AI Safety Collective - Crowdsourcing Solutions for Critical AI Safety Challenges

Lye Jia Jun
Singapore Management University

Dhruba Patra
Indian Association for the Cultivation of Science

Philipp Blandfort
Independent Researcher

Overview

Artificial Intelligence (AI) is rapidly advancing, with agents and multimodal AI systems becoming increasingly prevalent. These technologies hold great promise but also present significant challenges, particularly when it comes to ensuring their safety and reliability. Despite efforts like red teaming (testing AI systems for vulnerabilities) and ongoing work in the research community, it's likely that severe issues will persist due to the complexity of these systems.

Additionally, there's a vast, untapped potential among developers and experts, which remains largely underutilized because there are currently few public challenges or platforms dedicated to applied AI Safety.

Our AI Safety Collective idea aims to tap into this potential by creating a global, collaborative platform focused on AI Safety. On this platform, we want to promote challenges and bounties (similar to bug bounties but with a focus on AI Safety problems) to draw in contributions from a diverse community of experts, developers, and AI enthusiasts.

Problem

As AI technologies like large language models (LLMs), multimodal systems, and AI agents become more sophisticated and widely integrated into critical areas such as healthcare, finance, and security, the risks associated with their deployment are growing at an alarming rate. These systems are evolving faster than our ability to ensure their safety and reliability, creating a significant potential for harm.

Why This Problem is Critical:

1. **Hard-to-Test Systems:** AI systems, especially those involving complex interactions between different modalities (e.g., text, image, and voice), are incredibly challenging to test thoroughly. Even after rigorous testing processes like red teaming, hidden vulnerabilities often remain, leading to unpredictable and potentially dangerous outcomes.

2. **Persistent and Severe Issues:** Given the complexity of these systems, even well-tested AI can still harbor severe issues that are difficult to detect and correct. This is especially problematic as these technologies become more embedded in essential decision-making processes, where errors can have serious consequences.

3. **Underutilization of Expert Potential:** There is a wealth of expertise among developers and AI safety professionals that is not being fully leveraged. The lack of accessible, applied

challenges means that many experts do not have the opportunity to contribute to solving real-world AI safety problems, which stifles innovation and progress in this critical area.

Timing: At the current stage, we see AI-based solutions spreading rapidly in the industry, where AI safety expertise is scarce. At the same time, the trend is towards more complex AI solutions. In particular, it seems very likely that AI agents will be very common in industry applications in 2 years from now.

Solution

AI Safety Collective: Our platform will serve as a global hub where AI companies can post safety challenges, and AI Safety enthusiasts and developers from around the world can identify concrete issues, propose solutions & build solutions.

This platform will operate on a **bounty system**, incentivizing participants to contribute their knowledge and skills to solve critical AI safety issues. By leveraging the collective intelligence of a global community, we aim to create a more robust and comprehensive approach to AI safety.

Key Features:

- **Challenge Posting & Searching**: AI companies can post detailed safety challenges, specifying the problem, desired outcomes, and any specific requirements. For example, a company could offer a bounty for generating copyrighted content with an image generation model, or for identifying important failure cases of their AI agents. Participants are able to browse and search for challenges. They can tackle these issues, receiving monetary rewards for significant results. In addition, we feature challenges originally posted elsewhere (e.g. from workshops at academic conferences).

- **Bounty System**: Companies can create bounties for identifying issues or proposing specific solutions. This is similar to bounties in cybersecurity, while we use the term more liberally to include any prize money for challenges.

- **Leaderboard**: To further motivate participants, points can be collected for contributions and we display an overall ranking.

Strategy and Difference to Existing Solutions:

While our proposed solution has significant overlap with the AI Safety bounty approaches described by Rethink Priorities in [1], we don't exclusively deal with x-risk. Rather, we would start focusing on the non-catastrophic risks posed by companies in the industry, as this greatly increases the number of potential customers while reducing the complexity of the problems. We think that this difference makes it possible to start the project as a for-profit.

Technical Foundation:

- **Crowdsourcing Framework**: The platform will be built on a robust crowdsourcing framework modeling successful platforms like Kickstarter, HackerOne, etc., ensuring scalability and security.

- **AI-Driven Matching**: Machine learning algorithms (such as [2]) will suggest challenges to the most relevant developers based on their skills and experience.

Process

Timeframe	What will you do?
Next 3 months	First, we need to optimize the product-market fit and implement the core platform: - Get feedback from Rethink Priorities and non-profit AI safety orgs - Discuss with AI companies and AI safety experts which bounties and challenges are valuable and practical - Complete core platform development - Design bounty trials together with AI companies and developers/experts, with legal assistance for contracts - Launch bounty trials
2025	The focus here is on non-catastrophic AI safety risks since the potential customer base is much larger: - Extend the platform to include AI safety challenges - Kickstart targeted marketing to attract people from global AI safety experts and developer communities. - Intensive sales to get more bounties and challenges to the platform - Establish partnerships with red teaming and auditing companies, since their services can be recommended for a commission in case clients are not ready for bounties yet
2026	With sufficient traction, collaborations with frontier labs become feasible and the stakes can be increased: - Establish partnerships with evals organizations and frontier labs - Launch bounties and challenges focused on x-risk, which tend to require significant investments (see, e.g. evals-based bounties as explained in [1]) - Regular assessments of risks caused by bounties and challenges (e.g. to avoid a bounty causing anyone to create a highly capable misaligned AI that breaks free)
2027	At this stage, we should be able to automate operations and expand: - Integrate advanced analytics and AI-driven tools to automate and improve operations - Explore ways to extend the business, e.g. to add further community aspects

Impact on AI Safety & Key Risks

Positive Impact on AI Safety:

- **Creating Grounded Solutions**: By crowdsourcing practical AI safety challenges from the industry, our platform will foster the creation of diverse and grounded solutions, greatly advancing the field of AI safety.

- **Inspiring AI Safety Innovation**: The bounty system will accelerate the pace of innovation in AI Safety, as AI Safety experts compete to develop the best solutions.

- **Increased Accountability**: The platform will foster greater accountability in AI development by ensuring that safety challenges are addressed in an open and transparent manner, with solutions being rigorously evaluated by a diverse community of experts.

- **Attract Talents to AI Safety:** By offering meaningful challenges and competitive rewards, the platform will draw top-tier talent from across disciplines to the field of AI safety, enriching the community with fresh perspectives and expertise.

Key Risks and Mitigation:

- **Quality Control**: There is a risk that very few proposed solutions built could meet the required standards. We will mitigate this by implementing a rigorous peer review process for submissions (from companies & participants) and leveraging AI to filter out low-quality submissions.

- **Risks Caused by Bounties**: For advanced AI systems, if a bounty hunter manages to successfully trigger a dangerous behavior, this could have harmful consequences up to a catastrophic scale [1]. To mitigate this risk, bounties are categorized based on risk categories, and the overall risk assessment is regularly revisited.

- **Cold Start:** The platform is only interesting to participants if sufficient lucrative bounties and challenges can be found while posting them is only worth the effort if sufficient participants can be expected to work on them. We address this problem by starting with bounties requiring little preparation time from the companies, making use of existing AI Safety communities for distribution, and featuring challenges and bounties from other websites on our platform.

- **Market Fit**: The biggest risk is that industry organizations could be hard to convince that bounties or challenges help them to avoid issues. To mitigate this, we will talk to AI companies from a very early stage, establish partnerships with AI companies, and develop the first bounties and challenges in close collaboration with companies and AI safety experts.

By addressing these risks and leveraging the power of crowdsourcing, our **AI Safety Collective** is poised to make an extremely significant positive impact on the future of AI, ensuring that it develops in a safe, transparent, and ethical manner.

Appendix

Inspiration

- https://www.aicrowd.com/
- https://hackenproof.com/
- https://www.hackerone.com/
- https://www.microsoft.com/en-us/msrc/bounty-ai
- https://taskdev.metr.org/introduction/ (bounties for good submissions)

Academic Challenges

- https://www.llmagentsafetycomp24.com/
- https://llm-pc.github.io/
- https://trojandetection.ai/?ref=mlcontests

References

[1] https://rethinkpriorities.org/publications/ai-safety-bounties
[2] https://paperswithcode.com/task/recommendation-systems