Breaking the machine: jailbreaking AI and safeguarding the digital frontlines of vulnerable communities

by Fabien Leimgruber and Alexandru Lazar

Artificial intelligence is no longer confined to tech firms. In the nonprofit world, AI is increasingly developed and woven into the daily work of organizations serving vulnerable communities - allocating humanitarian aid, managing beneficiary data, monitoring and tracking project delivery, or coordinating emergency response. Many of these systems are hosted in the cloud, built by external third-party providers, and designed to process highly sensitive personal and behavioral data. As AI's capabilities expand, so does the ambition of those looking to exploit it. Jailbreaking has emerged as an accessible technique for bypassing safeguards, extracting hidden data, and manipulating outputs.[1] Combined with zero-day vulnerabilities, these attacks can undermine the very integrity of the digital infrastructure on which critical social services depend. A zero-day vulnerability in an AI system or tool is an undiscovered flaw in the system's code, model, or supporting infrastructure that threat actors can exploit before the organization or vendor is aware of it or able to patch it.[2] A jailbreak can uncover hidden weaknesses in the system's logic, giving malicious actors a way to exploit them. For the nonprofit sector, such security breaches are not abstract risks - they can translate directly into physical danger.

How jailbreaking differs from other cyber threats

Jailbreaking attacks the logic of the AI model itself. Instead of exploiting code or network weaknesses and breaking into a server, adversaries work to bend the system's understanding of language, context, and rules. This can mean changing prompts in ways that slip security filters, embedding hidden instructions inside seemingly harmless text, or using languages and dialects that evade existing safeguards. Once a method is discovered, it can be shared and replicated with remarkable speed. And unlike a fixed piece of code, AI models' vulnerabilities can shift depending on the data they process or the con-

1 Kranz, T. & Jonker, A. (2025), AI jailbreak: rooting out an evolving threat, IBM. https://www.ibm.com/think/insights/ai-jailbreak

2 Shastri, V. (2025), What is a zero-day exploit?, Crowdstrike. https://www.crowdstrike.com/en-us/cybersecurity-101/cyberattacks/zero-day-exploit/

text in which they operate. The risks go far beyond producing offensive or unsafe outputs. Jailbroken systems can be coerced into revealing proprietary prompts, leaking training data, or exposing bias patterns that adversaries can then exploit.

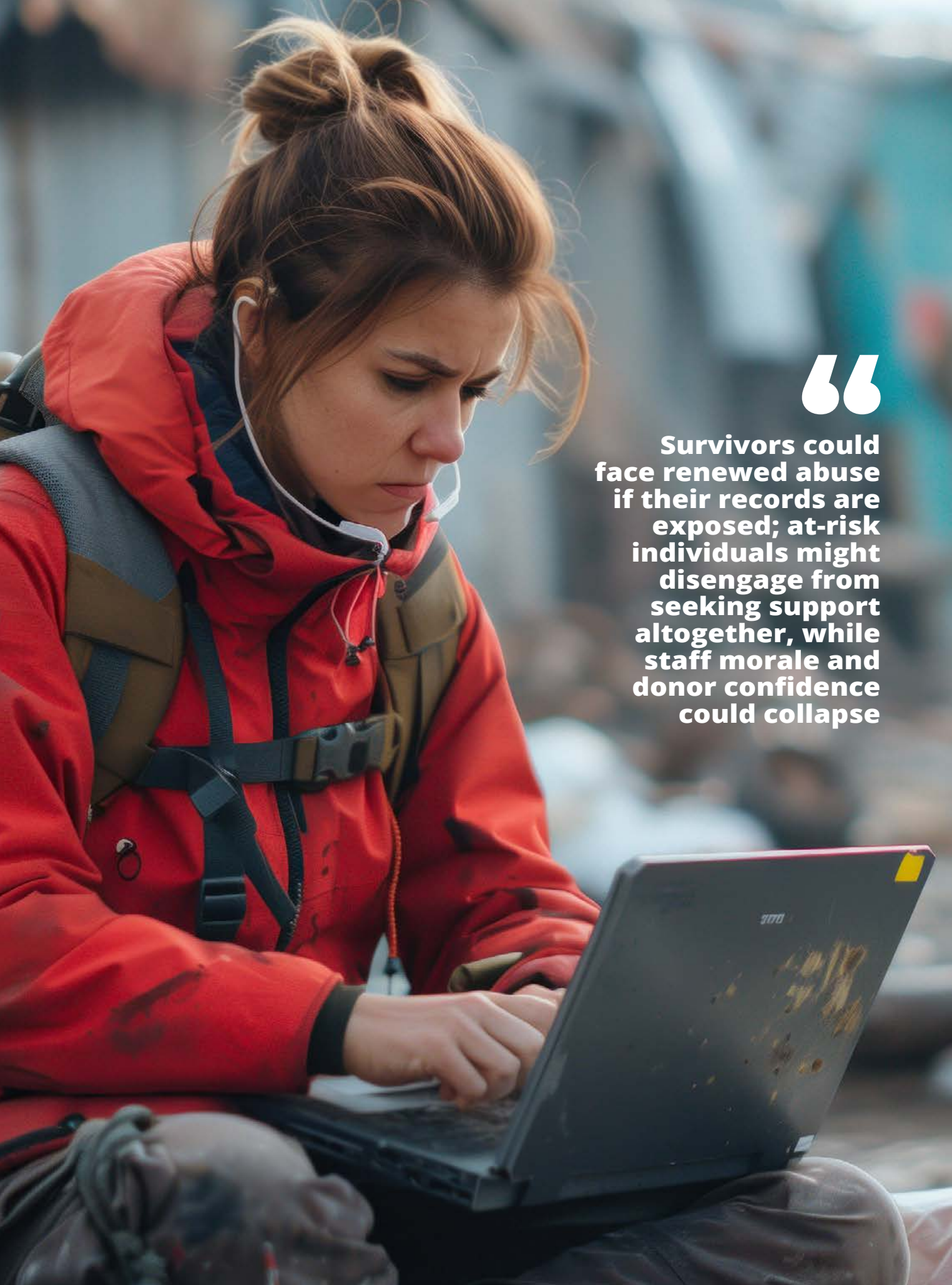
In 2023, a wave of public "persona" jailbreak prompts, most notably the "DAN" (Do Anything Now) prompt, spread rapidly across online communities. These prompts instructed conversational AI models like ChatGPT to assume an alternate persona supposedly freed from their normal content restrictions. By framing requests within this role-play, users were able to elicit disallowed or sensitive information that the models would ordinarily refuse to produce.
This form of jailbreak, driven by social prompt engineering rather than technical exploitation, demonstrated how easily safety filters could be bypassed through clever manipulation of language.

It led to numerous documented incidents in which AI systems generated harmful, misleading, or prohibited content despite embedded safeguards.[3]

Why nonprofits are uniquely exposed

A large corporation hit by an AI breach may suffer reputational damage and financial loss, but it often has dedicated security teams and the resources to recover quickly and resume normal business. Nonprofits do not enjoy such resilience and business continuity. They frequently rely on donated or discounted tools that they cannot fully configure or monitor. Mission urgency often drives rapid adoption of new technologies, leaving little room for deep security vetting. Threat actors, and in particular

3 Taylor, J. (2023), ChatGPT's alter ego, Dan: users jailbreak AI program to get around ethical safeguards, The Guardian. https://www. theguardian.com/technology/2023/mar/08/ chatgpt-alter-ego-dan-users-jailbreak-ai-program-to-get-around-ethical-safeguards



Survivors could face renewed abuse if their records are exposed; at-risk individuals might disengage from seeking support altogether, while staff morale and donor confidence could collapse

cybercriminals, see nonprofits as rich targets for three intertwined reasons:

1. Because of what they do: nonprofits often fill crucial gaps in social services, and as visible agents of change, they are frequent targets for actors who wish to discredit or neutralize them.
2. Because of what they know: these organizations collect and manage sensitive data about vulnerable populations, local networks, and operational plans, making them attractive to attackers seeking to steal, extort, or weaponize that information.
3. Because of what they have: nonprofits do raise funds, yet these funds are restricted for project delivery providing lifesaving support to millions of vulnerable individuals.

Only a very small part of a nonprofit's budget is invested in cybersecurity personnel and tools.[4] Therefore, this combination of high-value data, limited control, and stretched resources creates a dangerous gap. Since 2018, there have been over 590,000 threats impacting the nonprofit sector.[5]

As an example, consider a nonprofit operating a digital counseling platform for survivors of gender-based violence. To extend its reach, the organization deploys a custom AI system that screens initial requests, offers guided self-help resources, and directs urgent cases to human counselors. The model is trained on sensitive intake data, including narratives of abuse, geolocation details, and mental health assessments. A jailbreak attack against this system could unfold in several ways:

1. By tricking the AI into revealing snippets of its training data, an adversary could extract private testimonies, placing survivors at risk of being re-identified or located.
2. Threat actors might coerce the model into delivering

4 CyberPeace Institute (2023), CyberPeace Analytical Report:NGOs serving Humanity at risk: Cyber Threats affecting "International Geneva". https://geneva.cyberpeace.ngo/

5 CyberPeace Institute (2025), CyberPeace Tracer. https://cyberpeacetracer.ngo/

harmful instructions - encouraging self-harm instead of coping strategies, for example - undermining trust and endangering lives.

3. A successful attack could disable triage logic, flooding human counselors with false 'urgent' cases and delaying help for those in genuine crisis.

For the nonprofit sector, as in this example, the consequences extend far beyond reputational damage. Survivors could face renewed abuse if their records are exposed; at-risk individuals might disengage from seeking support altogether, while staff morale and donor confidence could collapse.

"

Since 2018, there have been over 590,000 threats impacting the nonprofit sector

Threat detection, policy and responsibility in the AI era

Defending against jailbreaking requires tools and methods different from traditional network monitoring. It means stress-testing models through continuous prompt injection simulations (deliberate attempts to trick the AI into ignoring its safeguards), building red-teaming exercises specific to AI (structured tests to mimic attacker strategies), and watching for patterns in model behavior that may indicate tampering.

For nonprofits, this is rarely achievable alone. As such, there is a pressing need for:

- Secure communication channels for sharing jailbreak techniques and alerts on AI zero-day vulnerabilities, as well as mitigation strategies across trusted partners, both from the private tech sector and nonprofit organizations.

- Safe spaces where nonprofit organizations can test their models against advanced attacks without risking live data.

- Investment in staff AI literacy to bridge mission delivery with AI risk awareness, even if they are not full-time security professionals.

Without these measures, the nonprofit sector risks falling further behind in a rapidly evolving threat landscape. In this context, the CyberPeace Builders program from the CyberPeace Institute is a vital initiative connecting nonprofits with pro bono cybersecurity and AI experts who provide practical support, such as cyber hygiene training, AI security awareness, and technical assistance.[6] Through this initiative, organizations can also access tailored resources on AI skills training for staff, AI responsible use, and policy templates - helping them strengthen digital resilience without diverting scarce funds from their missions.[7] Moreover, technical solutions are only part of the answer. Procurement policies must explicitly require AI security reviews, data minimization practices, and clear incident response plans that also include AI threat scenarios.

> **In cybersecurity, reacting is costly; anticipating is essential**

Since AI systems are built and deployed across borders, harmonizing standards internationally is essential - so that a frontline nonprofit in Nairobi benefits from the same safeguards as one in New York. Technology providers must also bear responsibility. The nonprofit sector should not be left with "lite" or "free" versions of AI tools that strip away critical protections. Providing secure, fully-featured systems should be considered a core ethical obligation.

Staying ahead

In cybersecurity, reacting is costly; anticipating is essential. For nonprofits working with sensitive data, staying ahead of AI jailbreak techniques is part of their duty of care. That means ongoing investment in staff training, governance, and partnerships that cut across sectors and borders. Jailbreaking and exploiting potential zero-day vulnerabilities in nonprofit AI systems is a reminder that, as these tools become smarter, defenses must become sharper.

In spaces where digital compromise can cause real-world harm, particularly to those least able to protect themselves, there is no room for complacency.

Security here is not just about protecting systems - it is about protecting lives.

ABOUT THE AUTHORS

Fabien Leimgruber is the Head of Cyber Resilience at the CyberPeace Institute. He joined the Institute in 2020, where he works to strengthen the cyber resilience of public-interest organizations and oversees the expansion of the CyberPeace Builders programme. Fabien has experience across the public, private, and nonprofit sectors. He previously led the Cyber Threat Intelligence Unit at Kudelski Security and later served as Information Security Awareness Adviser at the ICRC. He holds a Master of Law in Criminality and Information Security from the University of Lausanne.

Alexandru Lazar is a Nonprofit Cyber Resilience Specialist at the CyberPeace Institute, which he joined in 2021. At the Institute, he supports nonprofits in strengthening their cybersecurity resilience and helps coordinate the CyberPeace Builders, a network of corporate volunteers assisting public-interest organizations. He holds an Erasmus Mundus International Master's in Security, Intelligence and Strategic Studies from the University of Glasgow, as well as an MA in Politics and International Relations from the University of Aberdeen.

6 CyberPeace Institute (2025), CyberPeace Builders https://cpb.ngo/

7 CyberPeace Institute (2025), AI Skills for Nonprofits: A Collective Effort. https://cyberpeaceinstitute.org/news/ai-skills-for-nonprofits-a-collective-effort/

United Nations
Interregional Crime and Justice
Research Institute

Publications

Download UNICRI publications