ZEW policy brief

// NO.06 | APRIL 2024

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

Using Market Design to Improve Red Teaming of Generative AI Models

With the final approval of the EU's Artificial Intelligence Act (AI Act), it is now clear that general-purpose AI (GPAI) models with systemic risk will need to undergo adversarial testing. This provision is a response to the emergence of "generative AI" models, which are currently the most notable form of GPAI models generating rich-form content such as text, images, and video. Adversarial testing involves repeatedly interacting with a model to try to lead it to exhibit unwanted behaviour. However, the specific implementation of such testing for GPAI models with systemic risk has not been clearly spelled out in the AI Act. Instead, the legislation only refers to codes of practice and harmonised standards which are soon to be developed. In this policy brief, which is based on research funded by the Baden-Württemberg Foundation, we propose that these codes and standards should reflect that an effective adversarial testing regime requires testing by independent third parties, a well-defined goal, clear roles with proper incentive and coordination schemes for all parties involved, and standardised reporting of the results. The market design approach is helpful for developing, testing and improving the underlying rules and the institutional setup of such adversarial testing regimes. We outline the design space for an extensive form of adversarial testing, called red teaming, of generative AI models. This is intended to stimulate the discussion in preparation for the codes of practice, harmonised standards and potential additional provisions by governing bodies.

KEY MESSAGES

- The EU AI Act's provisions on general-purpose AI with systemic risks require adversarial testing. To ensure safe and reliable AI models, these adversarial tests should not only be conducted by the developers, but also by an independent third party.
- An extensive form of adversarial testing, called red teaming, can be an effective way to test AI models. To produce reliable information, a red teaming process needs a clear goal, well-defined roles, and incentive and coordination schemes in place.
- Following the AI Act, codes of practice and harmonised standards are developed to govern its implementation. This policy brief outlines the design space for the implementation of red teaming for general-purpose AI models. It is intended to stimulate a discussion on the incentive schemes and coordination mechanisms.

RULES FOR GENERATIVE AI IN THE AI ACT

AI Act requires the riskiest generative AI models to be evaluated via adversarial testing.

Surpassing human-level performance in answering questions on a wide range of expert topics, generating photorealistic imagery, and autonomously writing production-ready code: new capabilities of generative AI are being reported with remarkable regularity. Yet, new skills like generating flawless prose or having advanced knowledge in biochemistry can prove to be a double-edged sword: while they can support humans in writing text or speeding up drug discovery, these capabilities have also raised concerns, such as generative AI models proliferating hate speech or enabling malicious actors to develop new pathogens. To mitigate risks associated with emerging AI capabilities, it will be vital to have a process in place that identifies them reliably and quickly.

Exact implementation of adversarial testing is still to be determined.

Late in the legislative process of the European Union's Artificial Intelligence Act (AI Act), lawmakers were playing post-hoc catch-up with these fast-paced technological advancements. The AI Act was finally approved by the European Parliament in March 2024 after years of complex negotiations and numerous revisions. The final version of the regulation not only contains a risk-based approach, which places its strictest rules on the riskiest AI applications, but also separate provisions on "general-purpose AI" (GPAI) models, a classification that includes generative AI and acknowledges the technology's varied and far-reaching applications.

Most notably, according to Article 52d of the AI Act, GPAI models with systemic risk are required to undergo model evaluations, for which the ground rules are laid out in Annex IXa. Specifically, such GPAI models will need to be examined, where applicable, through adversarial testing. The exact implementation of adversarial testing will need to be determined in codes of practice and harmonised standards, which are yet to be developed.

RED TEAMING IS AN EXTENSIVE FORM OF ADVERSARIAL TESTING

Effective testing needs to study a model's behaviour.

Several strategies are available to evaluate an AI model, including analysing the model's training data, examining its architecture, or documenting its training techniques. While all of these are essential for a thorough model evaluation, they may not be sufficient for some classes of AI models. Unlike traditional software, many models are impossible to fully understand merely by examining their source code or training data. These are referred to as "black box" models. Instead, to gain a more comprehensive understanding of a black box model, it is necessary to directly observe the model's outputs in response to various inputs. In other words, it is important to study its behaviour. Therefore, it is laudable that the AI Act calls for adversarial testing of GPAI models, which is based on the principle of studying model behaviour.

During adversarial testing, testers intentionally craft inputs that cause the GPAI model to exhibit unwanted behaviour (e.g. Radharapu et al., 2023). Unwanted behaviour refers to the generation of an undesired or incorrect output given a certain input. For example, early versions of large language models could easily be prompted to provide incorrect information or reveal protected training data.

Red teaming can be an effective form of testing GPAI models.

Although there is no clear consensus on what the term red teaming means in practice, an interpretation close to its original meaning in cybersecurity suggests that it should include but also extend beyond adversarial testing. The term red teaming originates from military simulations and was later adopted by the cybersecurity community to describe attacks by a "red team" attempting to break into a computer system or network. Red teaming of an AI model not only involves crafting inputs to elicit unwanted behaviour, but also simulating different kinds of attacks, for instance changing the model weights or stealing the model itself (Ji, 2023). A comprehensive evaluation of a GPAI model should hence be based on red teaming.

RED TEAMING IS SUITABLE TO EVALUATE GENERATIVE AI, BUT NEEDS SOUND ECONOMICS

Red teaming can identify a wide range of unwanted behaviour, be done continuously and explore vast input and output spaces.

Red teaming as a tool for AI model evaluation has been growing in popularity since the rise of generative AI. Most of the major developers of frontier GPAI models state that they conduct internal red teaming (see e.g. Ganguli, 2022; OpenAI, 2024). Red teaming has several advantages as an evaluation mechanism. First, it allows for the discovery of a wide range of unwanted behaviours, such as the generation of false information, incitement to commit crimes or disclosure of proprietary information. Second, red teaming can be conducted continuously, making it well-suited to evaluate GPAI models that are constantly changing, even after their placement on the market. This is relevant because a GPAI model can acquire or lose capabilities with each modification, such as after fine-tuning it for a particular use case. Sometimes these changes are not intended by the developers and thus go unnoticed. In such cases, red teaming can serve as an early warning system. Third, red teaming can address the vast input and output possibilities of GPAI models. Smaller standardised test sets that are commonly used to evaluate predictive AI models are not suitable for models with such large input and output spaces. For example, in the case of language models, any combination of letters, numbers, and symbols can serve as possible input. Red teaming can be an efficient method to explore these large input and output spaces if designed as such.

Current internal red teaming efforts by AI developers lack clearly defined goals, rules and incentives.

For all the aforementioned reasons, it is commendable that developers of GPAI models conduct internal red teaming. However, the current organisation and reporting of red teaming efforts leave room for improvement as there is no standardised approach for conducting red teaming, even for models of the same type. This makes the comparison of results unnecessarily difficult. Mainly, current attempts typically lack a clearly defined goal, making it difficult to determine when a model has been sufficiently tested. It is also up to the developers to decide how to incentivise red teamers and how to report the results of the red teaming exercise. This invites cherry-picking of both the incentive mechanisms and the reported findings.

Market design provides toolbox to design red teaming.

To maximise the usefulness of red teaming, we believe it is necessary to clearly define the goal of the red teaming exercise, properly align the incentives of all parties involved, and establish rules governing coordination. This makes red teaming of GPAI models a fruitful field for economists, as these problems can be tackled from a market design perspective.

STRUCTURED RED TEAMING NEEDS A CLEARLY DEFINED GOAL

The goal of red teaming should be to determine the difficulty of eliciting unwanted behaviour.

In the following, we outline important implementation aspects of red teaming with respect to the specific challenges of testing GPAI models. From a market design perspective, red teaming can be interpreted as an information production process. At the core of this process are the red teamers producing information about possible attack vectors that could lead to unwanted behaviour from a GPAI model. This process should be purposefully designed to ensure the efficiency and usefulness of the produced information.

Most importantly, a clear goal for the red teaming process must be defined. Consistent with the view that model behaviour should be directly tested, we suggest that the goal of a structured red teaming process should be to determine the difficulty of eliciting unwanted behaviour in a particular usage context. In such a setting, difficulty can be measured by the effort required for red teamers to elicit unwanted behaviour from a GPAI model. Effort can be approximated by, for example, the necessary monetary expenditure or the skill levels of red teamers. Additionally, a well-defined usage context is imperative because GPAI models and their underlying models can typically handle a large variety of inputs and outputs. While some behaviours might be universally

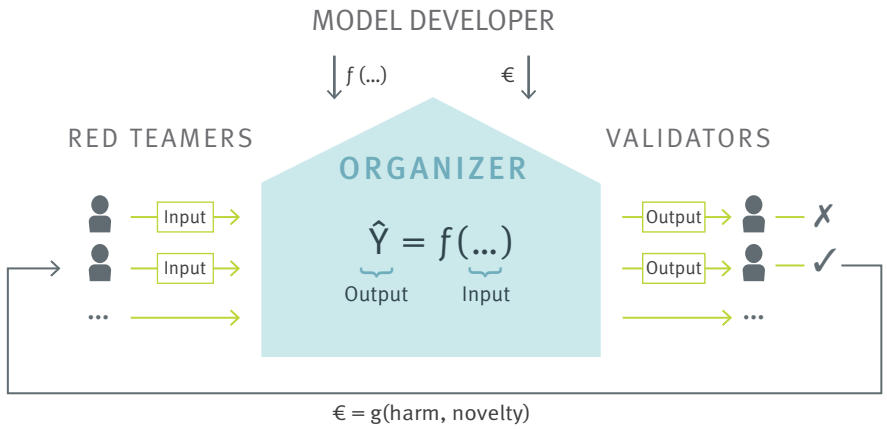
unwanted, such as personal threats or the output of protected content, for other behaviours, the context is critical in determining whether the behaviour is, in fact, unwanted. For example, fabricating information might be acceptable in a usage context of fictional writing, but constitutes unwanted behaviour if the usage context is related to political news.

Because of the large input and output spaces, unwanted behaviour can never be completely ruled out. Instead, a structured red teaming process can determine the likelihood of unwanted behaviour to occur for any given level of effort. This is analogous to product testing in other markets, where risks are minimised to a level deemed acceptable to society rather than eliminated entirely. The likelihood of unwanted behaviour in a usage context can inform a regulator about the level of risk associated with the GPAI model.

Red teaming could yield similar results to product tests in other markets.

FIGURE 1: ROLES AND PROCESS OF RED TEAMING GPAI MODELS

A model developer provides a GPAI model to be evaluated and commissions an independent organiser for this service. The red teaming organiser recruits red teamers and validators and provides the infrastructure for them to interact with the model. Red teamers are rewarded for creating inputs that lead to unwanted model outputs. Validators verify whether an output is unwanted and assess its severity. Rewards for red teamers increase with harmfulness and novelty.



ROLES IN A RED TEAMING PROCESS NEED TO BE SEPARATED

The definition of clear roles within the red teaming process is crucial for effectiveness. For this purpose, each role entails a set of tasks that requires its own incentive and coordination scheme. Here, we identify at least four roles in the red teaming process: the red teamers and the validators as producers of information as well as the organiser of the red teaming process and the developer of the GPAI model.

Separation of roles allows the design of effective incentive schemes.

Red teamers should be rewarded for successfully providing inputs that lead to unwanted behaviour, i.e., to an unwanted output given a certain input. Their reward structure should also take other factors into account to maximise the efficiency of the information production. When specifically rewarded for the novelty of an input leading to unwanted behaviour or the novelty of the behaviour itself, red teamers are incentivised to find new attack vectors and to explore the large input and output spaces of GPAI models. This provides a coordination scheme among – potentially very many – red teamers. Additionally, the reward structure should consider the severity of an unwanted behaviour to incentivise red teamers to focus on the attack vectors with the greatest potential for harm.

Validators verify that the inputs of red teamers are within a defined usage context, check whether outputs are unwanted and assess the severity of unwanted behaviour. In current industry players' internal red teaming efforts, the roles of red teamers and validators are typically not separated. Often red teamers are simply asked to evaluate their own success (e.g. Ganguli, 2022). This lack of independent validation makes it difficult to implement suitable incentive mechanisms for red teamers. Additionally, validators are essential in accounting for the ambiguity of outputs. Many modalities of GPAI model outputs, such as text or images, can be ambiguous, and their meaning can depend on the usage context. It is critical to take ambiguities and context into account when evaluating whether a model behaviour is unwanted.

Both red teamers and validators should be selected depending on the usage context. For instance, in a medical context, domain experts such as physicians or medical scientists have the knowledge necessary to red team and to validate. For a chat bot that is used by the general public, any user could potentially serve as a red teamer or validator (Deng, 2023). The selection of red teamers and validators should be carefully considered, because their skills and domain knowledge are critical factors for generating useful and reliable information. The importance of red teamer and validator selection has already been recognised by industry players. While earlier internal red teaming efforts often employed crowdworkers (Ganguli, 2022), more recent efforts are said to be conducted with the help of domain experts (OpenAI, 2024; Touvron, 2023).

The role of the organiser of a red teaming process is to act as an intermediary to bring together red teamers and validators on one side and the developers of GPAI models on the other side. The red teaming process requires suitable infrastructure for red teamers and validators to be able to interact with a model. This infrastructure should be provided by the organiser. Providing such a service would also require monetary compensation so that it could potentially be developed into a business model.

To avoid conflicts of interest, the developer of the GPAI model should be independent from all other roles. Ideally, the red teamers, validators, and the organiser would be fully independent from the developers. The developer of a GPAI model should be incentivised to participate in the red teaming process. It is conceivable that independent red teaming will become mandatory before placing a model on the market. Since models are often modified even after their market placement, post-placement red teaming in regular intervals could also become required.

All other roles should be independent from the developer of the AI model.

THE CASE FOR AN INDEPENDENT GPAI EVALUATION

If regulatory authorities are meant to fully trust red teaming evaluations, it seems important for organisers of red teaming to be independent third-party entities who possess the incentive to truthfully and thoroughly check the developers' models. If red teaming is done internally by developers or by closely affiliated entities, the aforementioned incentives and coordination schemes cannot be fully trusted. However, mandating independent third-party red teaming raises two questions: who should be tasked with organising red teaming? And who should bear the cost of red teaming?

Red teaming should be done externally by an independent third party.

With the proliferation of GPAI models and the regulatory requirements for safety evaluations, the demand for red teaming is likely to increase. Organising a red teaming process and bringing together skilled red teamers and developers of GPAI models is a task that creates value for both. As different models often exhibit similarities or are built on top of each other, red teamers might gain the skill to successfully elicit unwanted behaviours, which they could then transfer to evaluations of similar models. With this in mind, it is conceivable that a new, lucrative market for independent red teaming could emerge.

Red teaming could turn into a new service market.

Developers of GPAI models should bear the costs of red teaming.

The cost of red teaming could possibly draw from the market for financial audits, in which companies pay for their financial audits to be performed by external entities. Such a model could be applied to the market for GPAI model red teaming as well. Additionally, requiring developers to bear the cost of independent red teaming would incentivise them to test their models as well as possible beforehand, e.g. through prior extensive internal red teaming efforts. This is because red teamers would be rewarded for successfully eliciting unwanted behaviour. For a given level of red teaming effort, a well-tested model that produces fewer of these unwanted behaviours would be financially beneficial. Moreover, it is a fundamental economic tenet that, whenever private market activity causes negative externalities for society, the responsible parties should internalise the additional societal costs. Requiring the developers of frontier GPAI models to finance external independent red teaming of their models is an elegant way to have them internalise the social costs of this new technology. To address concerns about possible innovation-dampening effects for startups and small businesses, legislators could provide discounted capital and/or funding specifically designated for red teaming expenses.

LITERATURE

Deng, Wesley H., Boyuan Guo, Alicia Devrio, Hong Shen, Motahhare Eslami, and Kenneth Holstein. 2023. "Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice." Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 377, 1–18.

Ganguli, Deep, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, et al. 2022. "Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned". arXiv preprint arXiv:2209.07858.

Ji, Jessica. 2023. "What Does AI Red-Teaming Actually Mean?". Accessed April 4, 2023. https://cset.georgetown.edu/article/what-does-ai-red-teaming-actually-mean.

OpenAI. 2024. "GPT-4 Technical Report". arXiv preprint arXiv:2310.03693.

Radharapu, Bhaktipriya, Kevin Robinson, Lora Aroyo, and Preethi Lahoti. 2023. "AART: AI-Assisted Red-Teaming with Diverse Data Generation for New LLM-powered Applications". arXiv preprint arXiv:2311.08592

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, et al. 2023. "Llama 2: Open Foundation and Fine-Tuned Chat Models". arXiv preprint arXiv:2307.09288.

//

ZEW policy brief

Authors: Dominik Rehse (ZEW), dominik.rehse@zew.de · Sebastian Valet (ZEW & KIT), sebastian.valet@zew.de · Johannes Walter (ZEW & KIT), johannes.walter@zew.de

Publisher: ZEW – Leibniz Centre for European Economic Research
L 7, 1 · 68161 Mannheim · Germany · info@zew.de · www.zew.de/en · twitter.com/ZEW_en
President: Prof. Achim Wambach, PhD · Managing Director: Claudia von Schuttenbach

Editorial responsibility: Fabian Oppel · cvd@zew.de

Quotes from the text: Sections of the text may be quoted in the original language without explicit permission provided that the source is acknowledged.

© ZEW – Leibniz-Zentrum für Europäische Wirtschaftsforschung GmbH Mannheim