Evidence to the Science, Innovation, and Technology Committee

submitted by Ruchika Joshi

Governing AI based on Enterprise Revenue

Based on independent primary research, the following evidence submission by Ruchika Joshi encourages policymakers to achieve a socially desirable level of AI safety by influencing revenues of AI enterprises to be more responsive to safety efforts exerted by those enterprises.

Using a theoretical economic model combined with insights from expert interviews, this submission provides detailed recommendations under six categories of policy action:

1. Implement a penalty framework highly responsive to changes in observed AI failures
2. Establish an attainable and adaptive "no-fine safety level" i.e. the level of safety efforts by AI enterprises deemed by regulators as sufficient to impose no fine
3. Increase end-user awareness of safety efforts by AI enterprises
4. Support AI enterprises in mitigating risks associated with stochastic failures
5. Encourage revenue sharing and cross-subsidization between AI developers and deployers
6. Reduce costs of safety efforts by AI enterprises

1. About the Author

1.1. Ruchika Joshi specializes in AI safety and governance at the Harvard Kennedy School. Her research areas include AI explainability, distributing AI safety accountability, incentive-based AI regulation, and human oversight of AI systems. Ruchika has extensively worked with government, business, and civil society leaders across Asia and Africa to deploy data and technology for increased impact in health, education, employment, and public service delivery.

2. Background

2.1. Regulating artificial intelligence systems (AIs) has been the subject of extensive policy discussions. At its heart, a key tension policymakers face is how to reap the benefits of AI while minimizing safety risks, as articulated in the Bletchley Declaration[1] for AI safety agreed upon by 28 countries[2].

2.2. Given that AI has been argued to pose large-scale societal harms[3] some stakeholders argue that it should be more regulated. Others are concerned that excessive AI regulation may curtail firms' business incentives for innovation. Against this backdrop of diverging opinions on AI regulation, policymakers

[1] UK Government. (2023, November 2). Chair's Summary of the AI Safety Summit 2023, Bletchley Park. Retrieved from https://www.gov.uk/government/publications/ai-safety-summit-2023-chairs-statement-2-november/chairs-summary-of-the-ai-safety-summit-2023-bletchley-park

[2]UK Government. (2023, November 1).The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023. Retrieved from https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

[3] Critch, A., & Russell, S. (2023). Tasra: A taxonomy and analysis of societal-scale risks from ai. arXiv preprint arXiv:2306.06924.

continue to grapple with the right balance of regulating AI such that AI is safe for society while ensuring AI enterprises are incentivized to innovate.

2.3. Based on the current AI landscape, one way to navigate this perceived tension between AI safety and innovation is to **leverage the observation that the incentive for AI enterprises to invest in safety is not necessarily at odds with their incentive to innovate.** Examining the emerging business models around these AI systems indicates that there is likely a market incentive for AI enterprises to innovate *with* safety.

2.4. Policymakers could develop AI regulation that **strengthens AI firms' existing business incentives for safety efforts, and align the profit-maximizing levels of safety efforts with what is considered socially desirable by policymakers.**

2.5. In addition to aligning safety incentives with innovation, such an approach to regulation also mitigates challenges specific to AI technology such as limits to explainability, AI's inherent unknowability and unpredictability, and the need for specialized technical expertise.

2.6. In this submission, AI safety efforts refer to efforts to prevent and mitigate harms from AI at a societal scale[4]. However, it excludes the use of AI categorized as unacceptable risk in the EU AI Act or dangerous uses of frontier AI. Such AI models, including aspects of artificial general intelligence (AGI) and their applications, would likely need to be regulated within a separate framework.

2.7. Additionally, the framework proposed in this submission should be considered as *one* of many regulatory approaches to test and iterate over as new evidence

[4] Critch, A., & Russell, S. (2023). Tasra: A taxonomy and analysis of societal-scale risks from ai. arXiv preprint arXiv:2306.06924.

emerges. Both AI regulation and research on AI regulation are in nascent stages with a lack of clarity on the right path forward. Much of what constitutes appropriate regulation will depend on how the technology and society co-evolve as AI use becomes more widespread.

2.8. Since AI is a fast-changing technology, policymakers must consider the proposed approach as one strategic tool to employ in a multipronged approach to policy development. As new evidence emerges, they may respond by intensifying or retracting from any given approach within their policy toolkit[5].

3. Regulating AI based on business incentive for safety

3.1. This submission focuses on two main stakeholders in the AI market: AI developers (enterprises that train and create AI) and AI deployers (enterprises that use AI)[6]. It further assumes a simplified market relationship where AI developers sell the use of AI models to AI deployers who in turn sell custom AI applications to end-users. An end-user refers to an individual human or business that the deployed system ultimately affects[7].

3.2. This is in line with emerging business practices. Leading foundation models developed by companies like Open AI[8], Anthropic [9], Google[10], and Cohere[11] are available on API for deployers to purchase, who then are expected to sell custom

[5] Based on expert interviews.
[6] Our definitions of AI developer and deployer borrow from the latest version of the draft EU AI Act.
[7] Brown, I. (2023, June 29). Expert explainer: Allocating accountability in AI supply chains. Ada Lovelace Institute. Retrieved from https://www.adalovelaceinstitute.org/resource/ai-supply-chains/
[8] OpenAI. (2023, March 1). Introducing ChatGPT and Whisper APIs. OpenAI. Retrieved from https://openai.com/blog/introducing-chatgpt-and-whisper-apis/
[9] Anthropic. Build with Claude. Retrieved from https://www.anthropic.com/api#pricing
[10] Google. Priced to help you bring your app to the world. Retrieved from https://ai.google.dev/pricing
[11] Cohere. Scalable, affordable pricing. Retrieved from https://cohere.com/pricing

applications of these models such as specialized AI assistants, coding, and DevOps workflows, and drug discovery and education software to end-users[12].

3.3. For AI deployers to purchase AI models and sell custom AI applications built on these models to end-users, some baseline level of safety would likely need to be assured by AI developers[13]. Assuming greater safety efforts by AI developers reduce the rate of AI application failures, deployers would need to credibly believe that AI developers have invested some minimum level of safety effort into training and creating AI. Without such assurance, deployers risk exposing themselves to AI application failures at the end-user level.

3.4. Such failures can erode user trust crucial to product adoption at scale[14]. If end-users do not trust a deployer's custom AI application, they are unlikely to adopt the application. This dampens the business incentive of AI deployers to build custom AI applications, thereby reducing their demand for the foundational AI models from developers.

3.5. AI developers appear cognizant of such risks to their revenue, which may partly be driving their public efforts to increase AI safety. In 2023, Google, Anthropic, Open AI, and Microsoft launched the Frontier Model Forum to ensure the safety of frontier models[15]. Anthropic, an AI developer that has made its foundational models available for commercial use describes itself as "an AI safety and research

[12] Bloomberg. (2023, June 1). Generative AI to become a $1.3 trillion market by 2032, research finds. Retrieved from https://www.bloomberg.com/company/press/generative-ai-to-become-a-1-3-trillion-market-by-2032-research-finds/

[13] Based on expert interviews.

[14] Bach, T. A., Khan, A., Hallock, H., Beltrão, G., & Sousa, S. (2022). A systematic literature review of user trust in AI-enabled systems: An HCI perspective. International Journal of Human–Computer Interaction, 1-16.

[15] Microsoft. (2023, July 26). Anthropic, Google, Microsoft, OpenAI launch Frontier Model Forum. Retrieved from https://blogs.microsoft.com/on-the-issues/2023/07/26/anthropic-google-microsoft-openai-launch-frontier-model-forum/

company"[16], highlighting its safety focus. More recently, OpenAI offered its customers protection against copyright issues saying "We will now step in and defend our customers, and pay the costs incurred, if you face legal claims around copyright infringement"[17] [18], Adobe[19], Microsoft[20], and Google[21] have offered indemnification protections for customers of its AI services. These initial announcements are one example of how AI developers are keen to assuage deployers and, by extension, end-users that they are investing significantly in making AI models safer. AI developers appear to have a strong business incentive to invest in AI safety efforts.

3.6. By the same argument, AI deployers also have a similar business incentive to invest in AI safety efforts[22]. Some aspects of AI that make it more trustworthy[23] are likely to also depend on safety efforts made by AI deployers, alongside AI developers. To some extent, the level of AI safety in a given situation would likely depend on the complementary efforts of AI developers and deployers.

3.7. For instance, for AI-based hiring software, both the enterprise selling the foundation model underlying the software and the enterprise selling the hiring

[16] Anthropic. Making AI systems you can rely on. Retrieved from https://www.anthropic.com/company
[17] OpenAI. (2023, November 6). New models and developer products announced at DevDay. Retrieved from https://openai.com/blog/new-models-and-developer-products-announced-at-devday
[18] The Guardian. (2023, November 6). OpenAI's ChatGPT faces copyright lawsuits from customers. Retrieved from https://www.theguardian.com/technology/2023/nov/06/openai-chatgpt-customers-copyright-lawsuits
[19] Miller, R. (2023, June 26). Adobe indemnity clause designed to ease enterprise fears about AI-generated art. TechCrunch. Retrieved from https://techcrunch.com/2023/06/26/adobe-indemnity-clause-designed-to-ease-enterprise-fears-about-ai-generated-art/
[20] Smith, B. (2023, September 7). Co-pilot copyright commitment: AI legal concerns. Microsoft.Retrieved from https://blogs.microsoft.com/on-the-issues/2023/09/07/copilot-copyright-commitment-ai-legal-concerns/
[21] Suggs, N. & Venables P. (2023, October 12). Shared fate: Protecting customers with generative AI indemnification. Google. Retrieved from https://cloud.google.com/blog/products/ai-machine-learning/protecting-customers-with-generative-ai-indemnification
[22] Based on expert interviews.
[23]Bach, T. A., Khan, A., Hallock, H., Beltrão, G., & Sousa, S. (2022). A systematic literature review of user trust in AI-enabled systems: An HCI perspective. International Journal of Human–Computer Interaction, 1-16.

software to end-users are likely to have a role in ensuring that the enterprise is safe to use at a societal level. During the training and development of AI, AI developers are likely to have a larger role in ensuring AI safety. This would include using an unbiased training dataset, conducting extensive model testing, and choosing fairness principles. In contrast, during deployment to end-users, AI deployers are more likely to be able to influence safety through equitable product design choices that improve the end-user's ability to make unbiased hiring decisions[24].

3.8. Therefore, this submission assumes there is a business incentive for AI developers and deployers (jointly referred to as AI enterprises) to invest in safety.

3.9. This business incentive for safety shared by AI developers and deployers as part of their market relationship can be leveraged by policymakers to achieve the socially desirable level of AI safety.

3.10. Consequently, the AI safety policy challenge can be reframed as: **How can policymakers influence the business incentives of AI enterprises to align their profit-maximizing level of safety efforts with what is socially desirable?**

3.11. The socially desirable level of safety may depend on many factors but is ultimately some benchmark that policymakers would arrive at based on the social and political context. But, assuming such a socially desirable level of AI safety enterprise effort has been defined, policymakers could intervene to influence the profit-maximizing level of enterprise safety effort to be at that desired level. **This**

[24] Quiñonero Candela, J., Wu, Y., Hsu, B., Jain, S., Ramos, J., Adams, J., ... & Basu, K. (2023, June). Disentangling and operationalizing AI fairness at linkedin. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 1213-1228).

submission proposes that policymakers consider doing so by influencing AI enterprise revenues to be more responsive to their safety efforts.

3.12. Such a regulatory intervention to promote AI safety is unlikely to dampen firms' business incentives for innovation. Instead, it may motivate firms to consider greater AI safety as part of the same innovation that drives their business revenues for the following reasons:

3.12.1. By strengthening firms' revenue responsiveness to safety efforts, regulators can increasingly encourage firms to view safety investments as drivers of value creation instead of mere regulatory compliance costs.

3.12.2. In markets where consumers are increasingly concerned about the ethical implications and safety of AI, firms that innovate by incorporating safety into their products can differentiate themselves from competitors.

3.12.3. Firms that innovate by integrating safety into their products can also capture first-mover advantages. By setting safety standards in their industries, they can establish proprietary technologies, methodologies, and brand reputations that are difficult for competitors to replicate.

3.12.4. A proactive approach to safety in their innovation strategies can further enable firms to align more closely with current and future regulatory standards. This can reduce the costs and disruptions associated with compliance, avoiding penalties and enabling smoother market access.

3.13. Beyond alleviating the safety-innovation trade-off, this proposed regulatory approach introduces three other crucial advantages that address the unique characteristics of AI in its current stage. These include:

3.13.1. **Focusing on observable outcomes instead of causal explanations:** By encouraging policymakers to regulate based on firm outcomes i.e. observable AI failures, this approach reduces the need for causal explanations of those failures, which has been linked to gaps in accountability[25] [26] [27].

3.13.2. **Incentivizing safety through adaptive regulatory incentives:** Given AI's inherent unknowability and unpredictability[28], an incentive-based regulatory regime is likely to be more adaptive and therefore effective than direct regulation of AI enterprises[29].

3.13.3. **Encouraging technical experts to lead AI safety innovations:** The proposed framework reduces the need for policymakers to understand all the intricate technical aspects of AI to ensure its safety. Instead, it places the onus on enterprise technical experts to lead safety innovations that mitigate AI failures.

4. Policy implications of regulating AI based on enterprise revenue

Based on the regulatory approach proposed above, the author conceptualized market dynamics between AI developers and deployers as a symmetric two-player game, where both parties' revenues increase with their AI safety efforts. Further, a regulatory penalty function is assumed

[25] Bathaee, Y. (2017). The artificial intelligence black box and the failure of intent and causation. Harv. JL & Tech., 31, 889.
[26] Hohma, E., Boch, A., Trauth, R., & Lütge, C. (2023). Investigating accountability for Artificial Intelligence through risk governance: A workshop-based exploratory study. Frontiers in Psychology, 14, 1073686.
[27] Porter, Z., Zimmermann, A., Morgan, P., McDermid, J., Lawton, T., & Habli, I. (2022). Distinguishing two features of accountability for AI technologies. *Nature Machine Intelligence*, *4*(9), 734-736.
[28] White, J. M., & Lidskog, R. (2022). Ignorance and the regulation of artificial intelligence. Journal of Risk Research, 25(4), 488-500.
[29] Bova, P., Di Stefano, A., & Han, T. A. (2023). Both eyes open: Vigilant Incentives help Regulatory Markets improve AI Safety. arXiv preprint arXiv:2303.03174.

to apply based on the observable AI failure rate, safety efforts by AI enterprises, and an unpredictable 'AI black-box' variable. Based on the Nash equilibrium result of the model, this submission highlights six key policy implications to consider when regulating AI based on enterprise revenue.

4.1. **Penalty regime responsiveness:** For AI enterprises to invest more in safety efforts, the penalty regime imposed should be highly responsive to the decline in expected AI failures.

4.2. **Calibrating the "no-fine safety level":** The level of safety efforts by AI enterprises deemed by regulators as sufficient to impose no fine — referred to as the "no-fine safety level" — should not be set so high as to discourage investment in safety efforts.

4.3. **User awareness of AI enterprise safety efforts:** Increasing awareness among end-users about safety efforts exerted by enterprises can motivate enterprises to increase their safety efforts if their revenues increase with the increase in end-user awareness.

4.3.1. A highly successful user awareness campaign may eliminate the need for penalties if it makes the enterprise revenue sufficiently responsive to enterprise safety efforts.

4.3.2. Increasing end-user awareness may increase the equilibrium safety level in society due to market differentiation efforts by enterprises to be "most safety conscious" to capture revenue share informed by safety considerations.

4.3.3. Increasing user awareness may also reduce reliance by all stakeholders on past data about failures and may increase reliance on the actual safety efforts made, which are more likely to determine the long-run failure rate. Since AI is an emerging technology, data on AI failures is limited and has a stochastic component i.e. they can take extreme values in the short run. Instead, using data on safety efforts to make demand decisions may be a more suitable approach.

4.4. **Risk of extreme stochastic failures:** Regulation based on enterprise revenue may help mitigate enterprise risk of extreme stochastic failures. If AI enterprises exert high safety effort but observe an extreme kind of random failure, they can absorb that failure in the long run since their safety effort is determined by long-term averages of expected failure, which in turn affects end-user demand and consequently, net revenue. When a rare but extreme AI failure does occur randomly, firms' net revenue may be affected by greater penalty payments to the government in the short run. However, in the long run, it could be cushioned from such extreme incidents since end-user demand would likely remain relatively stable.

4.5. **Revenue sharing by AI enterprises:** Influencing revenue shared between AI developers and deployers may be a useful policy instrument to incentivize safety efforts. Since the safety efforts of AI enterprises are assumed as complementary, there may also be scope for cooperation between the two agents, which should be encouraged. In cases where there is an imbalance in the safety efforts between AI

developers and deployers, regulators could influence the revenue share or consider cross-subsidization.

4.6. **Cost of safety efforts by AI enterprises:** Reducing the cost of safety efforts may incentivize safety efforts. While firms have an inherent incentive to reduce costs of safety efforts, regulators must further encourage such measures especially when safety requires a high upfront investment that individual enterprises may not be able to make.

Since AI is rapidly evolving, such a market revenue-based approach to regulation is just *one* of many strategies that policymakers could employ as part of their toolkit. When new evidence emerges, regulators must adapt by either intensifying their commitment to or retracting from any specific policy approach.

5. Policy recommendations to incentivize safety efforts by AI enterprises

Based on the policy implications of the author's model, this submission proposes that policymakers consider the following recommendations to incentivize safety efforts by AI enterprises.

5.1. Implement a penalty framework highly responsive to changes in observed AI failures

5.1.1. **Tiered penalty system based on failure rates:** Implement a penalty system where fines are directly correlated with the observed rate of AI failures. This approach could have predefined tiers, where each tier

corresponds to a range of failure rates i.e. lower failure rates result in enterprises forgoing a smaller proportion of revenue in penalties and vice versa. Such a structure could incentivize AI enterprises to reduce their failure rates actively.

5.1.2. **Real-time monitoring and assessment:** Establish continuous monitoring systems that track the performance and safety record of AI applications in real-time. Such a system could also include reported failures from end-users. This data could inform the penalty regime, allowing for immediate adjustments in fines based on changes in the failure rates so that penalties are continuously aligned with the current safety performance of the AI enterprise.

5.1.3. **Variable penalty rates adjusted annually:** Adjust penalty rates annually based on the aggregated data of AI failures across the industry. If the overall trend shows improvement, baseline penalties could be reduced to reflect higher industry safety standards. Conversely, if failure rates increase, penalty rates could be escalated accordingly. This could encourage long-run industry-wide improvement in AI safety efforts.

5.1.4. **Incentives for reporting and rectifying failures:** Offer reduced penalties for enterprises that proactively report their failures and take swift action to rectify them. This could help bolster transparency and immediate response to failures, reducing the overall expected failure rate by ensuring that emerging lessons are disseminated across the industry.

5.1.5. **Link penalties to impact of failures:** Differentiate penalties not just on the frequency of failures but also on their severity and impact. This means that AI applications with the potential to cause significant harm could face steeper penalties for failures. This could help emphasize the importance of safety in high-stakes AI applications, encouraging enterprises to allocate safety efforts where they are most needed.

5.1.6. **Safety performance bonds:** Consider encouraging AI enterprises to post safety performance bonds, with the bond amount adjusted based on the enterprise's historical and current failure rates. A lower failure rate could result in a lower bond requirement, while a higher rate could increase the bond amount. This could motivate enterprises to have a financial stake in keeping failure rates low.

5.1.7. **Discounts on insurance premiums for low failure rates:** Collaborate with insurance companies to offer lower premiums on liability insurance for AI enterprises that demonstrate lower-than-average failure rates. This could provide a direct financial incentive for companies to invest in safety and reduce failures by decreasing operational costs.

5.1.8. **Penalty exemptions for investment in safety R&D:** Offer exemptions or reductions in penalties for enterprises that demonstrate significant investments in AI safety research and development. This could help acknowledge and reward proactive efforts to enhance safety beyond immediate operational practices.

5.1.9. **Dynamic penalty adjustments based on peer comparison:** In addition to absolute failure rates, adjust penalties relative to industry averages or benchmarks. Enterprises performing better than their peers in terms of failure rates could receive reduced penalties, while those performing worse could face harsher penalties. This could encourage a competitive approach to safety where companies strive to outperform each other in reducing failures.

5.1.10. **Feedback loop for continuous improvement:** Implement a system where the penalty framework is also subject to regular review and adjustments based on its effectiveness in reducing AI failures. Seeking stakeholder feedback from AI enterprises, safety experts, and end-users could help ensure the system remains fair, responsive, and effective at encouraging safety improvements.

5.2. Establish an attainable and adaptive "no-fine safety level"

5.2.1. **Stakeholder consultation:** Regularly engage with AI developers, deployers, industry experts, and consumer protection groups to assess and adjust the "no-fine safety level" based on technological advancements and societal expectations. This could be done through roundtables, online feedback platforms, and advisory committees.

5.2.2. **Dynamic benchmarking:** Consistently update safety benchmarks through technology review sessions. Additionally, establish protocols for making adjustments, to ensure that benchmarks remain relevant with the rapid pace of AI advancements and industry-specific needs.

5.2.3. **Transparent enterprise guidelines:** Provide clear, detailed guidelines on what constitutes sufficient safety efforts to meet the "no-fine safety level" benchmark. This could include examples of best practices and recognized safety standards in various AI application areas. Additionally, publish and regularly update comprehensive safety guidelines, conduct explanatory workshops, and maintain an accessible online repository of guidelines and best practices for AI enterprises.

5.2.4. **Safety incentive structures:** Offer tax incentives, grants, and public recognition programs for companies that exceed safety standards, encouraging continuous investment in and innovation of AI safety beyond the minimum requirements.

5.2.5. **Modular safety standards:** Recognize the diversity of AI applications by developing modular safety standards tailored to different risk profiles so that the "no-fine safety level" is appropriate for the specific context of each AI system. Develop and periodically revise sector-specific safety standards through dedicated working groups, allowing flexibility in compliance methods to accommodate diverse technological approaches and sector-specific risks.

5.3. Increase end-user awareness of AI enterprise safety efforts

5.3.1. **User-friendly transparency reports:** Encourage or mandate AI enterprises to publish annual transparency reports detailing their safety efforts, methodologies, and outcomes. These reports should be easily

accessible and understandable to end-users and other non-expert public audiences.

5.3.2. **Safety labels and certifications:** Develop a standardized safety certification or labeling system for AI products and services, akin to nutritional labels or energy efficiency ratings that inform users about enterprise safety efforts and standard compliance at a glance. Regularly updating these certifications could ensure companies maintain high safety standards to keep their ratings.

5.3.3. **Public education campaigns:** Launch public educational campaigns through media, public seminars, and institutional partnerships to raise awareness about the importance of AI safety and enterprise efforts towards it.

5.3.4. **User feedback mechanisms:** Implement robust feedback mechanisms that allow users to report their experiences, concerns, and suggestions regarding AI safety. This feedback could enable enterprises to adjust their safety efforts and for users to feel directly involved in the safety process. Publicly acknowledging and implementing user feedback could also help enhance trust and awareness.

5.3.5. **Safety-focused user engagement platforms:** Develop online platforms or forums dedicated to AI safety, where end-users can learn about safety efforts, engage with AI enterprises, and share their experiences. These platforms could feature webinars, Q&A sessions with safety experts, and tutorials on understanding AI safety metrics. Making safety a central topic

of conversation could increase user literacy on the subject and highlight the efforts of proactive companies.

5.3.6. **Collaboration with consumer advocacy and research groups:** Partner with consumer advocacy groups and researchers to review AI enterprises' safety efforts. As independent evaluators, these groups could provide an unbiased assessment of safety measures, shaping user trust and awareness.

5.3.7. **Incentives for market differentiation based on safety:** Provide incentives for enterprises to differentiate their products from competitors based on product safety. Encourage enterprises to build safety into the core design and marketing of their AI products to link product appeal with safety efforts. This could be in the form of tax breaks, awards, or public recognition. Encouraging competition among enterprises for developing and deploying the “most safety-enabled” AI products in the market could help raise the overall industry safety standards.

5.3.8. **Industry standards and best practices showcase:** Organize annual industry showcases or conferences where AI enterprises can present their safety innovations, share best practices, and discuss safety standards. These events would serve as platforms for companies to demonstrate their commitment to safety directly to end-users, industry peers, and regulators. Recognizing industry leaders in safety could help create aspirational benchmarks for others to achieve.

5.3.9. **Collaborative public-private partnerships:** Establish partnerships between governments, AI enterprises, and non-governmental

organizations to create AI safety public awareness campaigns. Systematically coalescing regulatory authority, technical expertise, and public trust in this manner could help embed the value of AI safety in the public consciousness.

5.4. Support AI enterprises in mitigating risks associated with stochastic failures

5.4.1. **Safety investment recognition system:** Administer a formal system to recognize and document the safety efforts of AI enterprises. This record could be referenced in mitigating penalties in the event of an unexpected failure, acknowledging the company's ongoing commitment to safety.

5.4.2. **Stochastic failure insurance:** Encourage or require AI enterprises to contribute to a collective insurance fund designed to cover costs associated with stochastic failures. Contributions to this fund would be scaled based on the company's safety investment level, offering better rates to those demonstrating higher safety efforts.

5.4.3. **Regulatory sandboxes for safety innovations:** Establish regulatory sandboxes that allow companies to test and refine innovative safety technologies in controlled environments without the usual regulatory constraints. Successful innovations could consequently be fast-tracked for wider adoption, promoting continuous improvement in safety efforts.

5.4.4. **Public reporting and transparency initiatives:** Mandate the public reporting of stochastic failure incidents and related safety efforts, encouraging that consumer demand is informed by a company's long-term safety commitment rather than isolated incidents.

5.4.5. **Incentives for continuous staff training:** Offer tax incentives, grants, or other benefits to encourage companies to continuously educate their staff so that AI safety remains a priority and evolves with emerging risks and technologies.

5.4.6. **Collaborative safety research initiatives:** Facilitate industry-wide collaborations on safety research, pooling resources and knowledge to tackle the challenges of stochastic failures. Such initiatives could accelerate the development of more robust safety solutions towards unpredictable risks.

5.5. Encourage revenue sharing and cross-subsidization between AI developers and deployers

5.5.1. **Revenue-sharing agreements:** Encourage or mandate the creation of revenue-sharing agreements that specifically link the revenue share of AI developers and deployers to their safety efforts. The agreements could include clauses that adjust revenue shares based on meeting predefined safety benchmarks, incentivizing both parties to invest in safety.

5.5.2. **Cross-subsidization funds:** Establish a fund that supports cross-subsidization between AI developers and deployers. By subsidizing the safety investments of one party by using contributions from the other, such a fund could encourage a cooperative investment in safety.

5.5.3. **Cooperative safety incentives and R&D grants:** Implement a system of financial incentives, such as tax breaks or grants, for AI enterprises that demonstrate cooperative safety efforts exceeding industry standards. Offer

grants or subsidies for joint safety R&D projects between AI developers and deployers. This could encourage developers and deployers to work together more closely on safety initiatives.

5.5.4. **Joint safety certification programs:** Develop certification programs that recognize outstanding cooperative safety efforts between developers and deployers. Certification could be tied to financial or reputational benefits, motivating enterprises to pursue joint safety strategies.

5.5.5. **Joint liability arrangements:** Introduce joint liability provisions for AI safety incidents, making both developers and deployers responsible for any failures. Such a legal framework could encourage both stakeholders to work together on safety measures to mitigate risks and potential liabilities.

5.5.6. **Collaborative knowledge platforms and guidelines:** Create and support platforms where AI developers and deployers can share safety-related knowledge, best practices, and technologies. Publish best practices as a roadmap for enterprises to strengthen their collaborative safety strategies.

5.6. Reduce costs of safety efforts by AI enterprises

5.6.1. **Standardized safety protocols:** Standardize baseline safety protocols and technologies across AI enterprises to promote economies of scale and reduce the cost of safety efforts as they are more widely adopted.

5.6.2. **Incentives for safety expenditure and cost-saving innovations:** Provide subsidies, tax credits, or grants for AI enterprises to cover any prohibitory fixed costs associated with developing or purchasing safety-enhancing

technologies. Additionally offer grants for projects focused on innovative safety solutions and cost-reduction techniques.

5.6.3. **Safety education and training funds:** Increase funding to educational institutions and vocational training programs specializing in AI safety and ethics. This could help produce a larger workforce of safety experts and reduce the cost of hiring qualified personnel for AI enterprises.

5.6.4. **Cross-industry public-private partnerships:** Facilitate partnerships between public research institutions and private AI enterprises across industries to co-develop safety technologies. By sharing resources, technologies, and expertise the cost and risk associated with safety R&D could be significantly reduced.

5.6.5. **Promote safety startups:** Support incubators and accelerators focused on nurturing startups that develop safety technologies or provide safety services. These programs could offer mentoring, funding, and resources, reducing the startup costs for new entrants, and encouraging innovation in safety measures.

5.6.6. **Regulatory streamlining for safety approval:** Simplify and streamline the regulatory approval processes for new safety innovations. Reducing the time and cost required to navigate regulatory hurdles could encourage firms to invest in innovative safety solutions.