hackerone

hackerone.com

April 16, 2024

The Honorable Lisa O. Monaco
Deputy Attorney General of the United States

Ms. Nicole Argentieri
Principal Deputy Assistant Attorney General, Criminal Division

Mr. John T. Lynch, Jr.
Chief, Computer Crime and Intellectual Property Section

U.S. Department of Justice
950 Pennsylvania Avenue, NW
Washington, DC 20530-0001

Dear Deputy Attorney General Monaco, Principal Deputy Assistant Attorney General Argentieri and Chief Lynch,

Subject: Request to Develop a Charging Policy under the CFAA to Protect Independent AI Trustworthiness Research

HackerOne is writing to respectfully request that the Department of Justice (DOJ) explore developing a charging policy to protect independent good-faith artificial intelligence (AI) red teaming or trustworthiness testing. The DOJ deserves credit for its support of good-faith security research protections under the Computer Fraud and Abuse Act (CFAA) and Section 1201 of the Digital Millennium Copyright Act. The DOJ has also commented in support of our effort to extend protections under Section 1201 to good-faith researchers who test generative AI models to expose and address biases in these models.[1] We urge the Department to now take steps to extend similar protections to good-faith AI trustworthiness research.

Although the DOJ helpfully issued a CFAA charging policy to decline prosecution of good-faith security research, this charging policy is focused on research performed for security purposes. Yet the emergence of AI has established a need to test systems for a broader range of trustworthiness metrics and potential harms beyond security – such as bias, discrimination, toxic content, misinformation, and other algorithmic flaws. This creates a potential gap in protections for independent researchers engaged in AI red teaming for non-security purposes.

HackerOne believes the contributions of independent good-faith research are a fundamental driver of holistic improvements to the technology ecosystem. Independent AI testing, or AI red teaming, serves a

[1] Letter from John T. Lynch to Suzanne Wilson, April 15, 2024 (available at https://www.copyright.gov/1201/2024/USCO-letters/Letter%20from%20Department%20of%20Justice%20Criminal%20Division.pdf).

key role in helping to maintain robust, equitable, and trustworthy AI systems. Such practices supplement the efforts of developers to test AI systems, deploy AI responsibly, and promote transparency, accountability, and safety in AI technologies.

Accordingly, we recommend that the DOJ consider the following measures:

1. Align the DOJ's CFAA charging policy to clearly include protections for good-faith AI red teaming for non-security harms. This should help ensure that individuals and organizations engaged in such research are not subject to prosecution under the CFAA, provided their actions are consistent with principles of good-faith research. Such a charging policy also serves to clarify that activity inconsistent with good faith, such as extortion or exfiltrating large volumes of data, continues to be restricted.

2. Encourage collaboration between the government, private sector, and civil society to develop consistent expectations and processes for independent AI red teaming. This may include community awareness and education about the role of good-faith research in strengthening trust, the boundaries of good-faith behavior, and promoting adoption of coordinated disclosure processes for algorithmic flaws.

3. Adapt existing legal frameworks for cybersecurity information sharing to facilitate the exchange of information regarding AI system trustworthiness. This would enhance collective understanding and response to AI risks beyond cybersecurity.

We appreciate the DOJ's ongoing commitment to protecting security research and urge similar protections for AI trustworthiness research. Thank you for considering this request.

Sincerely,

Ilona Cohen

Ilona Cohen
Chief Legal and Policy Officer
HackerOne