Vulnerability disclosure for AI safeguards. How open should programs be and what incentives are necessary?

BY ED PARSONS · MARCH 24, 2026 · LAST UPDATED ON APRIL 2, 2026

What you will learn

- How vulnerability disclosure applies specifically to AI safeguards and systems.
- The pros and cons of making AI disclosure programs more open/restricted.
- The kinds of incentives that motivate researchers.
- Which disclosure program structures can help organizations improve their AI security.

In a recent NCSC blog post on adapting vulnerability disclosure for AI safeguards, the authors posed a series of questions to researchers.

Intigriti, being a global crowdsourced security provider trusted by the world's leading organizations, has observed a significant rise in the number of Intigriti customers, including AI assets in the scope of their programs.

At the same time, a growing number of security researchers (or ethical hackers) are developing their AI-security skills, evidenced in submissions to the Intigriti platform.

On public programs specifically, NCSC asked:

- **How public and open should Safeguard Bypass Bounty Programs (SBBP) and Safeguard Bypass Disclosure Programs (SBDP) be?**
- **What incentives are most appropriate for the safeguard context?**

Let's first recap what safeguards are. As described in NCSC's post: "Safeguards are techniques developers use to prevent AI systems from producing policy-violating outputs or actions. These include model-level changes like refusal training or unlearning, and external tools like auxiliary classifiers. But these safeguards aren't foolproof and can be bypassed through techniques such as jailbreaking, agent hijacking, and indirect prompt injection."

In the context of the worldwide eruption of AI-enabled applications, effective safeguards are increasingly important. As the focus of security research and assurance activity shifts towards AI technology, the ability to test safeguards will become more important to deliver impactful crowdsourced security programs.

Public and Private Programs

Typically, there are four different types of crowdsourced programs, on a spectrum from public to private:

- **Public programs** are fully 'advertised' with all details, and they are accessible to any researcher who has registered to submit reports of vulnerabilities found.
- **Registered programs** are different only because they aren't publicly advertised. These programs are visible to all researchers who are registered on the platform. Any registered researcher can view the program details and create submissions. You can optionally restrict participation to ID-checked researchers only, which adds an extra layer of trust while keeping the program broadly accessible within the platform.
- **Application programs** are visible in the same way as a public program, but with only limited information provided. Researchers need to apply to join these programs; if accepted, the remaining details are then disclosed.
- Finally, **invitation-only programs (private programs)** are, by definition, not visible on crowdsourced cybersecurity platforms. Instead, a customer applies specific criteria against the known skills and other attributes of researchers to determine whom they would like to have on their program; these individuals are then invited to take part.

The NCSC blog rightly notes that "more open programs encourage diversity of submissions, whilst invite-only programs make it easier to manage risks". Furthermore, a "hybrid model, where any user can participate but trusted testers are given extra access and affordances, may provide the best of both worlds."

Given the relative scarcity of AI-security expertise, it could be argued that private programs are more likely to attract sufficient researchers with the requisite skills. Yet public programs are how many researchers learn and improve, increasing supply. Therefore, the selection of a Private or Public program is more of a strategic choice, and it may even be a case of test-and-learn.

Intigriti offers the following insights to organizations considering SBBP or SVDP on AI systems, to maximize impact:

- Public programs with constrained scopes can help mitigate risk while offering scale and the opportunity to identify researchers to be involved in private programs on other assets.
- Private programs offer more control and allow programs to grow at a sustainable rate whilst managing risk.
- Extract the maximum value out of each report. Beyond remediating and closing, take the lessons learned and apply them across the wider organization.

Incentives

We can reasonably hypothesize that AI safeguard-testing skills will become a standard or core skill among researchers if they are appropriately incentivized. Crowdsourced security service providers and the organizations that commission and administrate bug bounty programs tend to know the level of

bounties and other incentives necessary to attract the attention of suitable researchers, as organizations are, after all, competing for the attention of the best researchers.

The absolute reward value depends on the use of cases of the AI systems concerned. We must recognize that the value and importance of AI systems will increase. How it does so alongside the scarcity or otherwise of AI safeguard-testing skills, and how demand/supply dynamics play out, will influence the incentives that organizations need to offer to help protect themselves.

Reward value may vary by industry. For industry-specific insights, use our **Bug Bounty Calculator** to provide a baseline for AI-related vulnerabilities, including typical reward amounts and payout estimates.

But it's not solely about bounties. A multitude of other factors, such as excellent community engagement, fantastic triage, trusted verification, clear code of conduct, multi-layered encryption, and **many other elements**, need to be factored in.

Organizations considering SBBP or SBDP should consider the following measures to incentivise researcher activity:

- Run targeted campaigns on specific features that are higher risk / more impactful.

Offer extra rewards for:

- higher severity vulnerabilities in specific areas of concern,
- certain classes of issues (e.g., safeguard bypasses),
- worst-case scenarios/outcomes.
- Treat researchers well! If someone submits an excellent bug, ensure they receive meaningful recognition.
- Timely, transparent communication on delays and key decisions, such as validity or criticality, is essential to engage researchers.
- Make it as easy as possible for researchers to hack on your platform by offering resources, technologies, and processes.

Early AI security specialists are helping define this field. Their work is not only uncovering weaknesses but also shaping the emerging norms around what constitutes a meaningful bypass and how its impact should be assessed. It's important to recognize that as demand for safeguard-testing grows, those shaping methodologies and impact narratives today will effectively set the benchmarks for tomorrow.

In short, demand-supply dynamics for AI-specific skills are not yet benchmarked, and the specific incentive structure for safeguarding bypass skills is an ongoing area of research.

As a last observation: it's terrific that, at a time when "AI" is touted as the solution to so many challenges, NCSC's observations make clear the ongoing importance and value of human researchers in challenging AI-enabled systems. At Intigriti, we believe crowdsourced security research will remain essential to accelerate the discovery of high-impact vulnerabilities in an evolving technology landscape. Creative humans will continue to be an important element of AI bypass-testing.

For more on Intigriti's stance on AI, read '**How AI is leveraged to enhance the Intigriti platform**'.

Contributors

- Eleanor Barlow, Senior Cybersecurity Technical Writer, Intigriti
- Alex Olsen, Head of Hackers, Intigriti
- Ottilia Westerlund-Trew, Hacker Engagement Manager, Intigriti

AUTHOR

Ed Parsons

Ed Parsons is Chief Operating Officer for Intigriti. Before joining Intigriti, Ed was Vice President of the world's largest member association for cyber professionals and led an international cybersecurity consultancy, renowned for research and technical expertise. As a cybersecurity professional, Ed spent several years helping organizations investigate and respond to cyber threats from nation-states and organized crime groups. He is a Certified Information Systems Security Professional (CISSP) and a UK Chartered Cyber Security Professional.