To: Harley Geiger, Hacking Policy Council
Cc: US Copyright Office
From: Tim Boucher, Online Trust & Safety Professional
Date: 6 August 2024

On the Need for DMCA Exemptions for AI Red Teaming

As a professional online Trust & Safety researcher with expertise in Generative AI (see my prior submission on this topic, as part of the Ad Hoc Group of Artists Using Generative AI), I strongly urge the Copyright Office to adopt the DMCA Section 1201 exemptions proposed by the Hacking Policy Council regarding red teaming of AI systems for harms outside those of security. This section of the DMCA, in its present form, provides inadequate legal protections for independent researchers such as myself who may in good faith discover and disclose issues in artificial intelligence systems, especially in bias, discrimination, or the generation of toxic or non-consensual content, as in the case I document below. This lack of strong clear legal safe harbor for researchers such as myself has a real chilling effect on this work, disincentivizing essential AI red teaming research, and leaving these systems and their users less safe and less well-served.

Six months ago, I discovered a reproducible flaw in a major image generation system's latest model release, whereby the system would consistently produce non-consensual nude images in seemingly unlimited quantities, against the company's own Terms of Service. The flaw relates to inadequate technical guardrails, ineffective input/output filters, and content restrictions that are easily jail-broken by using semantically adjacent allowed concepts in text prompts (e.g., "beach party" instead of "nude"), and then requesting variations of the output images. This problem is potentially easy to exploit maliciously using uploaded pictures of private or public individuals to create targeted malicious deepfake nude images.

Given that the company does not have a responsible disclosure program, nor a bug bounty program, nor any private means of contacting the company for such issues, I made the risky decision to document the nature and scope of the issue, and to publish my findings online. I strongly believe that conversations about the proper functioning of high-impact, high-risk generative AI systems needs to happen in public, not behind closed doors where companies can simply ignore reported issues. I knew this might be problematic under the company's Terms of Service, but I was unaware at the time that I was also potentially opening myself up to further risk under the DMCA. If I had been aware of that risk at the time, I would not have continued with the publication of my results.

Two weeks later, a journalist was able to reproduce the issue I identified, and published an article documenting the persistent problem. This increased public exposure resulted in the immediate suspension of my account by the company without any explanation, and no possibility of appeal. Shortly after, a second journalist was able to verify that, despite my account suspension, the problem persisted and no apparent corrective action had been taken by the company.

I am not able to continue this research, because I now understand that if I were to create a second account to verify whether it has been fixed with additional jail-breaking tests, I would be opening myself up to further potential liability under the DMCA for circumventing an account suspension. Further, now that I have better knowledge of the stipulations of the DMCA in this area, I am extremely reluctant to pursue similar AI red teaming investigations on either this platform (if my original account were reinstated), or any other platform where I might encounter issues of this nature.

Due to the growing ubiquity of AI and automated decision-making systems, I am extremely concerned about the chilling effects this has on AI red teaming efforts by outside researchers such as myself. It causes us to second-guess whether we ought to do the right thing and disclose the issue for the well-being of everyone, or stay silent about our findings in fear of negative legal consequences to ourselves. Thus, I again urge the Copyright Office to adopt the DMCA Section 1201 exemptions proposed by the Hacking Policy Council for AI red teaming outside of purely security areas.

Written by

Tim Boucher
https://timboucher.ca/about
6 August 2024