest Employers to Emphasize Electrical Safety after 4 Workplace Deaths in Missouri, Kansas*, Occupational Safety and Health Administration (Nov. 9, 2021) https://perma.cc/TNU5-LXJQ (pointing to statistical data on 166 workplace deaths related to electrocution in 2019, reflecting a 3.75% increase over the previous year.).

[140] *See* Yang Shen and Xiuwu Zhang, *Towards a Low-Carbon and Beautiful World: Assessing the Impact of Digital Technology on the Common Benefits of Pollution Reduction and Carbon Reduction,* 196 ENV. MONITOR. ASSESS. 695, 700 (2024) (discussing the benefits of investments in low-carbon technologies in improving air quality and reducing environmental impacts).

air quality conditions.[141] Indeed, general technological modernization in workplace settings achieves both direct and indirect safety enhancements through updated equipment and improved manufacturing processes.[142]

The tax code promotes these safety-enhancing investments through accelerated cost recovery mechanisms. Immediate expensing—representing the most aggressive form of depreciation available—permits businesses to write off up to $1.25 million of qualifying property costs, with deductions phasing out dollar-for-dollar once total purchases exceed $3.13 million.[143] While this mechanism does not explicitly target safety-related expenditures, it enables immediate deduction of qualifying new and used equipment and facility purchases. [144] These deductions, though limited to taxable income, can carry forward to subsequent years under the same income and dollar constraints.[145] This accelerated expensing of new equipment and safety infrastructure—including fire protection and security systems—incentivizes businesses to prioritize safety enhancements, potentially reducing workplace hazards and improving overall safety conditions. [146]

Bonus depreciation provides another significant tax incentive mechanism, allowing an 80% immediate deduction of property costs in the service year, encompassing safety-oriented equipment like protective gear and safety guards.[147] The provision covers a broad spectrum of qualifying property: tangible assets, computer software, qualified improvements, and nonresidential real property enhancements including critical safety infrastructure like HVAC, fire protection, and security systems.[148] This mechanism complements immediate expensing by enabling additional deductions beyond standard expensing limits, thus accelerating cost recovery and indirectly promoting investments in newer, inherently safer equipment and facilities.[149]

Lastly, enacted in August 2022, the Inflation Reduction Act "(IRA") employed tax credits and incentives to drive investments in clean energy, reduce greenhouse gas emissions, all the while

---

[141] *See* Gaia J. Larsen, *Skewed Incentives: How Offshore Drilling Policies Fail to Induce Innovation to Reduce Social and Environmental Costs*, 31 STAN. ENVTL. L.J. 139, 148 (2012) (discussing safety technology investments for deepwater drilling, highlighting the urgent need for policies to enhance drilling safety and prevent future catastrophic spills.).

[142] *See* Sang-Heon Lee & Ji-Hoon Kim, *Technological Advancements in Industrial Safety: Intelligent Devices for Accident Prevention*, 9 SAFETY 35, 35–42 (2023), https://perma.cc/BAU2-K7W5 (exploring how updated technologies reduce industrial accidents and enhance worker safety).

[143] 26 U.S.C. §179. Rev. Proc. 2024-40, available at https://perma.cc/NW3Z-Z3BF (last visited Feb. 2, 2025).

[144] 26 U.S.C. § 179; Gary Guenther, *The Section 179 and Section 168(k) Expensing Allowances: Current Law, Economic Effects, and Selected Policy Issues*, CONG. RES. SER. 1–2 (Feb 7, 2024), https://perma.cc/XS3M-AFWY (last visited Feb. 2, 2025) (hereunder "The Section 179 and Section 168 CRS Report").

[145] 26 U.S.C. §179(b)(3)(B).

[146] Internal Revenue Service Announcement, *Depreciation Expense Helps Business Owners Keep More Money*, IRS (Mar. 16, 2020) https://perma.cc/VB9L-6P89 (last visited Feb. 9, 2025). For example, the Security Industry Association provides a detailed overview of these tax incentives in their fact sheet. *See* Security Industry Association, *New Tax Incentives for Security and Fire Protection Systems* (2018), https://perma.cc/ZZ33-2BKW.

[147] 26 U.S.C. §168(k) (providing additional allowance equal to the applicable percentage of the adjusted basis of the qualified property with a recovery period of 20 years or less).

[148] *Id.,* Internal Revenue Service, *Publication 946: How to Depreciate Property* 16 (2024), https://perma.cc/RJ6R-ELXQ.

[149] *Id.,* at 14.

promoting safe work practices.[150] These incentives operate through a two-tier structure: the act provides a base credit for eligible projects and a bonus credit for those meeting additional requirements, such as adhering to prevailing wage standards and apprenticeship programs.[151] For example, a renewable energy project installing solar panels could receive a 6% investment tax credit, which increases to 30% if labor and apprenticeship conditions are met.[152] Another mechanism is training requirement for apprentices by journeymen, which is meant to promote the training and development of a skilled workforce, further enhancing workplace safety standards.[153]

D. Safety Research Incentives

While the preceding analysis examined tax incentives for implementing existing safety technologies, a distinct challenge emerges in the domain of fundamental safety research. Innovation drives long-term economic growth and living standard improvements, with research and development (R&D) serving as the cornerstone of sustained technological progress. [154] However, this domain exhibits a classic market failure: social returns from research, particularly in fundamental areas characterized by uncertainty and non-rivalry, vastly exceed private returns.[155] This disparity creates a compelling case for government intervention, especially as the U.S. research funding landscape has shifted dramatically toward private sector dominance, with businesses now conducting more than two-thirds of all U.S. R&D activities. [156]

The tax system has historically addressed this market failure through two primary mechanisms: immediate R&D expensing and the R&D tax credit. Until recently, federal policy permitted full,

---

[150] The Inflation Reduction Act of 2022, Pub. L. No. 117-169, 136 Stat. 1818 (The Act incentivizes specific sectors, including renewable energy generation, energy-efficient upgrades, electric vehicle adoption, and domestic manufacturing of clean energy components. With a 10-year timeline for these credits, the IRA provides businesses with the long-term stability needed for sustainable planning and growth, fostering economic development alongside environmental progress.).

[151] *See* IRS Notice 2022-61, Guidance on Prevailing Wage and Apprenticeship Requirements Under Section 45(b)(6) and Other Provisions of the Internal Revenue Code, 87 Fed. Reg. 73,978 (Nov. 30, 2022).

[152] *See* Deborah Tam, *Prevailing Wage and Apprenticeship Requirements for Inflation Reduction Act Clean Energy Tax* Credits, Thomson Reuters (Dec. 2, 2022) https://perma.cc/V9TY-3ZBV (detailing President Biden's Inflation Reduction Act provisions for increased clean energy tax credits or deduction amounts if certain prevailing wage and apprenticeship requirements are met.).

[153] *See generally* Rosemary K Sokas et al., *An Intervention Effectiveness Study of Hazard Awareness Training in the Construction Building Trade,* 124 PUB. HEALTH REP. SUPP 1, 160 (2009), https://perma.cc/SH37-RHD3 (providing an overview on the role of apprenticeship programs in and discusses their impact on construction workforce development and safety).

[154] *See generally* Robert M. Solow, *Technical Change and the Aggregate Production Function*, 39 REV. ECON. & STAT. 312, 316 (1957) (demonstrating that technological innovation significantly contributes to increases in output and productivity, thereby enhancing living standards over time.).

[155] *See* Bronwyn H. Hall & Josh Lerner, *The Financing of R&D and Innovation*, *in* 1 HANDBOOK OF THE ECONOMICS OF INNOVATION. 609, 610 (2009), https://perma.cc/BMQ6-GNYU (discussing the critical role of both public and private R&D funding in fostering technological advancements). *See* Brett M. Frischmann & Mark A. Lemley, *Spillovers*, 100 COLUM. L. REV. 101, 115 (2006) (providing a comprehensive analysis of the economic dynamics of R&D spillovers and the associated market failures.).

[156] *See* National Science Board, *Research and Development: U.S. Trends and International Comparisons*, SCIENCE AND ENGINEERING INDICATORS (2022), https://perma.cc/R3EL-2GAA (providing an in-depth examination of the shifts in U.S. research expenditures, highlighting the changing roles of public and private sectors in funding and conducting R&D.).

immediate deductions for R&D expenditures to encourage private sector investment.[157] However, the Tax Cuts and Jobs Act of 2017 mandated five-year amortization for domestic research (fifteen years for foreign research), potentially deterring R&D investment, particularly among resource-constrained firms.[158]

The R&D credit, introduced in 1981, aims to incentivize increased research investment rather than merely subsidizing existing R&D activities.[159] The credit offers multiple pathways: a traditional 20% credit for qualified expenses above a base amount,[160] an alternative simplified 14% credit,[161] an energy research credit (20% flat rate),[162] and a basic research credit for university collaboration.[163] The Basic Research credit guarantees same advantages to companies when they outsource scientific research and engage in collaborations with universities.[164] In addition to the federal credit,

---

[157] *See* 26 U.S.C. §174.

[158] *See* 26 U.S.C. §174(a)(2)(B). The Tax Cuts and Jobs Act the government abruptly eliminated R&D expensing starting 2022 and left such expenses to be capitalized ratably over five years. Tax Cuts and Jobs Act of 2017, Pub. L. No. 115-97, 131 Stat. 2054. *See also* Richard Ray, *Amortizing Research & Development Expenditures Under the TCJA*, J. ACCOUNTANCY (2022), https://perma.cc/9SST-QKBE (explaining how the Act requires capitalization and amortization of R&D expenses after December 31, 2021). *See also* Alex Muresianu, Garrett Watson, *The Economic Impact of Restoring Immediate Expensing for R&D Costs*, TAX FOUNDATION 5 (2021), https://perma.cc/5KH7-C7GR (warning from significant implications and the incentives to participate in research activities for business taxpayers due amortization of R&D expenses). Nevertheless, the House enacted legislation on January 31, 2024, to reinstate the immediate expensing of research and development expenditures. Enacted in January 2024, the $78 billion Tax Relief for American Families and Workers Act undid the previous modification and restored the practice of expensing research and development expenses. Research and development expenditures would be fully deductible until 2025. Tax Relief for American Families and Workers Act of 2024, H.R. 7024, 118th Cong. (2024) (reinstating the immediate expensing of domestic research and development expenditures under I.R.C. § 174).

[159] *See generally* Mirit Eyal-Cohen & Ana Santos Rutschman, *Promoting Vaccine Innovation*, 82 OHIO ST. L. J. 1003, 1029 (2022) (surveying the history of the R&D credit and its efficiency in the context of pharmaceutical investments).

[160] *See* 26 U.S.C. § 41(a)(1).

[161] *See* 26 U.S.C. § 41(c)(4)(A). A company's alternative simplified credit (ASC) is equivalent to 14% of its QREs in excess of 50% of its moving average QREs over the preceding three years. If the taxpayer has no qualified research expenses in any of 3 preceding taxable years the alternative simplified credit rate is 6 percent of qualified research expenses. 26 U.S.C. § 41(c)(4)(B). *See* U.S. Gov't Accountability Off., GAO-10-136, *Tax Policy: The Research Tax Credit's Design and Administration Can Be Improved* (2009), https://perma.cc/99HR-3T64 (claiming that Although the alternative credit may make the calculation of the credit easier, it offers less marginal incentives to invest than the standard credit.). *See also* Daniel Karnis, *How the R&D Tax Credit Is Calculated*, 1 J. ACCT. 28 (2010) (examining the R&D Tax Credit calculation methods, including scenarios where firms have no prior research history.).

[162] *See* 26 U.S.C. § 41(c)(4)(A) (Twenty percent of a company's qualified research expenditures (QREs) on payments to nonprofit organizations for the purpose of undertaking energy research in the public interest is eligible for the energy research credit.).

[163] *See* 26 U.S.C. §41(e); According to the National Science Foundation, "basic research" is "any original inquiry for the progress of scientific knowledge that does not have a definite commercial purpose." *National Science Foundation, Annual Report* 1953, at 6 (1953), https://perma.cc/259S-D9PB.

[164] *See generally* Ufuk,Akcigit, Douglas Hanley, and Nicolas Serrano-Velarde, *Back to Basics: Basic Research Spillovers, Innovation Policy, and Growth*, 88 REV. ECON. STUD. 1, 10 (2021) (comparing basic to applied research credit and identifying the spillovers embed between private firms and a public research sector.). It is worth noting that such collaboration is encouraged also through patent donations, which offer a tax deduction for intellectual property firms transfer to non-profit organizations. *See generally* Lily Kahng, *The Taxation of Intellectual Capital*, 66 FLA. L. REV. 2229, 2267-77 (2014)

approximately 35 states have a research tax credit.[165] The latest tax spending report from the Joint Committee on Taxation estimates in 2025 a $22 billion loss in revenue as a result of the R&D tax credit.[166] While these credits can offset both income and payroll taxes and carry forward for twenty years,[167] their effectiveness faces several constraints, especially in the AI safety context.[168] Empirical studies show mixed results on R&D spending impact, with critics noting potential expense reclassification rather than new research investment.[169]

Most critically for AI safety, current tax incentives explicitly exclude quality assurance and safety testing from qualifying research expenses.[170] Clinical trials or similar testing may be eligible if they

(referencing extensive literature on taxation of intangibles); Xuan-Thao Nguyen & Jeffrey A. Maine, *Equity and Efficiency in Intellectual Property Taxation*, 76 BROOK. L. REV. 1, 1-8 (2010) (reviewing and criticizing tax rules relating to patents, copyrights, and trademarks).

[165] *See* MICHAEL D. RASHKIN, PRACTICAL GUIDE TO RESEARCH AND DEVELOPMENT TAX INCENTIVES: FEDERAL, STATE AND FOREIGN 1001 (2007) (providing a comprehensive analysis of state R&D tax).

[166] The Joint Committee on Taxation, *Estimates of Federal Tax Expenditures for Fiscal Years 2024-2028*, JCX-48-24, at 22 (December 11, 2024), https://perma.cc/U3TP-RFR4.

[167] *See* 26 U.S.C. § 39(a)(1). The fixed-base ratio is a historical percentage denoting the company's total "qualified research expenditures" over total gross receipts. In calculating the credit, the firm's base period research was not permitted to be less than 50% of the current year's research spending. The credit's statutory rate was initially set at 25 percent and applied only to increases in a firm's research spending over its average spending in a base period consisting of the previous three years. 26 U.S.C. § 41(c).

[168] Many expenses related to AI safety are ineligible as they involve routine data collection, routine quality-control testing, social science research, grant-funding research, or research conducted outside the United States. 26 U.S.C. § 41(d)(4)(B)-(H). The definition also consists of a "specified" R&D expense, which includes any amount paid or incurred in connection with the development of any software. *Id.* Contract research expenses are limited to 65 percent of any amount paid to any person (other than an employee of the taxpayer) for qualified research. 26 USC § 41(b)(3)(A).

[168] Many expenses related to AI safety are ineligible as they involve routine

[169] *See, e.g.,* Russell Thomson, *The Effectiveness of R&D Tax Credits*, 99 REV. ECON. STAT. 544, 547 (2017), https://doi.org/10.1162/REST_a_00559 (claiming long run every $1 of R&D tax credit translates to around $4 in new R&D investment); Antoine Dechezleprêtre, et al., *Do Tax Incentives for Research Increase Firm Innovation? An RD Design for R&D*, NBER Working Paper 22405, 28, (2016), https://perma.cc/7YUB-CTEU (finding evidence that for every €1 of tax subsidy there is an increase of €1.7 in R&D.); Wesley Yin, *Market Incentives and Pharmaceutical Innovation,* 27 J. HEALTH ECON. 1060, 1061 (2008) (demonstrating Tax credits can stimulate R&D); Irem Guceri & Li Liu, *Effectiveness of Fiscal Incentives for R&D: Quasi-experimental Evidence*, 11 AM. ECON. J.: ECON. POL'Y 266 (2019), https://perma.cc/53PD-SSPH (finding that $1 in additional private R&D spending per dollar foregone in tax revenue.); Jieun Choi, *Do Government Incentives to Promote R&D Increase Private R&D Investment?*, 37 WORLD BANK RES. OBS. 204 (2022) ("R&D incentives generally increase private R&D, but to a varying extent depending on incentive types, countries' income levels, industry and firm characteristics, and the design and implementation of the incentives."). *But see* Jennie S. Stathis, *The Research and Development Tax Credit Has Stimulated Some Additional Research Spending,* Government Accountability Office, Sept. 5, 1989, https://perma.cc/2S9C-ZSCW (concluding the research credit had just a little effect—for every $1 in tax subsidies, between $0.15 and $0.36 was spent on R&D); Russell K. Thomson, The Effectiveness of R&D Tax Credits: Cross-Industry Evidence (2013), http://dx.doi.org/10.2139/ssrn.2275094 (providing a cross-country analysis indicates that in the short term, industries increase R&D investment by only $0.24 for every dollar of tax revenue forgone, implying that a significant portion of the credited R&D expenditure might have been undertaken regardless of the tax incentives.); Robert Eisner, Steven H. Albert & Martin A. Sullivan, *The New Incremental Tax Credit for R&D: Incentive or Disincentive?*, 37 NAT'L TAX J. 171, 181 (1984) (reporting a limited impact of the research credit).

[170] *See* 26 C.F.R. §1.174-2(a)(3).

are part of the process to develop new technology or prove feasibility.[171] However, trials conducted for routine market testing or compliance are generally excluded.[172] This means that quality assurance activities, including post-market safety testing and compliance verification, generally fall outside the definition of qualified research expenses unless part of new technology development. [173] This creates a significant gap: existing incentives not only fail to prioritize AI safety research but may actually discourage such investment by providing equal or greater inducements for less safety-oriented innovation.

III. A Tax Framework for Safe AI Development

The preceding analysis illuminates a critical market failure in AI development: while private entities capture the benefits of capability advances, the risks and potential harms are broadly socialized. This misalignment creates systematic underinvestment in safety research and protocols relative to capability development. Current tax frameworks, particularly R&D incentives, exacerbate rather than ameliorate this dynamic by subsidizing capability research without differentiating safety-oriented initiatives. We propose leveraging fiscal policy to address this market failure through a tripartite framework that builds upon established precedents in energy efficiency, workplace safety, and environmental protection.[174] By integrating producer-side safety incentives, market-based certification mechanisms, and corrective tax measures, our approach harnesses existing administrative competencies while addressing the unique challenges of emerging AI systems. Importantly, all of these proposals are grounded in practices already employed, albeit in a diffused manner, by the tax system, and so they draw on existing institutional competencies.

---

[171] *But See* Thomson Reuters, *Section 174 Expenditures: What Qualifies and What Doesn't*, Thomson Reuters Tax & Accounting, https://perma.cc/4FV5-HBKA (explaining that quality control testing does not qualify as a research and experimental expense under Section 174).

[172] *See* 26 C.F.R. §1.174-2(a)(3) (disallowing section 174 treatment for certain activities, including: ordinary testing or inspection of materials or products for quality control, efficiency surveys, management studies, consumer surveys, advertising or promotions, acquisition of another's patent, model, production or process, or research in connection with literary, historical, or similar projects.).

[173] *See* Internal Revenue Service, Audit Techniques Guide: Credit for Increasing Research Activities (Research Tax Credit) IRC § 41—Qualified Research Activities, IRS.gov, https://perma.cc/8QV8-R5WY (disallowing "section 174 treatment for certain activities for ordinary testing or inspection of materials or products for quality control").

[174] It is an open-ended question how much money should be spent on reducing AI risks; the primary if initial attempt to answer this question offers a qualified estimate of 8% of GDP, which would amount to a stunning 2 trillion per year. Charles I. Jones, *How Much Should We Spend to Reduce A.I.'s Existential Risk?* (Jan. 26, 2025) (unpublished manuscript), available at https://perma.cc/BUW6-UAQW. But we do not take a positive stance on this question other than to note that there are strong reasons to spend considerable amounts on encouraging safety.

A Novel Incentive, Allocation, and Distribution Mechanism

Hemel and Ouellette identified a critical insight for technology governance: the critical importance of regulatory pluralism.[175] Traditional regulatory measures, such as liability rules, fines, and audits, respond to accidents after they happen. In contrast, tax levers like credits, deductions, and accelerated depreciation push forms to invest in preventive measures, weaving safety into their core culture and strategic planning. Together, both have ex-ante effects, contributing to what some have called the "Swiss cheese" model of safety, where it is recognized that safety depends not on the perfection of any specific security 'slice,' but rather by the stacking of several imperfect slices.[176] The balanced pluralistic approach likewise recognizes that no single tool can singlehandedly solve AI's risks, but taken together they make safety both imperative and economically viable for firms.

By linking tax benefits directly to verifiable AI safety expenditures—such as workforce training, alignment research, and safer product design—the government spurs private investment and helps American labs stay competitive in the heating global AI race.[177] These incentives reduce the cost of building robust safety features, encourage proactive measures, and ensure that safety becomes an organizational priority rather than an afterthought. While precise quantification of safety benefits remains challenging and would require time to develop, delaying implementation until perfect information becomes available would effectively privilege capability development. The framework thus calls for baseline incentives now, combined with flexible mechanisms to refine these incentives as our understanding of AI safety matures.

1. Business Tax-Incentives for Investments in AI Safety

Research incentives constitute a critical yet underutilized governance mechanism for advancing AI safety protocols through strategically calibrated fiscal interventions. These instruments operate through three distinct channels—research credits, expensing rules, and basic research incentives—each addressing specific market failures in safety-oriented research and development.[178] By systematically modifying the underlying cost structure of safety research, targeted tax credits can catalyze investment across crucial technical domains, including: (1) alignment research and verification methodologies, (2) adversarial robustness testing and validation, and (3) interpretability frameworks and monitoring systems.[179] This tripartite approach leverages existing administrative

---

[175] *See* Hemel & Ouellette, *supra* note 23, at 544; Daniel J. Hemel & Lisa Larrimore Ouellette, *Law and the New Dynamic Public Finance*, 2020 WIS. L. REV. 645, 650 (2020) (advocating for a pluralistic approach to regulatory policy).

[176] *See* James Reason, *Human Error: Models and Management*, 320 BMJ 768, 768–70 (2000) ("[common safety measures] are more like slices of Swiss cheese, having many holes . . . The presence of holes in any one "slice" does not normally cause a bad outcome. Usually, this can happen only when the holes in many layers momentarily line up").

[177] *See, e.g.,* Mariarosaria Comunale and Andrea Manera, *Fiscal Policies for Managing AI's Economic Impacts,* IMF Working Paper No. 24/065 (2024), https://perma.cc/7GS7-9968 (surveying the role of tax incentives in promoting responsible adaptation to technological advancements in AI).

[178] *See, e.g.*, Bronwyn H. Hall & John Van Reenen, *The Impact of the Research and Development Tax Credit on Innovation: A Meta-Analysis of Causal Evidence*, 12 AM. ECON. J. ECON. POL'Y 1, 1–25 (2020) (finding that R&D tax incentives significantly boost firms' innovation outputs, including increased R&D spending and quality-adjusted patenting, with sustained effects over several years, particularly benefiting financially constrained firms).

[179] On various safety measures, see *supra* note 20.

competencies while addressing the systematic underinvestment in safety research that characterizes current AI development trajectories.

The pharmaceutical sector offers an instructive comparative framework. Contemporary pharmaceutical R&D expenditures are directed toward both novel therapeutic discovery and rigorous safety validation through clinical trials—activities that generate pure public goods in the form of generalizable knowledge. The existing R&D tax credit and Basic Research credit frameworks create targeted incentives for broad research initiatives,[180] while specialized mechanisms like the Orphan Drug tax credit address specific market failures in rare disease research.[181] This regulatory architecture demonstrates the potential for precisely targeted fiscal interventions to address systematic underinvestment in socially beneficial research domains.

Current R&D credit qualification criteria present significant limitations for AI safety research. Routine data collection, cleaning, and processing, as well as the implementation of existing AI tools without experimentation, are currently considered operational activities and excluded.[182] The exclusions extend systematically across multiple domains: research conducted outside the U.S., third-party funded research, market research costs, legal compliance expenses, infrastructure investments (like GPUs or cloud computing), general employee training, quality assurance testing, and aesthetic or interface design without technical uncertainty resolution.[183]

The case for similar interventions in AI safety is particularly compelling given the technology's universal impact radius.[184] Unlike pharmaceutical products which primarily affect direct consumers, AI systems generate broad externalities affecting both users and non-users through their societal deployment and network effects.[185] We propose adapting the pharmaceutical model to create an "AI Safety Research Tax Credit" modeled after the Orphan Drug Credit.[186] This credit would reward expanded types of safety-oriented research activities including red team testing, explainability

---

[180] *See* 26 U.S.C. § 41(a), (e). The Basic Research Credit offers for-profit firms a similar tax credit for payments made to nonprofit organizations for their collaborative research. 26 U.S.C. § 41(e)(6). *See generally* MICHAEL D. RASHKIN, RESEARCH AND DEVELOPMENT TAX INCENTIVES: FEDERAL, STATE, AND FOREIGN 330 (3RD ED., 2007) (providing an in-depth examination of the legislative intent and practical application of the R&D Tax Credit, highlighting its role in fostering significant advancements in various industries.).

[181] *See* 26 U.S.C. § 45C (providing tax incentives for qualified clinical testing expenses related to the development of drugs for rare diseases or conditions, also known as orphan drugs that affect small patient populations). *See also* Mark A. Lemley, Lisa Larrimore Ouellette & Rachel E. Sachs, *The Medicare Innovation Subsidy*, 95 N.Y.U. L. REV. 75, 120 (2020) (exploring the advantages of qualifying innovation research in orphan drugs).

[182] *See* 26 U.S.C. §41. *See also* Warren Averett, *Cloud Computing Services and the R&D Tax Credit* WARREN AVERETT INSIGHTS (Mar. 24, 2023) https://perma.cc/8VHP-H9MN (explaining that while cloud computing costs directly tied to research may qualify as QREs, expenses for market research, legal compliance, and depreciable property like GPUs are disqualified).

[183] *See* Internal Revenue Service, *Audit Techniques Guide: Credit for Increasing Research Activities,* IRC § 41 - Qualified Research Activities, C. Exclusions, https://perma.cc/8QV8-R5WY (detailing the exclusions form qualified research expenses); Jennifer Frost, et al., *The Research Credit and Funded Research*, THE TAX ADVISER (Mar. 1, 2023) https://perma.cc/FVS8-QCXL (noting that research funded by grants, contracts, or third parties is excluded from eligibility for the research credit).

[184] *See e.g.,* Tokson & Arbel, *AI X-Risk*, note 108.

[185] *Id*.

[186] See *supra* note 7 and accompanying text.

requirements, training monitoring methodologies, and robust guardrail systems.[187] Operating through direct reduction of tax liability, the credit will create immediate financial benefits for firms prioritizing AI safety research, including a structure benefiting firms that lack scale, scope, and age.[188]

The framework could implement differentiated credit rates across research phases. During quality assurance and testing phases, elevated credit rates (e.g., 50% for startups, 25% for established firms) would create enhanced incentives for thorough safety validation. This temporal variation in credit rates reflects the particular importance of safety verification in later development stages. Moreover, the Basic Research Credit could be redesigned to specifically encourage research institutions and public-private partnerships to focus on investigating AI safety. Such targeted incentives would foster collaborations between academic institutions, independent research organizations, and private companies to address critical challenges in AI safety through development of safety protocols, algorithmic robustness studies, and risk mitigation strategies.[189]

A second proposed mechanism leverages expensing rules to create differentiated business incentives between safety and capability investments. Expensing rules significantly influence corporate behavior and tax compliance, [190] with accelerated depreciation reducing complexity while incentivizing investment.[191] Currently, businesses can claim deductions under Section 168 (bonus depreciation) or Section 179 (immediate expensing) for AI products meeting eligibility requirements. Bonus depreciation enables 50% first-year cost deduction for tangible property rather than following statutory recovery periods,[192] with AI-related hardware (servers, GPUs, edge devices) typically qualifying for 5-7 year depreciation schedules.[193]

We propose expanding this framework to permit immediate expensing of qualified safety-related expenditures while mandating extended amortization periods for pure capability investments. This dual approach creates complementary incentives: reducing effective costs for safety investments through immediate deductibility of research, testing protocols, and monitoring systems, while encouraging longer-term stability considerations through extended amortization requirements for

---

[187] *See supra* note 20.

[188] *See generally*, Mirit Eyal-Cohen, *The Cost of Inexperience,* 69 ALA. L. REV. *859* (2018) (demonstrating how firms that lack scale, scope, and age are disadvantaged via heavier regulatory burdens).

[189] *See, e.g.,* Paolo Bova, Alessandro Di Stefano & The Anh Han, *Both Eyes Open: Vigilant Incentives Help Regulatory Markets Improve AI Safety*, ARXIV (Mar. 6, 2023) (proposing to design non-tax government incentives to build regulatory markets that deter reckless behavior in AI development).

[190] *See* Israel Klein, *Contemptuous Tax Reporting*, 2019 WIS. L. REV. 1161, 1161 (arguing that R&D tax incentives often lead to abusive tax practices).

[191] *Id.* at 1170.

[192] *See* 26 U.S.C. §168(k); 26 C.F.R. §1.168(k)-2. See also *Bonus Depreciation Regulations Favorable to Taxpayers, Tax Adviser* (Feb. 1, 2020), https://perma.cc/D9SX-A48P (explaining the implications of regulations regarding depreciation deduction under Section 168(k)).

[193] *See* Gary Guenther, *The Section 179 and Section 168(k) Expensing Allowances: Current Law, Economic Effects, and Selected Policy Issues,* CONG. RSCH. SERV., RL31852, 12 (Feb. 7, 2024), https://perma.cc/3L8L-NELB (detailing the economic effects of Section 179 and bonus depreciation allowances for 5 years property like GPUs).

capability-focused investments.[194] The framework would qualify safety-oriented investments—including testing frameworks, alignment research, and monitoring systems—for immediate expensing, while requiring longer depreciation or amortization for investments purely targeting increased model size or computational capacity. This preferential treatment extends to development costs like software, prototypes, licenses, and patents that would otherwise face extended amortization periods. The extended amortization period for capability investments serves as a natural brake on the "move fast and break things" mentality that has characterized much of AI development, subtly reshaping organizational decision-making toward more responsible and sustainable practices.

A significant challenge emerges from the current tax treatment of AI systems themselves. Under existing law, while tangible components like servers and GPUs qualify for bonus depreciation,[195] custom-developed AI software must typically be amortized over 15 years as an intangible asset. This extended amortization period is said to significantly impede financial returns for companies investing in innovation such as AI systems. [196] To address this barrier, we propose extending immediate expensing benefits to certified safe AI systems. Rather than waiting many years to recover development costs, companies that meet rigorous safety certification standards would qualify for immediate deduction of their AI software development expenses. This modification creates a powerful incentive for safety-conscious development while reducing the financial burden on companies committed to responsible AI practices.

The framework can be further enhanced through strategic use of payroll tax deductibility.[197] Recent proposals advocate increasing the maximum payroll tax liability offset by the R&D credit, raising gross receipts eligibility barriers, and extending the startup claim period from four to eight years.[198] Making the credit refundable would particularly benefit high-tech industries and firms lacking scale, scope, and age—entities that often lack the financial sophistication to navigate complex

---

[194] *See* Amy I. Kinkaid, Charles E. Federanich, Pease Bell, *Planning Opportunities: Section 179 Expensing vs. Bonus Depreciation, Tax Adviser* (Dec. 1, 2024), https://perma.cc/T7EK-KMCW (analyzing the distinct advantages of Section 179 expensing and bonus depreciation under Section 168(k), focusing on taxable income limitations and capital expenditure benefits).

[195] Servers and GPUs qualify for bonus depreciation under IRC § 168(k) and Section 179 expensing because they are classified as tangible personal property with a recovery period of 5 years under MACRS, making them eligible for accelerated depreciation incentives if used at least 50% for business purposes. *See* 26 U.S.C. § 168(k), *Additional First-Year Depreciation Deduction (Bonus Depreciation)* (2024), https://perma.cc/69YG-W3FQ (last visited Jan. 30, 2025).

[196] *See* Mary Cowx, Rebecca Lester, and Michelle L. Nessa, *The Consequences of Limiting the Tax Deductibility of R&D* (July 23, 2024) https://dx.doi.org/10.2139/ssrn.4998845 (last visited Jan. 30, 2025) (providing evidence about the detrimental effects of limiting innovation tax incentives.).

[197] *See* Wendy Landrum and Ted Butler, *Gross Receipts Definition Plays Key Role in R&D Credit Limit for Startups*, THE TAX ADVISER (Oct. 2017), https://perma.cc/2ESN-AKCH (explaining the $5 million gross receipts limit and the requirement of no gross receipts for any tax year before the five-tax-year period ending with the current tax year); *Qualified Small Business Payroll Tax Credit for Increasing Research Activities,* INTERNAL REVENUE SERVICE, https://perma.cc/TN5P-NSZ8 (providing guidance on eligibility criteria, election procedures, and timing for claiming the payroll tax offset).

[198] The American Innovation and Jobs Act, S. 866, 118th Cong. (2023) proposed by Republican U.S. Senator Todd Young of Indiana and Democratic U.S. Senator Maggie Hassan of New Hampshire proposed to increase it from $250,000 to $500,000, and then to $750,000 over the next decade.

R&D documentation requirements.[199] This modification would make the AI safety research credit more accessible to startup enterprises with limited income tax obligations by providing immediate liquidity through payroll tax refunds, including Medicare and unemployment insurance payments.[200]

The Basic Research Credit framework offers a promising mechanism for advancing *foundational* AI safety research.[201] Unlike conventional R&D credits that primarily incentivize applied research with clear commercial trajectories, the Basic Research Credit specifically targets fundamental scientific inquiry devoid of immediate profit potential.[202] This structural alignment between credit design and the public goods characteristics of AI safety research warrants systematic examination. The institutional framework established through basic research payments—defined as transfers from corporations to qualified educational or tax-exempt organizations pursuant to written agreements—creates a structured mechanism for systematic knowledge generation. [203] While the existing calculation methodology introduces complexity to an already intricate incentive system, this complexity serves crucial policy objectives.[204] The differentiated treatment of basic research reflects its distinctive characteristics: extended temporal horizons, heightened uncertainty coefficients, and more diffuse societal benefits relative to applied research paradigms.[205]

We therefore propose expanding this framework to create a specialized "Basic AI Safety Research Credit." This mechanism would preserve the essential architectural features of the basic research credit—particularly its emphasis on pre-commercial investigation—while incorporating specific provisions for AI safety research. The proposed credit would incentivize formal research partnerships between commercial AI developers and qualified research institutions, establishing structured channels for knowledge transfer and collaborative investigation. This institutional

---

[199] *See generally* Eyal-Cohen, *supra* note 188.

[200] There was a proposal made in the Research and Development Tax Credit Expansion Act of 2019 to enlarge the number of companies eligible for a refund of payroll taxes by increasing the maximum amount of gross sales from $5 million to $10 million and increase the refundable element of the credit for new and small enterprises from $250,000 to $500,000. *See* Pinky Shodhan et al., *The Research Credit: Payroll Tax Offset, Tax Matters,* J. ACCOUNTANCY (Jan. 1, 2023), https://perma.cc/XE73-UYG9.

[201] The Basic Research Credit is calculated as the taxpayer's basic research payments over its qualified organization base period amount. The portion of the basic research payments which does not exceed the taxpayer's qualified organization base period amount is treated as contract expenses for purposes of the R&D tax credit, which can be claimed concurrent with the basic research credit. *See* 26 U.S.C. § 41(e)(4) and (5).

[202] *See, e.g.,* Sohvi Leih & David J. Teece, *Basic Research,* in THE PALGRAVE ENCYCLOPEDIA OF STRATEGIC MANAGEMENT (Augier, M., Teece, D.J. (eds) 2018) https://perma.cc/Y9EA-BFNT (defining basic research as "systematic study directed toward fuller knowledge or understanding of the fundamental aspects of phenomena and of observable facts without specific applications towards processes or products in mind").

[203] 26 U.S.C. § 41(e) (allowing businesses to claim a tax credit for qualified basic research expenses paid to universities and other scientific research organizations under certain information-sharing agreements.).

[204] *See* International Monetary Fund, *Effectiveness of Fiscal Incentives for R&D: Quasi-Experimental Evidence*, IMF Working Paper No. 17/84 (2017) (analyzing the impact of tax incentives on research and development activities using quasi-experimental methods), https://perma.cc/MC4T-MBX2; Jason J. Fichtner, *Can a Research and Development Tax Credit Be Properly Designed*?, MERCATUS CTR. (2014) (examining the challenges of the R&D tax credit especially its ambiguity and uncertainty), https://perma.cc/KZ8R-YZ34.

[205] *See* Bettina Becker, *Public R&D Policies and Private R&D Investment: A Survey of the Empirical Evidence*, 29 J. ECON. SURV. 917 (2015), (explaining the standard economic view of basic R&D as an under-supplied public good)

arrangement deliberately leverages proven complementary organizational capabilities: commercial entities contribute practical expertise and computational infrastructure, while academic institutions provide theoretical depth and commitment to open scientific inquiry.[206]

However, significant limitations exist in current basic tax credit qualification criteria for AI safety research.[207] These exclusions span multiple domains: non-technological activities like ethical studies and market research; research collaborations conducted outside the U.S.; [208] funded research where funding parties retain substantial rights; routine data collection and quality assurance testing;[209] administrative and managerial costs related to safety project oversight; capital expenditures for fixed assets;[210] personnel training costs; implementation of safety measures without experimentation; system improvements lacking technical uncertainty resolution; and legal, patent, marketing, and customer support efforts.[211] These exclusions systematically constrain the development of comprehensive safety frameworks.

The preceding analysis illuminates the diverse array of fiscal mechanisms within the existing tax framework that can be mobilized to promote safe AI development while requiring minimal systemic adaptation. From basic research credits to targeted safety incentives, these instruments leverage established institutional architectures and administrative competencies to address the critical market inefficiencies of underinvestment in AI safety research. While implementation presents certain administrative and operational challenges—explored in subsequent sections—none proves insurmountable, suggesting the framework's viability as a governance mechanism for promoting responsible AI development. As the discussion transitions to examining precise targeting mechanisms for safe AI practices, it becomes crucial to consider how fiscal incentives can be calibrated to advance genuine safety objectives rather than merely accelerating technological

---

[206] *See* Tamir Togoontumur & N. S. Cooray, *Does Collaboration Matter: The Effect of University-Industry R&D Collaboration on Economic Growth*, 15 J. KNOWL. ECON. 9482 (2023) (finding that university and industry collaborations on R&D have a strong positive effect on economic growth); Maria Cohen, Gabriela Fernandes & Pedro Godinho, *Measuring the Impacts of University-Industry R&D Collaborations: A Systematic Literature Review*, J. Tech. Transfer (2024) (finding in a meta-review various impact channels of collaborations)

[207] *See, e.g.,* Richard Ray, and Nicholas C. Lynch, *Qualifying Expenses for the Expanded Research and Development Credit*, CPA J. (Nov. 18, 2019), https://perma.cc/AFG2-WTZP (discussing limitations of the R&D tax credit, including the exclusion of routine data collection, quality control testing, and activities not directly tied to technological principles).

[208] *See* Treas. Reg. § 1.41-4A(d)(2) (stating that taxpayers must retain substantial rights to research results to qualify for the R&D tax credit); *See also* Populous Holdings, Inc. v. Commissioner, T.C. Memo 2020-71 (U.S. Tax Ct. 2020) (holding that research funded by a third party where the funder retains exclusive rights does not qualify for the R&D credit).

[209] *See* Sophia Shah and A.J. Schiavone, *A Process of Experimentation: Production Expenses for the Research and Development Tax Credit,* 96 TAX ADVISER 112 (Sept. 2024) (discussing the exclusion of routine data collection, quality assurance testing, and administrative costs from qualifying R&D tax credit expenditures).

[210] *See* Internal Revenue Service, *Audit Techniques Guide: Credit for Increasing Research Activities § 41—Qualified Research Expenses,* IRS.Gov, https://perma.cc/VF5N-UX3M (explaining that capital expenditures for fixed assets, such as servers or lab spaces, and costs for training personnel or implementing safety measures without experimentation are not considered qualified research expenses).

[211] *See* Internal Revenue Service, *Instructions for Form 6765*, IRS.gov, https://perma.cc/TU7M-E8U9 (outlining forms for activities excluded from the R&D tax credit, including non-technological activities like market research and research conducted outside the United States).

capabilities, thereby addressing the fundamental challenge of closing the capability-safety gap in AI development.

2. Spurring Consumer Demand for Safe & Reliable AI Products

On the individual consumer and household side, a tax credit could be created for purchasing AI products certified as reliable and safe, similar to the existing Energy Efficient Home Improvement Credit.[212] The new "AI Reliability Credit" would incentivize producers to certify, and consumers to invest in, AI technologies that meet rigorous safety and reliability standards, such as mitigating bias, protecting user data, or operating transparently. This fiscal apparatus will provide a credit equal to a 30% of the costs of "qualified AI reliable products" like smart home devices, AI-powered personal assistants, or other consumer-facing AI applications that have been certified by approved regulatory or independent organizations for meeting AI reliability and safety benchmarks.

The Energy Efficient Home Improvement Credit allows manufacturers to certify their products for the energy-efficient tax credit by providing a Manufacturer's Certification Statement confirming their product meets IRS requirements.[213] The mechanism for an AI Reliability Credit would mirror this process. Domestic AI producers will provide certification for their eligible products and label them with an "AI Reliability" mark to inform consumers of their qualification for the credit. At the time of purchase, individuals would retain documentation such as receipts and certification details to claim the credit on their annual tax returns and in case of IRS audit.[214] The credit amount could be a percentage of the purchase price, capped at a reasonable maximum per product or household produced in the U.S.. For example, a $3,000 credit cap for an AI-certified product could encourage adoption without significantly impacting the federal budget.[215] By making AI reliability an appealing and salient factor for consumers, the AI Reliability Credit would not only promote safer AI products locally but also drive international market competition toward higher standards of safety.[216]

These proposed measures will be tied to the certifications mentioned earlier from independent organizations in collaboration with regulatory bodies that establish clear and robust standards for AI safety. For example, customized AI products designed while investing in way to mitigate bias, prevent misuse, or enhance transparency could qualify for the tax incentive if they meet the

---

[212] *See* 26 U.S.C. § 25C (The Energy Efficient Home Improvement Credit allows taxpayers to claim a credit equal to 30% of the costs for qualified energy-efficient improvements made to their principal residence.).

[213] *See* Internal Revenue Service, *Treasury and IRS Issue Guidance for the Energy Efficient Home Improvement Credit*, IRS.GOV (Oct. 24, 2024) https://perma.cc/2BSK-ZFKY (detailing requirements for taxpayers to claim the credit, including retaining receipts, certification details, and, starting in 2025, product identification numbers for qualified items).

[214] *Id.* (noting consumers keep such certifications their records in case of an IRS audit. Certification standards are typically guided by the U.S. Department of Energy or Environmental Protection Agency).

[215] *Id.*

[216] A number of studies on certifications and consumer demand find that credible certification mechanisms spur demand. Mario F. Teisl *et al.*, *Can Eco-Labels Tune a Market? Evidence from Dolphin-Safe Labeling*, 43 J. ENVTL. ECON. & MGMT. 339, (2002) (demonstrating labels shift consumer behavior); Giovanna Piracci et al., *On the Willingness to Pay for Food Sustainability Labelling: A Meta-Analysis*, 55 AGRIC. ECON. 329, 340 (2024) (concluding, from a meta review, that consumers are willing to pay 29% more for sustainability labels on food, but warning that this masks significant variance).

certification requirements.[217] By rewarding companies that invest in safety and ethical AI practices, this policy could drive widespread adoption of responsible AI practices across industries.[218] Thus, through shifting consumer demand, this measure could align private firm incentives with broader societal goals of ensuring AI systems are safe, reliable, and beneficial to users. As consumer demand for safe and reliable AI products grows, it becomes equally important to consider implementing Pigouvian levers for unsafe AI development and practices to ensure accountability and deter harmful behaviors.

3. Penalizing Unsafe AI Development

Parties that engage in unsafe behavior sometimes externalize the risk to other parties.[219] A standard solution, due to Pigou, is the use of corrective tax measures in situations where markets do not satisfactorily resolve these issues. [220] Corrective taxes are meant to impose the external cost on those who engage in unsafe behaviors, and thus "internalize" the harm they create.[221] Unlike the command-and-control approach, penalties grants the regulated party the freedom to decide if, how much, and how to participate in the regulated activity.[222] For example, in the case of pollution-causing actions there is a uniform rate or fees are imposed on polluters for them to internalize their actions .[223]

Legal scholars have shown the utility of corrective taxes to encourage safety. Cass Sunstein, for example, has proposed implementing a tax on hazardous workplaces within the framework of workplace safety to enhance the effectiveness of OSHA standards.[224] Similarly, Jonathan Masur and Eric Posner propose imposing taxes as the optimal method for addressing negative externalities, such

---

[217] On plausible market consequences, *see* supra note 216.

[218] *See supra* note 216

[219] This is the basic tort model of accident risk, see Steven Shavell, *Liability for Accidents*, in 1 HANDBOOK OF LAW AND ECONOMICS 139 (A. Mitchell Polinsky & Steven Shavell eds., 2007)

[220] *See* William J. Baumol, *On Taxation and the Control of Externalities*, 62 AM. ECON. REV. 307, 309-11 (1972) (developing modern framework for Pigouvian taxation). *But see* Victor Fleischer, *Curb Your Enthusiasm for Pigovian Taxes,* 68 VAND. L. REV. 1673, 1674 (2015) (arguing that "Pigouvian taxes are likely to be the optimal regulatory instrument only when (1) the harm is (or is properly analogized to) global pollution, and where the harm does not vary significantly based on the source, or (2) the variation in marginal social cost is easily observed and categorized, as with traffic congestion charges.").

[221] *See* Ottmar Edenhofer, Max Franks & Matthias Kalkuhl, *Pigou in the 21st Century: A Tribute on the Occasion of the 100th Anniversary of the Publication of The Economics of Welfare*, 28 INT'L TAX PUB. FIN. 1090, 1092 (2021) (viewing Pigou's legacy of corrective taxes "in the fundamental concepts of externalities and their correction... which are taught in elementary economics courses.").

[222] *See* Omri Ben-Shahar and Kyle D. Logue, *Outsourcing Regulation: How Insurance Reduces Moral Hazard*, 111 MICH. L. REV. 197, 232 (2012) (pointing to insurance and taxes as interchange forms of regulatory measures choice of response).

[223] *But See* Forastiere Francesco, Hans Orru, Michal Krzyzanowski & Joseph V. Spadaro, *The Last Decade of Air Pollution Epidemiology and the Challenges of Quantitative Risk Assessment*, 23 ENV'L HEALTH 98 (2024), available at https://perma.cc/GF93-69HY (arguing that such uniform standards without considering source-specific risk levels lead to inefficiencies in policy implementation).

[224] *See* Cass R. Sunstein, *Administrative Substance*, 1991 DUKE L.J. 607, 640 (1991) ("A tax on employers for maintaining dangerous conditions, greater reliance on workers' compensation and on disclosure of risks to workers, and more active bargaining and employee involvement in the process of monitoring workplace safety, are all promising techniques.").

as pollution.[225] They argue that this taxation approach surpasses command-and-control regulations, which can be rigid and inefficient, and trading systems, which may face implementation challenges and market failures.[226] This framework not only promotes safety but also fosters innovation as businesses seek cost-effective ways to reduce their tax burden by adopting safer, cleaner practices.[227]

Building on these theoretical foundations, we propose implementing corrective taxes in the AI development context through a comprehensive penalty framework. The tax system would impose graduated penalties on firms that develop or deploy AI systems later determined to pose significant public safety risks or violate established ethical standards. These penalties would operate through two primary mechanisms: direct tax surcharges and the recapture of previously granted tax benefits. For instance, companies deploying AI systems that demonstrate significant preventable risks—such as leading to critical infrastructure failure or misalignment that leads to large loss of life—would face both immediate tax penalties and the potential recapture of prior R&D credits and expensing benefits.[228] This dual approach ensures that firms internalize the full social cost of unsafe development practices while creating strong *ex ante* incentives for responsible innovation.[229] This measure not only holds companies accountable for prioritizing safety but also ensures that public funds are not inadvertently subsidizing harmful AI practices. By linking tax benefits directly to past compliance with safety regulations, policymakers can create a strong financial disincentive for neglecting safety in AI development, reinforcing the importance of responsible innovation. Monies collected from such tax penalties will help reverse negative spillovers of unsafe AI practice by supporting research and development of AI safety standards.[230]

The tax system has demonstrated significant capability in designing and implementing effective penalties through established administrative frameworks. [231] In the oil and gas industry, for instance,

---

[225] Jonathan S. Masur & Eric A. Posner, *Toward a Pigouvian State*, 164 U. PA. L. REV. 93, 95 (2015) (arguing that Pigouvian taxes are superior to either command-and-control or trading systems and that regulators with authority to impose Pigouvian taxes should undertake that measure). *See also* Shawn E. Fields, *Regulatory Trading*, 90 U. CHI. L. REV. 1095, 1096 (2023) (examining the uses of trading in environmental and natural resources law and that environmental problems tend to have larger costs and benefits making them more worthwhile to incur the costs of a trading regime).

[226] Masur & Posner, *supra* note 225 at 99.

[227] *See* Inga Hardeck, Kerry K. Inger, Rebekah D. Moore & Johannes Schneider, *The Impact of Tax Avoidance and Environmental Performance on Tax Disclosure in CSR Reports,* 46 J. AM. TAX'N ASS'N, 83 (2024) (exploring the relationship between tax avoidance, environmental performance, and corporate social responsibility).

[228] For discussion of similar recapture provisions in other contexts, *see* Daniel N. Shaviro, *Selective Limitations on Tax Benefits*, 56 U. CHI. L. REV. 1189, 1213-15 (1989) (analyzing recapture mechanisms and targeted restrictions on tax advantages, such as deductions and credits, to curb tax avoidance). *But see* James Hurley, *Tax Authority Accused of Abusing its Power by Cancelling R&D credit claims,* TIMES (Aug. 19, 2024), https://perma.cc/Z39Y-YKV3 (highlighting concerns regarding the revocation of R&D tax credits without proper inquiry).

[229] This approach builds on established economic theory regarding optimal deterrence. *See* Gary S. Becker, *Crime and Punishment: An Economic Approach*, 76 J. POL. ECON. 169, 180-85 (1968) (claiming that adjusting economic incentives, such as fines or taxes, can influence crime rates by altering the cost-benefit analysis of potential offenders).

[230] *See supra* section III.1.

[231] *See, e.g.,* 26 U.S.C. § 831(b) (allowing small insurance companies to elect to be taxed only on their investment income, rather than their underwriting income, but non-compliance with its regulations can lead to the denial of tax benefits); The Tax Adviser, *IRS Penalties, Abatements, and Other Relief* (2024), https://perma.cc/99SS-3GTM (discussing generally IRS penalties for non-compliance with tax laws, including revocation of tax benefits).

deductions for intangible drilling costs may be challenged or revoked if expenses are mischaracterized or if companies violate environmental regulations.[232] Similarly, developers receiving Low-Income Housing Tax Credits must adhere to affordable housing requirements; failure to meet these obligations, such as maintaining affordability or safety standards, results in credit recapture.[233] Employment-related incentives, including the Work Opportunity Tax Credit, can be revoked when employers violate labor laws or engage in discriminatory practices. [234] Renewable energy and environmental tax credits provide another instructive example, as these benefits remain contingent on meeting specific compliance milestones.[235]

Building on these established frameworks, we propose a two-pronged approach combining ex ante investment requirements with ex post enforcement mechanisms. As a preventive measure, firms in the AI industry must allocate a minimum portion of their development research budget (say, 25%) toward safety-focused activities to qualify for tax benefits, including wage deductions and R&D credits. Qualifying safety expenditures would encompass robust testing protocols, adversarial attack mitigation, and alignment validation frameworks. On the enforcement side, firms that experience significant safety failures—such as critical system misalignment or demonstrated harm to users—would face both credit recapture and potential tax penalties calibrated to harm severity.

The implementation challenge lies primarily in validating safety-related expenditures, particularly given information asymmetries between regulators and firms.[236] However, the tax system has demonstrated capacity to perform similar validations in contexts of evolving standards and industry expertise advantages.[237] The key lies in maintaining relatively permissive standards during initial claims while reserving more stringent scrutiny for post-incident investigations or targeted audits. This approach allows the tax system to develop relevant expertise organically while creating strong incentives for accurate reporting and substantive safety investment.

The tax system's experience with corporate tax avoidance offers instructive parallels for AI safety enforcement.[238] Section 357's treatment of liability-laden property transfers demonstrates how tax

---

[232] 26 U.S.C. 263(c); IRS Pub. 5652, *Oil & Gas Audit Technique Guide*, https://perma.cc/WXX8-4EHQ.

[233] *See* National Housing Law Project, *LIHTC Preservation & Compliance*, *National Housing Law Project*, https://perma.cc/R7RM-KEV2 (explaining the recapture of Low-Income Housing Tax Credits for failure to maintain affordability and safety standards).

[234] *See* Andie Kramer, *Energy Tax Credits for a New World Part IX: Overview of Changes to Traditional Tax Equity Financing*, NATIONAL L. REV., Oct. 8, 2024, https://perma.cc/UWD8-5J5V (discussing changes to U.S. energy tax credits, including compliance requirements and recapture provisions under new legislation).

[235] *See, e.g.*, Tracey M. Roberts, *Picking Winners and Losers: A Structural Examination of Tax Subsidies to the Energy Industry*, 41 COLUM. J. ENVTL. L. 63, 94 (2016) (describing the history of renewable energy tax subsidies).

[236] *See* Eric A. Posner, *Controlling Agencies with Cost-Benefit Analysis: A Positive Political Theory Perspective*, 68 U. CHI. L. REV. 1137,1177 (2001) (discussing the various controls Congress and the President has agencies contributing to agency inefficiency).

[237] *See* Sean, McGuire, Thomas C. Omer and Dechun Wang, *Tax Avoidance: Does Tax-Specific Industry Expertise Make a Difference?* 87 ACCOUNT'G REV. 975 980 (2012) (arguing that Tax-specific industry expertise of the external audit firm influences its clients' level of tax avoidance.).

[238] *See generally* Richard G. Greiner, Paul L. Behling, and J. Denny Moffett, *Assumption of Liabilities and the Improper Purpose—A Re-examination of Section 357 (b),* 32 TAX LAW. 111, 113 (1978) (discussing the implications of Section 357(b) in corporate tax avoidance*);* Alissa Bruehne & Martin Jacob, *Corporate Tax Avoidance and the Real Effects of Taxation: A*

law can effectively address sophisticated avoidance strategies through a combination of clear statutory triggers and flexible administrative standards.[239] Specifically, Section 357(b)'s principal purpose test and Section 357(c)'s quantitative thresholds create a framework that both deters abuse and provides clear guidance for compliance.[240] This model of combining bright-line rules with standards-based oversight could be particularly valuable for AI safety regulation, where technical complexity and rapid innovation require similar flexibility.[241]

The preceding analysis demonstrates how tax policy can serve as a dynamic governance mechanism for emerging technologies. By combining ex ante investment requirements with calibrated enforcement measures, the tax system can help bridge the critical gap between private incentives and public safety imperatives in AI development. This approach leverages existing administrative competencies while creating new frameworks for safety validation and compliance monitoring.

E. The Case for Fiscal Levers

Tax policy presents distinctive advantages for promoting AI safety through its capacity to harness existing institutional frameworks while preserving market dynamics. [242] Unlike traditional command-and-control regulation, which can impose rigid constraints and potentially stifle innovation, tax-based mechanisms offer organizations flexibility in achieving their objectives while aligning private incentives with public welfare.[243]

The tax system's effectiveness in promoting AI safety operates through three primary mechanisms: cultural transformation, expertise mobilization, and equitable distribution of benefits. First, tax incentives fundamentally reshape organizational culture by making safety investments financially advantageous. [244] When firms receive tax credits for safety-focused research, compliance

---

Review, SSRN (Jan. 7, 2020), https://dx.doi.org/10.2kramer139/ssrn.3495496 (This paper synthesizes empirical research on corporate tax avoidance, exploring its effects on corporate behavior and tax compliance).

[239] 26 U.S.C. § 357(b) (stating liabilities assumed in a transaction with the principal purpose of tax avoidance are treated as taxable boot, negating the tax-free treatment of the exchange). *See* Karen C. Burke, *Contributions, Distributions, and Assumption of Liabilities: Confronting Economic Reality*, 56 TAX LAW. 383, 385 (2003) (examining how Section 357(b) prevents tax avoidance by treating liabilities assumed in transactions with a principal purpose of tax avoidance as taxable boot).

[240] *See* Boris I. Bittker, *The Corporation and the Federal Income Tax: Transfers to a Controlled Corporation*, 1959 WASH. U. L. Q. 1, 15 (February 1959) (describing the legislative history of section 357(b)).

[241] *See* Richard G. Greiner Paul L. Behling, and J. Denny Moffett, *Assumption of Liabilities and The Improper Purpose—An Examination of Section 357(b*), 32 TAX LAWYER 111 (1978) (discussing the mechanisms and complexities of section 357(b)).

[242] *See* Alan J., Auerbach & James R.Hines, *Taxation and Economic Efficiency*, 47 J. ECON. LIT. 1252 (2009) (analyzing how different tax policies can affect resource allocation, economic behavior, and overall societal welfare).

[243] *See e.g.,* Antoine Dechezleprêtre et al., *Do Tax Incentives Increase Firm Innovation? An RD Design for R&D, Patents, and Spillovers*, 15 AM. ECON. J. ECON. POL'Y 87 (2023) (providing causal evidence that R&D tax incentives positively impact firm innovation and generate spillover benefits for related firms).

[244] *See, e.g.,* Jacob Nussim and Anat Sorek, *Theorizing Tax Incentives for Innovation, 36* VA. TAX REV. 25, (2017) ("[Cash based transfers] may also facilitate knowledge sharing early on, and may prevent innovation races."); Jonathan Remy Nash, *Taxes and the Success of Non-Tax Market-Based Environmental Regulatory Regimes* in CRITICAL ISSUES IN ENVIRONMENTAL TAXATION 733, 735 (Chalifour et al. eds., 2008) (arguing that tax concerns and tax structures can have significant effects upon the function and ultimate success of market-based environmental regulatory regimes).

protocols, and governance frameworks, they internalize these priorities at both operational and strategic levels. This internalization extends beyond immediate compliance, fostering industry-wide safety consortia and creating positive network effects through initiatives like ethics training for AI developers and dedicated safety teams.

Second, tax incentives effectively mobilize private sector expertise while preserving competitive dynamics.[245] The framework harnesses organizational expertise through multiple channels: tax credits reward safety research investments, encourage collaboration with non-profit organizations and academic institutions, and create natural partnerships where private companies maintain ownership of safety initiatives while benefiting from public support. [246] This proves particularly crucial in the global AI race, where maintaining U.S. competitiveness requires careful calibration of safety requirements against development imperatives.

Third, the tax system's distributive function treats AI safety as a public good, allocating costs across taxpayers while concentrating benefits in safety-enhancing research and development. [247] This approach proves especially valuable for fundamental safety research that may lack immediate commercial appeal but provides critical societal benefits.[248] Complementing these positive incentives, tax penalties serve as crucial corrective mechanisms –when firms face tax consequences for safety failures, they naturally allocate greater resources toward risk mitigation and safety protocols.

Knowledge sharing represents another crucial advantage of tax-based safety promotion. Drawing from successful models like Information Sharing and Analysis Centers in cybersecurity, [249]

---

[245] *See, e.g.,* Mark A. Cohen and Paul H. Rubin, *Private Enforcement of Public Policy*, 3 YALE J. ON REG. 167, 187 (1985) (discussing the idea of safety regulation as socially efficient tool).

[246] *See, e.g.*, Daniel, Bradley, Connie X. Mao, and Chi Zhang, *Do Corporate Taxes Affect Employee Welfare? Evidence from Workplace Safety,* 42 J. ACCOUN'G PUB. POL'Y 42 1413, 1415 (2023) (finding workplace safety decreases when firms experience a tax increase).

[247] *See, e.g*. Richard H. Pildes & Cass R. Sunstein, *Reinventing the Regulatory State*, 62 U. CHI. L. REV. 1, 101 (1995) (discussing the distributive effects of tax incentives accompanied by efforts to diminish effects on the poor); Peter Mieszkowski, *Tax Incidence Theory: The Effects of Taxes on the Distribution of Income,* 7 J. ECON. LITE. 1103, 1103 (1969) (analyzing the tax incidence is the investigation of the distributive effects of taxes).

[248] *See, e.g.,* Diego d'Andria & Ivan Savin, *A Win-Win-Win? Motivating Innovation in a Knowledge Economy with Tax Incentives,* 127 TECH. FORECASTING & SOC. CHANGE 38, 38–56 (2018) (analyzing the effectiveness of tax incentives in fostering innovation within knowledge-based economies and exploring their broader economic impacts); Michael Keen & Jenny E. Ligthart, *Information Sharing and International Taxation: A Primer,* 13 INT'L TAX & PUB. FIN. 81, 81–103 (2006) (discussing the role of taxpayer-specific information exchange between national tax authorities in enhancing transparency and accountability). *But see* Mckay Jensen, Nicholas Emery-Xu, Robert Trager, *Industrial Policy for Advanced AI: Compute Pricing and the Safety Tax*, ARXIV 2 (Feb. 22, 2023), https://perma.cc/48YL-PDHX (defining "safety tax" as the marginal cost of deploying an AI system that is aligned with human values compared to an equivalent but unaligned system, representing the tradeoff between safety investments and performance-driven incentives.).

[249] *See, e.g.,* Maryland's Buy Maryland Cybersecurity Tax Credit provides businesses with a tax credit of up to 50% of the net purchase price of cybersecurity technologies and services from qualified Maryland-based providers, capped at $50,000 per tax year, to encourage investment in cybersecurity measures within the state. Md. Dep't of Com**.,** Buy Maryland Cybersecurity

tax incentives can foster transparency and collaboration through rewards for adopting shared safety standards or participating in third-party certifications.[250] These public-private partnerships enable corporations to share insights from safety research with the broader community while protecting proprietary information.[251] By rewarding participation in pre-competitive research alliances and contributions to open-source safety tools, the framework creates positive spillover effects that benefit the entire AI ecosystem.

The framework's effectiveness extends beyond direct safety promotion to fostering sustainable innovation ecosystems. Consumer-side tax credits for certified safe AI products create market demand for responsible development practices, while organizational tax benefits encourage long-term investments in safety infrastructure and expertise. Private corporations, leveraging their domain expertise,[252] can determine the most effective implementation of safety measures without external micromanagement. This flexibility ensures that investments are tailored to the unique challenges of AI safety while providing firms latitude in how they innovate and address risks.[253]

F. The Administrative Challenge

Although current safety tax incentives are widely used to shape behavior across various sectors,[254] their application to AI safety presents distinct challenges that warrant careful consideration. Critics raise several compelling objections that have deeply informed our framework's development. The most immediate concern is political economy: providing tax benefits to an already profitable technology sector may face opposition from both policymakers and the public. [255] Any reduction in tax revenue would need to be offset, either through increased burden on other taxpayers or reduced

---

Tax Credit, https://perma.cc/3WYT-Y4XM. Oklahoma's Software/Cybersecurity Workforce Tax Credit offers a tax credit of up to $2,200 annually for qualifying employees with degrees from ABET-accredited institutions working in software or cybersecurity roles, aiming to attract and retain skilled professionals in the state. Okla. Dep't of Com., Software/Cybersecurity Workforce Tax Credit, https://www.okcommerce.gov.

[250] *See* Siglé, Marie-Léandre, Sjef van Erp, Thomas van Hulten & Maarten Pieter Schinkel, *The Cooperative Approach to Corporate Tax Compliance: An Empirical Assessment*, 48 J. INT'L TAX 1141, 1154 (2022) (analyzing how cooperative compliance programs between tax authorities and corporate taxpayers enhance compliance and foster a culture of collaboration).

[251] *See* Matthew Stepp & Robert D. Atkinson, *Creating a Collaborative R&D Tax Credit,* INFO. TECH & INNOVATION FOUNDA. June 2011, at 1 https://perma.cc/4JZU-K8UV (finding that nations offering more generous R&D tax credits achieve higher rates of university-business collaboration than the United States)

[252] *See supra* note 18 and accompanying text.

[253] *See, e.g.,* Alan J. Auerbach, *Measuring the Effects of Corporate Tax Cuts*, 32 J. ECON. PERSP. 97, 100 (2018) (examining the impact of the Tax Cuts and Jobs Act of 2017 on corporate investment and resource allocation); Dan Andrews & Federico Cingano, *Public Policy and Resource Allocation: Evidence from Firms in OECD Countries*, 29 ECON. POL'Y 253 (2014) (analyzing how public policies, including taxation, influence resource allocation among firms in OECD countries).

[254] *See supra* Part II.

[255] *See e.g.,* Nancy C. Staudt, *The Political Economy of Taxation: A Critical Review of a Classic*, 30 L. & SOC'Y REV. 651 (1996) (reviewing Henry's Simons seminal work on the definition of income but adding that the political economy of tax policy has significant implications for income tax design); Ethan Ilzetzki, *Tax Reform and the Political Economy of the Tax Base*, 42 INT'L REV. POL. ECON. 132, 135 (2021) (analyzing the influence of political institutions on tax base reforms), available at https://perma.cc/HRU3-K3Q8.

public services. [256] Moreover, effective administration of these incentives poses unique oversight challenges that strain institutional capacity. The complexity of tax mechanisms, and tax incentives for R&D specifically, lies in their intricate design, which must balance fostering innovation and compliance while avoiding inefficiencies, misuse, or unintended consequences, and often involves complex audits and administrative processes to ensure proper implementation and oversight.[257] Companies must identify and substantiate expenses such as wages, supplies, and contract research related to eligible research activities, often necessitating the expertise of tax professionals.[258] The need to maintain comprehensive documentation and navigate evolving IRS guidelines adds layers of complexity that may overshadow the credit's potential benefits.[259] Lastly, the lack of salience and internalization of tax benefits at the upper C-suite level often results in missed opportunities to align strategic decision-making with available incentives, as these benefits may be perceived as peripheral or poorly integrated into broader organizational goals.[260] Together, these political, distributional, and administrative concerns demand thoughtful structuring of any proposed incentive scheme.

The political economy challenge, while significant, must be situated within the broader landscape of existing innovation policy and institutional dynamics.[261] As discussed above, firms already access various tax incentives to support their research activities, with most of these benefits flowing to capability development rather than safety research.[262] To put the point somewhat crudely, we are already pouring money on the industry, but perhaps too much on the wrong part of it.

---

[256] *See, e.g.,* Jaeger Nelson and Kerk Phillips, *The Economic Effects of Financing a Large and Permanent Increase in Government Spending* CONG. BUDGET OFFICE 8 (2021), https://perma.cc/XNP4-6DB (analyzing the impact of income tax cuts on the national deficit, concluding that tax cuts lead to higher deficits without sufficient offsetting revenue from economic growth); Paul N. Van de Water, *Tax Reform Must Not Lose Revenues and Should Increase Them*, CENTER ON BUDGET AND POLICY PRIORITIES, https://perma.cc/QR49-EKS3 (arguing that tax cuts should be offset by revenue increases or spending cuts to avoid exacerbating budget deficits and national debt); William G. Gale & Samuel A. Shapiro, *The Effects of Income Tax Changes on Economic Growth*, BROOKINGS INSTITUTION (June 9, 2016), https://perma.cc/VUL2-QZPF ("Reforms that improve incentives, reduce existing distortionary subsidies, avoid windfall gains, and avoid deficit financing will have more auspicious effects on the long-term size of the economy…").

[257] For example, the credit is not available for research funded via government or private grants. *See* 26 U.S.C. § 41(d)(4)(H). Moreover, companies claiming the credit cannot "double dip," thus, they must reduce immediate expensing & the Orphan Drug Credit for the amount of the credit. *See* 26 U.S.C. § 280C(c)(1).

[258] *See* Klein, *supra* note 190 at 1166 (surveying the role of the research credit in income misreporting arguing that due to its complexity large corporations employ full-time professionals to handle their taxes); John Deininger, Jared Boucher & Tom Windram, *Documenting Qualified Research Activities for the Research Tax Credit*, 51 TAX ADVISER 260 (2020) https://perma.cc/R5TX-6ZSZ (discussing the necessity for companies to meticulously document qualified research expenses).

[259] *See* Daniel J. Hemel & Lisa Larrimore Ouellette, *Beyond the Patents-Prizes Debate*, 92 TEX. L. REV. 303, 326 (December 2013) ("Estimates of the effectiveness of the R&D credit vary widely.")

[260] *See* Alex Raskolnikov, *Revealing Choices: Using Taxpayer Choice to Target Tax Enforcement*, 109 COLUM. L. REV. 689, 700 (2009) (referring to income taxes as the top reason for highest risk area of financial reporting and worry among the C-suite).

[261] On the political drivers of R&D policies, see James C. Hearn, T. Austin Lacy & Jarrett B. Warshaw, *State Research and Development Tax Credits: The Historical Emergence of a Distinctive Economic Policy Instrument*, 28 ECON. DEV. Q. 166 (2014).

[262] See *supra* note 13 and accompanying text.

Nonetheless, scholars have raised legitimate concerns about political capture,[263] noting that tax preferences are particularly susceptible to abuse by special interest groups in areas of bipartisan agreement like innovation policy.[264] Because these are reasonable concerns, our framework incorporates several structural safeguards against such risks. First, our proposal is flexible enough to allow policymakers redirect existing incentives toward safety research rather than creating new benefits, addressing both budget neutrality concerns and limiting opportunities for rent-seeking behavior. Second, the framework distributes oversight responsibility across multiple mechanisms, including tax incentives, direct grants, regulatory exemptions, and public-private partnerships, thereby constraining the discretion of any single agency. Third, we propose a fairly constrained mandate for AI safety, as elaborated below. This integrated approach, combining narrow targeting with distributed oversight, simultaneously addresses concerns about political capture while advancing the critical goal of redirecting resources toward systematically underinvested safety research.

A critical implementation challenge lies in precisely identifying which segments of the AI development chain warrant targeted tax incentives. Given AI's pervasive integration across sectors, nearly any firm could plausibly claim qualification for safety-related benefits without meaningful differentiation. We propose focusing initial incentive structures on foundational research and model training activities, rather than downstream applications or product development. This upstream prioritization finds theoretical support in two key dynamics: first, foundational research operates at a greater remove from market pressures that might otherwise drive safety considerations; second, advances in fundamental safety protocols at the architectural level generate positive spillover effects throughout the development ecosystem.[265] This narrower scope helps mitigate some of the political concerns noted above.

A more fundamental challenge emerges in distinguishing genuine safety research from what we term "safety-washing"— superficial or misleading commitments to AI safety that mask continued prioritization of capability advancement.[266] This challenge manifests along two distinct dimensions.[267] First, at the definitional level, we lack precise terminology to differentiate between investments in general AI capabilities and specific safety measures, a distinction further complicated by recent research suggesting positive correlations between model capabilities and certain safety

[263] *See, e.g.,* Zachary Liscow, *Redistribution for Realists*, 107 IOWA L. REV. 495, 524 (2022) ("Policies could look the way they do for many reasons, including political capture.").

[264] *See, e.g.,* Cato Institute, *Special Interests & Corporate Welfare*, in *Cato Handbook for Policymakers* (9th ed. 2022), https://perma.cc/E5YE-29HE (discussing how special-interest groups often secure narrow benefits from the government, leading to policies that may not align with the general public interest);

[265] The three point seat belt is a case in point, invented by Volvo in 1959 and released to the entire market, saving millions of lives. Alexander Stoklosa, *The Three-Point Seatbelt Turns 60, and It's a Damn Hero*, CAR AND DRIVER (Aug. 21, 2019), https://perma.cc/8PAZ-JKZC.

[266] *See* Ren et al., *supra* note 76.

[267] *Id.*

characteristics.[268] Second, we face significant empirical hurdles due to the absence of well-established safety endpoints and regulatory frameworks.[269]

OpenAI's 2023 safety initiatives starkly illustrate these challenges. The high-profile recruitment of Scott Aaronson to their safety team initially generated significant optimism within the AI safety community.[270] Yet this apparent commitment proved largely ceremonial—Aaronson was tasked primarily with developing AI writing detection tools, a narrow technical challenge focused on academic plagiarism that never materialized into meaningful products.[271] This episode exemplifies how leading AI firms can leverage prestigious appointments and safety rhetoric while maintaining their singular focus on capability advancement, effectively using safety initiatives as reputational cover rather than vectors for substantive reform.

Our framework addresses these challenges by building upon the tax authorities' established competencies rather than creating new oversight bodies. While subject-matter agencies typically possess greater technological expertise than tax authorities, [272] and direct funding mechanisms like grants are often considered superior to tax preferences,[273] the IRS offers distinct administrative advantages in three key areas:

First, the Internal Revenue Service has already developed sophisticated protocols for evaluating technical research claims across various complex sectors, including biotechnology and advanced manufacturing.[fn] Under Section 41, tax authorities regularly assess whether activities constitute systematic investigation through experimentation to resolve technical uncertainty.[fn] This existing framework provides a natural foundation for evaluating AI safety research through documented protocols and clear research objectives.

Second, current R&D verification systems rely on contemporaneous documentation requirements that can be readily adapted for AI safety research. [274] Just as firms must maintain

---

[268] *See* Hendrycks & Mazeika, *supra* note 5.

[269] *See* Ren et al., *supra* note 76.

[270] *See e.g.,* peterbarnett, Scott Aaronson is joining OpenAI to work on AI safety, LessWrong (June, 6, 2022) https://perma.cc/M6BW-XADZ. Aaronson himself, in his widely read blog, anticipated this very question, "Should you worry," he asks rhetorically, "that OpenAI is just hiring me to be able to say 'look, we have Scott Aaronson working on the problem,' rather than actually caring about what its safety researchers come up with?", and responds that that while he can't prove that this isn't a concern, he "was impressed by [their] detailed, open-ended engagement . . . sort of like how it might look if they actually believed what they said."). Scott Aaronson, *OpenAI!,* Shtetel Optimized (June, 17th, 2022).

[271] *See* Liron Shapira, *Scott Aaronson Makes Me Think OpenAI's "Safety" Is Fake, Clueless, Reckless and Insane, DoomDebates* https://perma.cc/D3CW-GCGY. Aaronson posted a response on social media, largely agreeing with the critique: "nothing I did at OpenAI and nothing I said on my podcast should make [a person worried about AI x-risk] less terrified (if anything, the contrary)." Liron Shapira, https://perma.cc/6PG8-PXSZ.

[272] For example, The National Institute of Standards and Technology has hired Paul Christiano as its head of AI Safety, a highly respected safety researcher and a former OpenAI engineer https://perma.cc/U4RX-TX9W.

[273] *See, e.g.,* Jacob Nussim & David A. Weisbach, *The Integration of Tax and Spending Programs*, 113 YALE L.J. 955, 1023 (2004) (discussing the inefficiencies and deadweight loss associated with using the tax system for regulatory purposes).

[274] *See, e.g.*, Jeff Drew, *New R&D Credit Documentation Requirements Clarified*, 233 J. ACCT. 1, 12 (Jan. 25, 2022), https://perma.cc/5A6L-59Z4 (This article discusses updated IRS documentation requirements for the R&D tax credit, highlighting the importance of maintaining adequate records to support claims).

detailed records for traditional R&D credits,[275] AI developers would document specific safety activities including alignment testing protocols, robustness evaluations against adversarial attacks, red-teaming exercises and outcomes, safety-relevant model behaviors during training, and systematic investigation of failure modes and mitigation strategies.

Third, the evaluation framework can leverage emerging industry standards and technical benchmarks to create concrete qualification metrics.[276] These would include implementation of specific safety protocols (such as adversarial training and bounded optimization), achievement of quantifiable safety benchmarks (like robustness scores and alignment metrics), development of safety monitoring systems, contributions to open-source safety tools, and systematic documentation of model behaviors and interventions.[277] By adhering to predetermined AI safety priorities established by technical committees, the framework maintains administrative efficiency while ensuring tax authorities are not overburdened with complex technical determinations.[278]

To further streamline implementation, we propose several administrative simplifications: standardizing eligibility criteria and reliability definitions, adopting pre-certification systems for upfront verification, automating documentation through digital platforms, and shifting toward output-based incentives tied to measurable outcomes rather than solely input-based claims.[279]

---

[275] *See, e.g.,* IRS, *Audit Techniques Guide: Credit for Increasing Research Activities*, IRS.Gov (June 2022), https://perma.cc/E54P-UXZ3 (providing detailed guidance on evaluating claims for the R&D tax credit).

[276] On developing safety protocols, see *supra* note 20.

[277] There are a number of evolving standards and regulatory measures on AI safety. They include ISO/IEC 23894:2023, *Artificial Intelligence – Guidance on Risk Management,* INT'L ORG. FOR STANDARDIZATION 2023, available at https://perma.cc/B6PW-N2PX; ISO/IEC JTC 1/SC 42 Standards on AI Trustworthiness, addressing robustness and functional safety, *see* ISO/IEC JTC 1/SC 42, *Artificial Intelligence,* INT'L ORG. FOR STANDARDIZATION, available at https://perma.cc/3N84-VAQX; ISO/IEC 42001:2023, an AI management system standard that establishes governance frameworks ISO/IEC 42001:2023, *Artificial Intelligence – Management System,* INT'L ORG. FOR STANDARDIZATION 2023, discussed in *How the ISO and IEC are Developing International Standards for the Responsible Adoption of AI*, UNESCO, available at https://perma.cc/92TP-MVXT; IEEE Ethically Aligned Design and 7000-series standards, which set technical benchmarks for AI transparency, bias mitigation, and fail-safe mechanisms. *See* IEEE, *Ethically Aligned Design* and IEEE 7000-series Standards, available at https://perma.cc/2ZWF-M4WA; NIST AI Risk Management Framework 1.0 (2023), a structured framework to assess and mitigate AI-related risks, NAT'L INST. OF STANDARDS & TECH., *AI Risk Management Framework 1.0* (2023), available at https://perma.cc/DQ3H-HV8T; OECD Principles on AI (2019), setting international AI governance norms, ORG. FOR ECON. CO-OPERATION & DEV., *OECD Principles on Artificial Intelligence* (2019), available at https://perma.cc/K8VA-QC7W; EU AI Act (2024), a binding regulatory framework that mandates safety, robustness, and transparency for AI systems (Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence, COM (2021) 206 final (Apr. 21, 2021), available at https://perma.cc/X7Q3-FXH9; and the Partnership on AI's guidelines, which provide industry-driven best practices for responsible AI deployment. *See* Partnership on AI, *Our Work*, available at https://partnershiponai.org/our-work/.

[278] *See* Atherine M. Sharkey and Kevin M. K. Fodouop, *AI and The Regulatory Paradigm Shift at the FDA*, 72 DUKE L.J. ONLINE 1, 106 (2022) (claiming the IRS developed in-house technical expertise to automate dynamic regulatory tasks, demonstrating the importance of integrating technical and regulatory subject matter expertise for effective AI implementation.).

[279] *See e.g.,* Joshua D. Blank and Leigh Osofsky, *Democratizing Administrative Law*, 73 DUKE L.J. 1615, 1620(2024) (proposing to increase the democracy deficit in tax administration by applying administrative law principles on agency communications with the general public similarly to interactions between agencies and sophisticated parties).

Regular program reviews would ensure these incentives remain effective and manageable without unnecessary administrative burdens.[280]

Overall, implementation of incentives to cutting edge research presents some inescapable difficulties. We believe that the proposal offers a realistic path of administration that, while sensitive to difficulties, offers a meaningful and practical path forward. If we are to encourage safety research and implementation, we must start building towards the appropriate regulatory apparatus. Existing tools within tax authorities offer a promising way forward that is—relatively speaking—less demanding in terms of subject matter expertise than alternative oversight proposals.

Conclusion

This Article has argued that tax policy can serve as a powerful and underutilized tool for promoting AI safety. We began by examining the growing capability-safety gap in AI development—where advances in AI capabilities have rapidly outpaced our ability to ensure these systems operate safely and reliably. At the heart of this gap lies what we termed the social misalignment problem: while the rewards from powerful AI systems accrue privately to their developers, the risks and potential harms are broadly socialized across society.

Drawing on extensive precedents from energy efficiency, workplace safety, and environmental protection, we demonstrated how tax policy has historically helped resolve similar misalignment challenges. Our framework proposes three interlocking mechanisms: producer-side incentives that reward safety research and development, consumer-side credits that stimulate demand for certified safe AI products, and corrective tax penalties that internalize the social costs of unsafe development practices. Rather than creating entirely new administrative structures, this approach leverages the tax system's existing competencies in monitoring research activities and verifying compliance.

The framework's utility extends well beyond AI safety. [281] It offers a blueprint for using fiscal policy to address market failures in emerging technologies where private incentives diverge from public welfare. Whether in biotechnology, nanotechnology, or other domains where innovation

---

[280] *See, e.g.,* Yifat Aran*, Making Disclosure Work for Start-up Employees,* 2019 COLUM. BUS. L. REV. 867, 931 (2019) (calming certain disclosure documents impose a high financial and administrative burden on startups); Pontus, Braunerhjelm, Johan E. Eklund, and Per Thulin, *Taxes, the Tax Administrative Burden, and the Propensity for Entrepreneurship,* 56 SMALL BUS. ECON. 681, 690 (2019) (finding that high tax administrative burdens negatively impact entrepreneurial activities, particularly in the early stages of a business.).

[281] *See, e.g.,* Murray Petrie, Richard Allen, *The Crucial Role of Fiscal Policy in Averting Environmental Catastrophe* IMF PFM BLOG (Dec. 6, 2021) https://perma.cc/2FZN-C8CP (discussing how green fiscal policies, such as green taxes and public spending, can contribute to environmental sustainability and support fiscal strategies for sustainable goals); Chris Brown, *Manage Cybersecurity as Part of the ESG Strategy*, DIRECTORS & BOARDS (Jan. 19, 2024) https://perma.cc/JL8D-N3V2 (emphasizing the role of integrating cybersecurity into Environmental, Social, and Governance (ESG) frameworks to protect infrastructure and maintain public trust); OECD, *Fiscal Sustainability of Health Systems* (2021), https://www.oecd.org/en/publications/fiscal-sustainability-of-health-systems_880f3195-en.html (highlighting the importance of robust fiscal policies in strengthening health system resilience); World Economic Forum, *How Fiscal Policy Can Help Save Forests* (2021), https://perma.cc/UQ3A-GECA (exploring how fiscal reforms can influence forest conservation and ecosystem health by incentivizing sustainable land use and reducing deforestation).

carries both tremendous promise and significant risk, carefully calibrated tax incentives can help align private sector behavior with social imperatives. [282] By making safety investments financially advantageous while penalizing reckless development, [283] tax policy can foster cultures of responsible innovation across multiple technological frontiers. [284]

Critics may argue that tax incentives alone cannot guarantee safe AI development. We agree—no single regulatory tool can fully address the complex challenges posed by transformative technologies. However, tax policy offers distinct advantages over traditional command-and-control regulation, particularly in fast-moving technical domains where regulators face significant information and expertise asymmetries. By harnessing market mechanisms and firm-level knowledge while preserving innovation incentives, tax policy can play a crucial role in a broader regulatory ecosystem.

The urgency of addressing AI safety cannot be overstated. As systems grow more capable and autonomous, the stakes of ensuring their reliable and beneficial operation continue to rise. Our framework offers a practical path forward, one that recognizes both the tremendous promise of AI technology and the critical importance of developing it safely and responsibly. Give me a lever and a place to stand, Archimedes said, and I can move the world.

---

[282] *See* Mirit Eyal-Cohen & Ana Santos Rutschman, *Promoting Vaccine Innovation*, 82 OHIO ST. L. J. 1003, 1029 (2022) (discussing crowing out socially valuable goods such as pandemic preparedness).

[283] *See* Christos Makridis, Christos Makridis, Anne Boustead, and Scott Shackelford, *Navigating the cybersecurity labyrinth: Defining 'reasonable' standards for businesses*, BROOKINGS INST. (Feb. 22, 2024) https://perma.cc/D3EC-4W95 (discussing the challenges businesses face in implementing "reasonable" cybersecurity standards, and the potential for tax incentives or regulatory clarity to encourage better security practices).

[284] *See, e.g.*, Matthew Wilson*, Government Market Power and Public Goods Provision in a Federation,* 28 INT'L TAX & PUB. FIN. 1234, 1234 (2020) (examining the impact of centralization and decentralization on public goods provision).