Evaluating Privacy Risks of Text-to-Image Generative Models

Executive Summary

Aadyaa Maddi
Swadhin Routray
Yifeng Zeng
Zili Zhou

December 12, 2023

1 Introduction

The development of Generative Artificial Intelligence (Gen AI) has led to its widespread use across various digital industries by businesses and consumers. Consequently, a growing need exists to educate users, especially minors and their guardians, about the risks and adverse effects of using Gen AI software. This project addresses privacy concerns about data collection and usage in Gen AI, particularly in scenarios where teenagers generate and share photo-realistic images. Our main goal is to map out information flows from the prompter of the Gen AI tool to the end-user and identify potential privacy risks associated with each of these flows. We will draw ideas from existing regulations, policies, and frameworks and propose mitigation strategies to improve privacy protections. Our study will also include user feedback, obtained by conducting a thorough user study, to prioritize risks and select effective mitigation strategies. We hope that the work presented herein will help inform the development and deployment of Gen AI tools by platform operators and other providers of Gen AI technologies.

2 Problem Statement

Recent work has identified a range of threats with text-to-image generative models. These models can generate content that perpetuates existing stereotypes about "race, ethnicity, culture, gender, and social class" [1]. Models can be misused to create inappropriate content or infringe on copyrights [1]. Furthermore, they can be used to spread misinformation, which can adversely impact individuals, communities, and societies [1].

Our project extends this line of work by focusing on privacy threats to users, especially teenagers, who interact with platforms with text-to-image generative models. Privacy threat modeling is a technique that can help identify potential privacy threats to a system and then define mitigations to prevent their effects. It can also help platforms evaluate their current practices, prioritize threats, and identify areas for improvement throughout the software development life cycle.

With the growing use of text-to-image generative models on various platforms, it is crucial to continuously evaluate these systems to prevent privacy harm. Informed by existing privacy threat modeling frameworks and a user study we conducted to understand parental perspectives

of privacy concerns to teenagers, our project provides platforms with a comprehensive privacy risk evaluation framework for the responsible deployment of text-to-image generative models.

3 Methodology

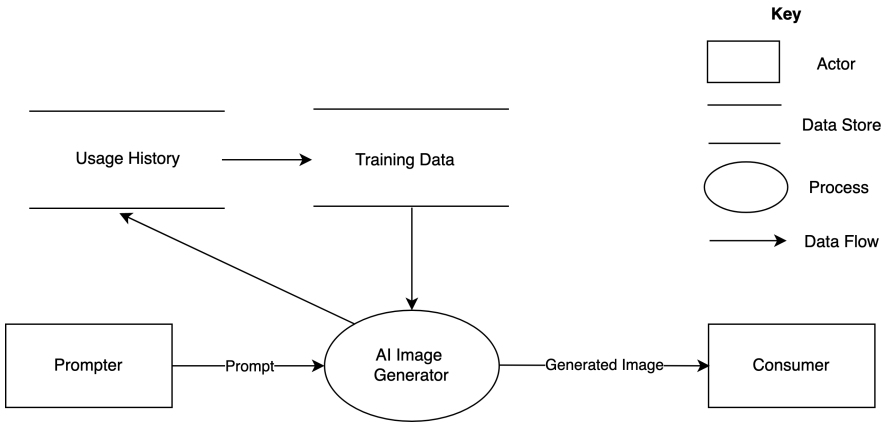

Figure 1: A Data Flow Diagram (DFD) of an example text-to-image generative model system under evaluation for privacy risks.

First, we model the system under evaluation by defining requirements or assumptions and breaking it into its constituent elements. Next, we determine elements (or groups of elements) and applicable stakeholders whose privacy can be impacted and map all possible threats to stakeholder-element(s) combinations. At this stage, creating scenarios or use case descriptions for each threat can help prioritize them by determining their level of impact and likelihood of occurrence.

In the final step, we assign mitigations to the threats and verify that these strategies are effective. In the report, we explain how we used these three steps to evaluate the privacy risks of text-to-image generative models. We consider the use case where teenagers are interacting with these models. Organizations or platforms that intend to deploy such models for other use cases can follow the same threat modeling process.

4 Interviews

To better understand the issues and concerns that parents and guardians have about their children's usage of new technologies, such as Gen AI, we conduct a user study. We conduct 60-minute semi-structured user interviews that aim to answer the three main research questions put forth in this paper. This user study was conducted after informed consent was obtained from the participants. Each interview was divided into two parts.

In the first part of the interview, participants were asked questions about their children, their children's activity on social media, gaming, and chat applications, such as Snapchat, TikTok, Discord, Steam, iMessage and Facebook among others, their understanding of privacy, and their concerns about their children's privacy online, along with a few questions that queried their knowledge of any online incidents that have occurred on these social media applications in their children's or their friend circle.

For the second half of the interview, participants were asked questions about their familiarity with new-age technologies, specifically, Gen AI. They were asked about their understanding of the term, their children's and their familiarity with text-to-image based generative models, their concerns associated with these models, from both their children's and their point of view, their knowledge of whether their children were using these Gen AI platforms, their knowledge of any incidents that involved the use of Gen AI, and their expectation of mitigations from these Gen AI models. This part of the interview also involved placing participants in scenarios that involved their children in vulnerable settings to see the different concerns that parents and guardians can identify. The interview concludes with a few questions about mitigations that the parents and guardians expect and want in the future from the Gen AI platform.

5 Results

We collect information to answer our research questions by conducting a within-subjects user study through 60-minute semi-structured interviews with parents of teens and children who are currently enrolled in middle or high school. We successfully validate our proposed risk evaluation framework through these interviews, since we find all user concerns and preferred mitigations are covered in our proposal.

Through our interviews and literature review, we outline some recommendations for various stakeholders. Platforms should educate their users about Gen AI models in their products. They should explain how these models work, how user data is used to train or improve them, and how they could be misused - in clear and accessible language. Furthermore, platforms should also implement infrastructure that allows users to submit complaints about misuse by other users, requests to view and manage their activity, and requests to opt out of the platform's data practices. When it comes to protecting children's privacy online, parents should have the option to make such requests on behalf of their kids. Platforms should ensure that the settings for Gen AI parental controls are similar to those already in place for existing parental control features. This will help users familiar with these settings navigate and use them efficiently, as with some of our participants who had previously used parental control software.

6 Acknowledgement

Prof. Norman Sadeh Professor, Co-Director Privacy Engineering Program, Carnegie Mellon University.

Dr. Tiffany Hsu Quantitative UX Researcher, Meta.

References

[1] C. Bird, E. Ungless, and A. Kasirzadeh, "Typology of risks of generative text-to-image models," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 396–410, 2023.