Beyond CVSS: rethinking scoring systems amidst AI Safety and Security

BY ELEANOR BARLOW · AUGUST 6, 2026

Want to read the latest version? Visit the link below:
https://www.intigriti.com/blog/business-insights/rethinking-cvss-scoring-systems-amidst-ai-safety-and-security

CVSS open framework, rapid recap

- Stands for Common Vulnerability Scoring System.
- Owned by a US-based non-profit organization, the Forum of Incident Response and Security Teams (FIRST).
- The purpose is to help response teams quickly and easily calculate the severity of cybersecurity vulnerabilities based on metrics.
- Latest version: (4.0) designed to assess multiple environments and dimensions, including exploitability and impacts.

The heart of this discussion

The crux of this article explores the fact that while CVSS is still the right tool for many AI security findings, AI safety needs customer-specific, outcome-based scoring.

While CVSS remains a strong standard for technical vulnerabilities, including many AI security findings, AI systems can also create safety findings including harmful outputs, unsafe tool use, misleading responses, or policy-violating behaviour that creates real business risk without mapping cleanly to confidentiality, integrity, or availability.

For those cases, organizations need a custom, outcome-based severity model that reflects the product, the audience, and the customer's worst-case harms.

Scoring systems explored

For context, the two most common types of severity scoring systems are CVSS and a basic 6-point scale (None, Low, Medium, High, Critical, Exceptional). Read Understanding signal-to-noise for vulnerability management success for more on this.

Traditional bounty tables map technical severity to payout. This means that the CVSS gives you a base score regarding the CIA triad (Confidentiality, Integrity, Availability) impact, plus exploitability factors.

When it comes to bug bounty programs, a "Critical" finding pays X, a "High" finding pays Y, and so on. The model works because the underlying classes of bugs have a stable, system-level impact that is measurable objectively.

But with the potential for AI agents to produce harmful outputs while remaining technically secure, this approach requires adaptation for AI Safety.

CVSS and AI Security

AI Security protects the system itself. This includes its data, its capabilities, and the actions it can take. Those issues usually look like traditional vulnerabilities and can be scored with a traditional CVSS. For instance, if a researcher identifies that a customer service chatbot has a ‘lookup my order’ tool wired into the back office, and if a prompt injection causes a customer-service chatbot to call an order-lookup tool without proper authorization and return another customer’s order details, confidentiality is impacted, and CVSS can be used.

CVSS and AI Safety

AI Safety focuses on whether the model or AI-enabled workflow can produce harmful, misleading, biased, or unacceptable outcomes. Those issues do not fit CVSS because they are about brand, regulatory, and human harm and impact, not technical severity.

This is different from the CVSS v4.0 Safety metric, which concerns predictable human injury arising from exploitation of a technical vulnerability; here, AI safety refers to harmful or unacceptable model behavior and its business or user impact.

Three key elements need to be factored in when it comes to AI safety.

Challenge 1: deployment context

CVSS includes environmental metrics, but those metrics are still designed for technical vulnerabilities. They do not, on their own, fully capture AI safety harms such as brand exposure, regulatory sensitivity, user trust, or context-specific misuse outcomes.

For instance, if a model produces hateful language on demand, it has not violated Confidentiality, Integrity, or Accessibility, not from a technical standpoint anyway. No tool has been misused. The harm is reputational to the brand, and the impact depends completely on the customer/viewer. Yes, it may breach content policies in most companies, but the CVSS score is not applicable. A CVSS score would not meaningfully represent the impact unless the behaviour can be mapped to a technical vulnerability affecting confidentiality, integrity, or availability.

What needs to be measured here is the customer-defined safety severity to rate how low or high the brand exposure would be because of said hateful language.

For instance, if the model produced the hateful language when marketing a children’s book, that would have fundamentally different brand exposure and impact than if it appeared in an internal Notion page.

CVSS environmental scoring still operates inside a technical-vulnerability framework and does not directly price reputational, regulatory, financial, or AI-behavioural harm.

Only those working in the environment will be able to highlight which vulnerabilities would have the greatest impact on their environment.

Challenge 2: business and regulatory context

Definitions should not be universal; instead, the reward model must encode the customer's worst-case definition. For instance, if a Banking firm and an Entertainment application had the same vulnerability, identical in every way, and ranked the same CVSS score, the severity would be completely different for each company.

This works the other way around: if a specific finding is deemed Low severity across the board, one industry or company type may still be hit much harder than another because the relevant context was not factored in. This leads to a distorted view of risk, where findings that appear low severity on paper may have a much greater real-world impact for certain industries or company types.

Challenge 3: outcome severity versus bypass elegance

In a traditional bug, the exploit is the path, and the impact is the destination. In a safety finding, eliciting the prohibited output is the goal.

CVSS is still useful for technical AI security findings, but it is not sufficient on its own for AI safety findings where the impact is brand, regulatory, user-harm, or business-context driven. AI security findings can often be scored with CVSS, while AI safety needs a severity model designed around product, users, and worst-case outcomes.

What is important is to reward the severity, repeatability, and real-world exposure of the harmful outcome, not just the cleverness of the jailbreak. A blunt jailbreak that crosses a line is still a finding and can be more useful than an elegant bypass that produces output the customer does not actually care about.

What actually matters when scoring AI safety findings

AI safety bounty tables should be outcome-based rather than CVSS-score-based: severity is defined by the harm category, affected audience, real-world exposure, and the customer's worst-case scenario.

The goal is to identify financial exposure, regulatory exposure, brand damage, and customer harm. Not just an analysis of how clever the hacker was in getting there.

At Intigriti, CVSS is still used for security findings, but a co-designed severity table for safety findings reflects how much it would have cost the customer in real life.

> **"If you were to ask an AI system in 300 BC if the Earth was round, it would have mocked you. It would have said, "Absolutely not, you're wrong because of XYZ". Simply because it didn't have all the context, it didn't have all the research yet. When it comes to security, there are new vulnerabilities right now that AI cannot find. The importance here is contextual risk assessments. You, as a company, have so many vulnerabilities coming in; it can be hard to decide what to fix first. Our model focuses on context and speed."**
>
> **Inti De Ceukelaire, Founding Member, Intigriti**

At Intigriti, we score vulnerabilities with a table built with you, against your worst-case scenarios, and combine techniques and scoring so that we show not only the most interesting findings, but the ones with the greatest Return on Security Investment (ROSI).

For more information, read 'AI Security and Safety'

Or, for a scoping call, reach out here.

AUTHOR

Eleanor Barlow

Eleanor Barlow is a London-based Senior Cyber Security Technical Writer at Intigriti, with 9+ years' experience reporting on and writing for the cyber and tech sector. She specializes in data-driven content on cybersecurity and bug bounty intelligence, helping organizations benefit from the latest trends and insights.

REQUEST A DEMO

intigriti.com/demo

VISIT THE WEBSITE

intigriti.com

GET IN TOUCH

hello@intigriti.com