Rehse, Dominik; Valet, Sebastian; Walter, Johannes

Research Report

Safe AI made in the EU

ZEW policy brief, No. 22/2024

Provided in Cooperation with:
ZEW - Leibniz Centre for European Economic Research

Suggested Citation: Rehse, Dominik; Valet, Sebastian; Walter, Johannes (2024) : Safe AI made in the EU, ZEW policy brief, No. 22/2024, ZEW - Leibniz-Zentrum für Europäische Wirtschaftsforschung, Mannheim

This Version is available at:
https://hdl.handle.net/10419/308835

ZEW policy brief

// NO.22 | DECEMBER 2024

// Dominik Rehse (ZEW), Sebastian Valet (ZEW),
Johannes Walter (ZEW)

Safe AI Made in the EU: Proposal for an EU Safe Generative AI Innovation Program

We propose an EU Safe Generative AI Innovation Program to address a market failure in generative AI development. While developers can capture significant value from generative AI capability improvements, they bear only a fraction of potential safety failure costs, which leads to underinvestment in the technological breakthroughs necessary to make generative AI safe. The EU should establish explicit incentives for the necessary technological breakthroughs, complementing its existing policy responses to the rapid proliferation of generative AI.

We propose a milestone-based incentive scheme where pre-specified payments would reward the achievement of verifiable safety milestones. This "pull" funding mechanism would aim to create predictable development paths for safety improvements, similar to how scaling laws have guided capability advances. The scheme would use robust safety metrics and competitive evaluation to prevent gaming while ensuring meaningful progress. Success would be measured through a combination of specific safety dimensions (like factual accuracy and harm prevention) and broader performance metrics, validated through adversarial testing and public comparative evaluation.

The program's design would be technology-neutral and it could be open to all qualified institutions, with rewards calibrated through incentive-compatible elicitation mechanisms. This approach mirrors other applications of outcome-based funding, such as advance market commitments in vaccine development. It might also provide the breeding ground for "Safe AI made in the EU".

KEY MESSAGES

- Current market incentives are insufficient to ensure the development of safe generative AI systems, as companies bear only a fraction of potential safety failure costs while capturing most of the benefits from capability improvements. This leads to underinvestment in making the technological breakthroughs necessary for making generative AI safe.
- The EU's current regulatory approach is mostly based on restrictions and penalties, while existing innovation initiatives focus on general capabilities rather than safety – creating a gap in positive incentives for safety innovation.
- A milestone-based incentive program using robust safety metrics and competitive evaluation could create predictable development paths for safety improvements while remaining technology-neutral.

MARKET FAILURE IN GENERATIVE AI SAFETY

Negative externalities are not internalised in generative AI development.

The development of generative AI currently suffers from a fundamental market failure. While companies can capture significant value from improvements in AI capabilities, they bear only a fraction of the costs associated with safety failures. Negative externalities include eroding trust in digital information due to disinformation campaigns of unprecedented scale, reducing the barriers to accessing dangerous information, e.g. details on biochemical weapons or cybersecurity vulnerabilities and systemic risks from AI applications in healthcare, the judicial system or digital content moderation, where even small errors can lead to large-scale harm due to their widespread use.

Making generative AI safe to use as policy goal.

The policy goal should be to make generative AI "safe" to use. This would be the case if negative externalities could be entirely avoided or at least consciously managed. However, in the current market environment, generative AI technology is adopted widely and very rapidly, boosted with multibillion private and public investments (Giattino et al., 2024), without being safe in this sense. This goes against the explicit EU policy goal of "ensuring that AI is human-centric and trustworthy" (European Commission, 2024).

TECHNOLOGICAL BREAKTHROUGH NEEDED FOR GENERATIVE AI TO BECOME SAFE

Safety not naturally built into current generative AI technology stack.

To make generative AI safe, a technological breakthrough or multiple technological breakthroughs are needed (Anthropic, 2023; d'Avila Garcez, 2020; LeCun, 2022). Current generative AI is created by learning patterns in high-dimensional data, interpolating between the examples it has been trained on while being able to recombine these patterns in sophisticated and sometimes surprising ways. Safety-aspects are not naturally built into the technology by first principles. For instance, factual correctness of statements produced by generative AI cannot be guaranteed, even if its users explicitly demand it. While the body of knowledge of Wikipedia is almost certainly part of the training corpus of most generative AI algorithms producing text, text created by such algorithms does not naturally adhere to the same editorial standards. Pre-training fixes such as curating the training corpus and post-training safety measures such as fine-tuning the algorithm and putting guardrails around it cannot fully compensate for the fundamental safety deficits of the core technology.

ROBUST EU POLICY RESPONSE, BUT LACKING FOCUS ON REWARDING TECHNOLOGICAL BREAKTHROUGH

Different policy responses can create incentives to innovate.

The main categories of potential policy responses to this state of affairs are banning, taxing or rewarding. In principle, all three responses could lead to the necessary technological breakthroughs. Banning would entail restricting or forbidding generative AI in total or for certain application areas until pre-defined safety standards are met, potentially creating incentives for companies to develop safer alternatives as a first-mover advantage. Taxing for potential or realized negative externalities can come in different forms. This includes product liability mechanisms with which producers are held liable for realized negative externalities of their products. This liability creates a de facto tax on unsafe systems, incentivizing companies to invest in safety. In turn, rewarding innovations through mechanisms like research grants tries to accelerate the development of safety breakthroughs more directly.

Multi-pronged regulatory approach by the EU.

The EU's response includes elements of all three categories. In light of the rapid development of the market for generative AI and the great uncertainty about the nature and size of its negative externalities, this is a sensible policy choice. While a more purist regulatory approach focussing only on one or two of the categories would be favourable in terms of reducing regulatory complexity and avoiding unnecessary red tape, not relying exclusively on any single mechanism provides redundancy and robustness.

EU bans and taxes unsafe generative AI.

The EU AI Act is mostly concerned with banning and taxing. AI, including generative AI, is entirely forbidden in certain application areas. Generative AI is also regulated more explicitly, with substantial parts of the regulation referring to harmonized standards and codes of practice which are still under development. This includes documentation and testing requirements, laying the foundation for traditional product liability mechanisms. This particularly concerns the recently updated EU Product Liability Directive and the EU AI Liability Directive, which is still under discussion, both of which are essentially taxing approaches.

EU rewards the development and adoption of generative AI in general, not safe generative AI in particular.

The EU also states that it wants to foster and reward the development of safer generative AI with various initiatives initiated during the last legislative term. However, they seem to be more focussed on fostering the adoption of generative AI in general rather than developing safe generative AI in particular. For instance:

- "AI Factories" or the "Common European Data Spaces" might increase access to computing power, data and talent, but do not focus on safety of generative AI as such.
- Testing generative AI in "Regulatory Sandboxes" might provide some useful insights into real-world safety issues, but don't explicitly create incentives for developing safe generative AI.
- Horizon Europe's "European Innovation Council" and "InvestEU" are meant to provide funding for AI in general, but currently do not focus on increasing safety of generative AI as such.
- "GenAI4EU" focusses on the productive use of generative AI in the EU, but does not address the broader safety challenges of generative AI.
- The "Large AI Grand Challenge" did not focus on safety of generative AI and did not fund start-ups stating the intention to make fundamental progress in this area.

While these initiatives may help strengthen the EU's position in generative AI development, where currently less than a handful of companies compete internationally, they focus on capability building and deployment, rather than safety innovation, which leaves the core externality problem unaddressed.

REWARDING THE DEVELOPMENT OF SAFE GENERATIVE AI WITH THE EU SAFE GENERATIVE AI INNOVATION PROGRAM

Explicit incentive for making generative AI safe.

Going forward, the EU should more strongly focus on creating explicit rewards for breakthrough innovations with respect to generative AI safety, as a complement to existing reward schemes. We propose to launch a Safe Generative AI Innovation Program designed as a milestone-based incentive scheme for the development of safe generative AI. The program should be technology-neutral, meaning that it should not favour any specific technology or approach, and should reward actual successes, not only effort. Its primary goal should be to give the EU access to safe generative AI.

Program designed as a structured competition.

The program would work as follows: The EU would define milestones in terms of safety and other performance metrics. An exemplary safety metric could be the level of factual correctness of the content produced by a generative AI model if a user explicitly demands it (i.e., unwanted "hallucinations"). The developer whose model reaches a milestone first, receives a predefined reward.

Other developers could still compete for the next milestone. The size of the reward should be set such that a critical mass of developers would seriously compete for reaching it first.
The program should be open to all firms, universities and other institutions that want to participate. Depending on digital sovereignty considerations, the program could require that the participating institutions are based in the EU. Similarly, the program could require that all models developed under the program are open sourced to make the developed safety techniques widely available. However, the primary goal should remain the focus, with secondary objectives included only if they do not alter incentives, to avoid "mission creep".

Board of experts advising on the program's design.

As part of the governance structure, a board of experts advising on the design of the program should be established. This particularly concerns the design of the milestone structure, the definition of the safety and performance measures as well as the milestone rewards. Below, we present a starting point for the discussions of the board.

DESIGNING EXPLICIT INCENTIVES FOR THE DEVELOPMENT OF SAFE GENERATIVE AI

Explicit incentives through milestone contracts.

The core difference to existing EU funding schemes for generative AI lies in the focus on creating explicit incentives for developing safe generative AI and the use of milestone contracts for efficiency. Milestone contracts are agreements where a funding agency commits to pay a pre-specified amount when a generative AI developer achieves certain development milestones, with payments increasing as more ambitious milestones are reached. These contracts reduce development risk by providing guaranteed revenue upon milestone completion while maintaining flexibility in how the milestones are achieved.

Milestones are common in generative AI development.

The core insight making milestone contracts interesting for incentivising the development of safe generative AI is that while calibrating competitive generative AI models requires significant resources, these resources can be managed to follow a predictable path, called "scaling laws". As the number of parameters and the amount of data used to train a model increase, the performance of the model usually improves, even though at a decreasing rate. Leading generative AI developers were able to attract billions of dollars in funding because they could demonstrate that they could manage to scale along such a predictable path, making such investments less risky.

Safety milestones to incentivise the development of safe generative AI.

Incentive schemes for developing safe generative AI should focus on finding similar predictable paths, where safety can be increased by developers in a way that is scalable and predictable. This might involve changing the model architecture, the training data, or the training process. If a predictable path is found, it can be used to scale up safety measures in a cost-efficient manner. Milestone contracts are a way to create incentives for finding and walking along such predictable paths. Each milestone consists of access to a model with a certain level of safety.

Potential tradeoffs between safety and other performance dimensions.

An increase in safety might well come at the expense of lower performance in other dimensions. To illustrate this point: The safest generative AI is one that does not generate anything at all, but a competitive level of performance is necessary for a model to be of actual use. This creates a tradeoff between safety and other dimensions of model quality that needs to be explicitly addressed in the incentive scheme.

BUILDING BLOCKS OF INCENTIVE SCHEMES

Implementing the requirements for designing incentive schemes for predictable paths and addressing the performance–safety tradeoff is challenging but not insurmountable. The first challenge consists of measuring safety and other performance dimensions. Without sound measures, it is difficult to track progress and design incentives schemes such as milestone contracts. The second challenge is to determine how large the reward linked to reaching each milestone should be.

MEASURING SAFETY AND OTHER PERFORMANCE DIMENSIONS

Imperfect measures are good enough, as long as they are sufficiently robust.

Precisely measuring aspects of safety such as factual correctness, if it is explicitly demanded from a generative AI model, and other performance dimensions is generally difficult and an active area of research (Ren et al., 2024; Vidgen et al., 2024; Zeng et al., 2024). However, when justifying the incremental but ultimately large investments in existing generative AI models, even imperfect measures seem to have been sufficiently accurate to guide large capital allocations. In general, it is important for any performance measures to be robust to "Goodhart's Law", which states that when a measure becomes a target, it ceases to be a good measure. For example, if we measure generative AI safety by a model's ability to avoid certain keywords, developers might simply optimise their models to avoid those specific words while still producing unsafe content in other ways.

Measuring safety via well-structured red teaming is likely to be robust.

Two classes of measures are necessary. The first class consists of measures of safety. This requires deciding which safety dimension to focus on. For instance, one focus could be on factual correctness, another one on harmful content. Given such a category of misbehaviour, it is necessary to design a testing procedure that consists of challenging a given model to misbehave and validating the extent to which misbehaviour can be observed. This is commonly referred to as "adversarial testing", particularly "red teaming". Relevant aspects of designing such testing procedures have already been discussed in an earlier ZEW policy brief (Rehse, Valet and Walter, 2024). It is particularly important to align the incentives of all parties involved and to provide the necessary coordination devices for efficient and effective red teaming. Given that different challengers and validators might have different approaches to challenging and validating, it is reasonable to expect that collective measures based on these approaches are relatively robust and unlikely to be gameable. The practical implementation of red teaming testing procedures – whether manual (e.g. Quaye et al., 2024), semi-automated (e.g. Deng et al., 2024) or fully automated (e.g. Zifan et al., 2024) – remains an active area of research. They all could contribute to measuring safety in a transparent and well-organized testing procedure. Similarly, ongoing work of the "US Artificial Intelligence Safety Institute Consortium" (AISIC), which – among other things – tries to develop AI safety measurement techniques could also be of help for developing robust safety measures.

Measuring other performance dimensions could follow established robust approaches.

The second class of measures consists of measures of other performance dimensions, for instance, a model's ability to generate high-quality text, images or audio. Many such measures are already available, but they are not yet fully standardised (Hendrycks et al., 2020; Sawada et al., 2023; Srivastava et al., 2022). Public comparative evaluation, such as Chatbot Arena (formerly LMSYS), appears to be a pretty robust measure. Chatbot Arena is a peer evaluation system where generative AI models compete directly against each other in blind tests across a wide range of tasks provided by the public. This again is difficult to game, for instance, by narrowly optimizing the model to perform well in this particular context. The result is a ranking of models that can be used to measure relative performance.

DETERMINING THE REWARD LINKED WITH REACHING EACH MILESTONE

Ballpark estimates of milestone rewards are good enough.

The reward associated with each milestone should be large enough to motivate the development of safe generative AI but not much larger than that, in order not to spend more than necessary and not to distort the existing market for generative AI development unnecessarily. However, given the great potential benefits of safe generative AI, the "perfect should not be the enemy of the good". If in doubt, rewards should be set on the high side to avoid underinvestment.

Choice-matching can provide the information base for setting milestone rewards.

The magnitude of each milestone's reward should be based on potential developers' expectations of what it takes to make them reach the milestone. One pragmatic way to get a glimpse of the distribution of market expectations is to elicit it in an incentive-compatible way using a choice-matching mechanism (Cvitanić et al., 2019). In a first step, the organising EU body would signal the intent to commit significant funding to the incentive scheme. As a second step, all potential participating AI developers would be asked to preregister for the incentive scheme. Only preregistered participants would be eligible for receiving the next milestone's reward.

As a third step, a random subset of the preregistered participants would be surveyed for their expected minimum reward to reach the milestone first. As part of the survey, the respondents would also be asked for their willingness-to-pay for state-of-the-art Graphical Processing Units (GPUs) or for another auxiliary good. The relevant good should be chosen such that respondents expect that other respondents answering the first question similarly would also answer the second question similarly. For instance, in the case of GPUs, respondents who need higher rewards to reach milestones would likely also have higher valuations for GPUs due to their more resource-intensive development plans. Following this reasoning, each respondent will also be told that they will be offered the good at the willingness-to-pay of the respondent with a similar stated minimum reward. Ideally, they would be required to buy the good at this willingness-to-pay. This creates an incentive for stating the minimum reward truthfully and accurately, since misstating it would match the respondent with others who have different true valuations, potentially leading to unfavourable GPU prices. For instance, overstating the minimum reward would match with respondents who truly need higher rewards and likely have higher GPU valuations, with the result that the GPUs would be offered at too high a price to be attractive. The opposite holds for understating the minimum reward. GPUs are particularly attractive as an auxiliary good, because they are likely to be needed for any approach to develop safer generative AI. An alternative auxiliary good might be novel data, for instance, non-public data amendable to calibrate generative AI models.

The last step would then consist of deciding on the size of a milestone's reward based on the elicited distribution of minimum rewards. As the reward size is increased, more and more respondents would be willing to participate in the incentive scheme. Deciding on the cutoff requires judgement and will also be based on the budget available for the program. The elicitation mechanisms could be reused after reaching a given milestone to determine the reward for the next milestone.

Imprecise information on needed milestone rewards is better than no information.

Such elicitation mechanisms are not without their challenges. For instance, the respondents have to understand the incentive mechanism in order for it to work. They might also be willing to collude with others to manipulate the outcome, for instance, to systematically push up the milestone rewards. However, careful design of the elicitation mechanism could mitigate these issues and lead to good approximations of market expectations for the necessary minimum rewards.

THE BIGGER PICTURE: MORE PULL MECHANISMS FOR INNOVATION IN THE EU

More “pull” funding.

The proposed Safe Generative AI Innovation Program would be a “pull” mechanism for innovation. While push mechanisms like research grants pay for inputs, pull mechanisms like milestone contracts pay for outputs and outcomes. This is particularly relevant for AI safety, where we know we need socially valuable innovation but incentives of private actors seem to be too small to discover who is best placed to develop it and how exactly it should be developed.

Advantages of pull mechanisms.

Pull mechanisms provide several key advantages over push mechanisms for incentivising safe generative AI development:

- They reduce demand uncertainty by signalling clear market demand for safety innovations.
- They place technological risk on the innovating firms who know their capabilities best.
- They remain solution-agnostic, allowing different approaches to safety.
- They only require payment when results are achieved.
- They can be designed to reward scaling successful approaches.

Successful precedents.

Pull mechanisms have already proven successful in other domains with similar market failures, such as for the development of a pneumococcal vaccine through an advance market commitment under which a fixed price was guaranteed for purchases of a functional vaccine , lowering the risk of the investment and thereby encouraging private sector involvement (Kremer, Levin and Snyder, 2020). Pull mechanisms are also promising to tackle climate change (Arnesen and Glennerster, forthcoming). Such mechanisms also have precedents in developing computer technology. For instance, the Grand Challenge by the US Defense Advanced Research Projects Agency (DARPA) led to multiple breakthroughs in the development of autonomous vehicles. By adapting such pull mechanisms to generative AI safety, the EU has an opportunity to pioneer a new model for responsible innovation that aligns private incentives with public benefit.

Getting the “Safe AI made in the EU” agenda back on track.

The Safe Generative AI Innovation Program might also re-energise AI innovation in the EU. Two years after the launch of ChatGPT, the EU still has very few serious contenders in the global market for generative AI development. By creating clear incentives for safety innovation, this program could help European companies develop a competitive advantage in an increasingly important market dimension.

Embedding into existing EU structures to be discussed.

The ideal embedding of the program within the existing structure of the EU institutions needs to be discussed in greater detail. Milestone contracts are common in public procurement, for instance, when the European Commission purchases customized software products for its own use. Pull funding mechanisms are also used by DARPA for innovation projects, which the Joint European Disruptive Initiative (JEDI) and the European Innovation Council (EIC) as well as some member states’ initiatives try to emulate. Tasking them with implementing the Safe Generative AI Innovation Program might be a good match.

REFERENCES

Anthropic, P. B. C. Core Views on AI Safety: When, Why, What, and How. (2023).

Arnesen, William and Rachel Glennerster. Market Shaping to Combat Climate Change. In New Directions in Market Design, edited by Irene Lo, Michael Ostrovsky and Parag Pathak (forthcoming).

Cvitanić, Jakša et al. Honesty via choice-matching. American Economic Review: Insights 1.2 (2019): 179–192.

d'Avila Garcez, Artur, and Luis C. Lamb. Neurosymbolic AI: The 3rd wave. arXiv e-prints (2020): arXiv-2012.

Deng, Boyi, Wenjie Wang, Fuli Feng, Yang Deng, Qifan Wang and Xiangnan He. Attack prompt generation for red teaming and defending large language models. arXiv preprint arXiv:2310.12505 (2023).

European Commission. European approach to artificial intelligence. Retrieved from: https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence (2024).

Giattino, Charlie, Edouard Mathieu, Veronika Samborska and Max Roser. Artificial Intelligence. Retrieved from: https://ourworldindata.org/artificial-intelligence (2024).

Hendrycks, Dan et al. Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300 (2020).

Kremer, Michael, Jonathan Levin and Christopher M. Snyder. Advance Market Commitments: Insights from Theory and Experience. AEA Papers and Proceedings 110. 269–73. (2020).

LeCun, Yann. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. Open Review 62.1 (2022): 1-62.

Quaye, Jessica, Alicia Parrish, Oana Inel, Charvi Rastogi, Hannah Rose Kirk, Minsuk Kahng, Erin Van Liemt et al. Adversarial Nibbler: An Open Red-Teaming Method for Identifying Diverse Harms in Text-to-Image Generation. In 2024 ACM Conference on Fairness, Accountability, and Transparency, pp. 388–406 (2024).

Rehse, Dominik, Sebastian Valet and Johannes Walter. Using market design to improve red teaming of generative AI models. No. 06/2024. ZEW policy brief (2024).

Ren, Richard et al. Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress? arXiv preprint arXiv:2407.21792 (2024).

Sawada, Tomohiro et al. Arb: Advanced reasoning benchmark for large language models. arXiv preprint arXiv:2307.13692 (2023).

Srivastava, Aarohi et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615 (2022).

Vidgen, Bertie et al. Introducing v0. 5 of the ai safety benchmark from mlcommons. arXiv preprint arXiv:2404.12241 (2024).

Zeng, Yi et al. Air-bench 2024: A safety benchmark based on risk categories from regulations and policies. arXiv preprint arXiv:2407.17436 (2024).

Zou, Andy, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043 (2023).

//

Imprint

Authors: Dominik Rehse (ZEW), Sebastian Valet (ZEW), Johannes Walter (ZEW)

Publisher: ZEW – Leibniz Centre for European Economic Research
L 7, 1 · 68161 Mannheim · Germany · info@zew.de · www.zew.de/en · x.com/zew_en

President: Prof. Achim Wambach, PhD · Managing Director: Claudia von Schuttenbach

Editorial responsibility: Fabian Oppel · kommunikation@zew.de

ZEW

Leibniz Association