[redacted]

CST 300 Writing Lab
9 October 2024

The Ethics of GenAI:
Generative AI and Images of Children

As AI technology has become more prevalent than ever, the generative aspect of AI that allows users to create images and videos had significantly evolved. It now allows creation of highly realistic images, which in some instances are hard to distinguish from real ones. These features of Generative AI can be also used to take a photograph of a real person and imbed it into other images or videos, creating a highly realistic visual image. The effect of Generative AI as it pertains to images or photographs of real children creates an ethical concern due to how it affects child online anti-exploitation and safety efforts by federal and state law enforcement agencies across multiple social media platforms. According to a recent research conducted by Stanford University Internet Observatory, the proliferation of AI-generated CSAM presents a significant challenge due to a situation when it "generates hundreds of thousands of reports to online social platforms, NGOs handling CSAM cases, and law enforcement, thus overloading the ability of companies and organizations to effectively handle reporting and investigations" (Stroebel et. al, 2023).

The capabilities of Generative AI technology in merging an existing photograph of a real person into explicit content, creates additional challenge for law enforcement in determining whether the depicted abuse in fact contains a real child. The report also indicates a significant threat that the spread of AI-generated CSAM poses due to capability of Machine Learning-based models to be trained from already existing real depictions of abuse, allowing to generate CSAM of previously victimized children, thereby increasing the distribution of such material.

The technological origin of this issue arises from the creation of the so-called "Stable-Diffusion" models, which deploy a generative artificial intelligence technology that enables the production of photorealistic images from text and image prompts. Diffusion models work by breaking down images into noise and learning how these images "diffuse", then reserving the deconstruction process to rebuild a new image, deriving from the content learned from fed datasets. The most prominent examples of such models are Stability's Stable Diffusion, OpenAI's DALL-E, Google's Imagen and Midjourney (Stryker, 2024). The issues began when it was discovered that the original releases of some of these diffusion models did not include proper filtering, vetting or any limitation of materials that was used to train them. As reported by investigative researcher Eryk Salvaggio in the Tech Policy article titled "LAION-5B, Stable Diffusion 1.5, and the Original Sin of Generative AI", the contents of the datasets that were used to train Stable Diffusion and Midjourney came from the LAION-5B dataset, which is a collection of captions and links to 2.3 billion images (Salvaggio, 2023). Upon analyzing the content of the LAION-5B dataset, the researcher at Stanford University Internet Observatory, David Thiel, discovered that it contained more than 1,000 URLS containing verified images of CSAM. The presence of material that depict exploitation of children in a dataset used to train Generative AI models is not only alarming, but it is also indicative of a glaring gap in the policy making space where the lawmakers are increasingly running behind on emerging technologies.

Despite the follow-on versions of Stable Diffusion, including 2.0, making changes in its filtering and anti-abuse configurations to improve safety measures, the original Stable Diffusion version 1.5 remains in use and it remains capable of creating CSE images of real children, including previously victimized. This continues to pose a challenge to law enforcement agencies and the organizations such as NCMEC (National Center For Missing and Exploited Children).

The stakeholders in the first group regarding this issue include state and federal law enforcement, NGOs such as Thorn and NCMEC (National Center For Missing and Exploited Children), the victimized children and their advocates. The stakeholders in the second group are the producers, distributors and consumers of the Generative-AI made images and videos containing children, those who utilize photographs of real children to produce Stable Diffusion-made CSAM, who as group altogether position themselves are proponents of the 1st Amendment's right to free speech and expression, as it relates to creation of GenAI-made realistic CSAM.

The first group of stakeholders expresses their values as rooted in the protection of children's safety and is focused on conducting identification, localization, search and rescue operations, as it pertains to the children identified in the CSAM images and videos. This group of stakeholders is focused on combating human trafficking, online exploitation, extortion and other forms of crimes against children. The position that this group takes on the issue is a call for better regulation of the AI-generated CSAM, as it poses a threat to overwhelm the ability of law enforcement to combat exploitation crimes across the nation-wide network of agencies that comprise ICAC (Internet Crimes Against Children) Task Force. The claims that this group of stakeholders utilizes is a claim of policy, where it argues that the existing law *18 U.S.C. § 2256(8)* is establishing a framework for prosecution of such acts that depict realistic imagery, to include computer-generated imagery of explicit material of a minor. U.S. Code § 2252A1(8) prohibits any visual depiction of CSA that is virtually indistinguishable from a minor engaging in sexual conduct, with its definition specifically referencing computer-generated material as being in scope (Stroebel et. al, 2023). The policy claim of the first stakeholder group further indicates that there is more that the policy-making bodies can do in order to address the influx of AI-generated CSAM.

The second group of stakeholders expresses their values as based in their belief that the 1st Amendment rights to free speech and self-expression are indicative of their ability to utilize the Generative AI tools to produce, distribute and consume Stable Diffusion-made CSAM, due to it being a form of art expression. This group of stakeholders argues that the producers, distributors and consumers of AI-generated CSAM should not be constrained in utilizing the photographs of real children to create such images and videos, due to the resulting images being "virtual" and not real, as they do not consider it harmful while not inflicting actual physical abuse. This group of stakeholders expresses their position as only considering harm if its physical, while dismissing the notion that utilization of photographs of real children can cause them mental harm or have other negative consequences for the children depicted in the distributed explicit images. This group raises a policy claim, referring to the court case *Ashcroft v. Free Speech Coalition, 535 U.S. 234* (2002), where the U.S. Supreme Court declined to uphold the law passed by the Congress (Child Pornography Prevention Act of 1996), which banned depiction of explicit content that "appears to be" child sex abuse material. In its decision, the court noted that this issue may be revisited in the future as technology continues to evolve.

The question that arises out of the ethical issues surrounding the utilization of Generative AI for the purposes of production, distribution and consumption of CSAM is multi-faceted. First, there is a legal approach to this matter, with both sides claiming that they have policy and legal basis to support their position. Second, there is a question about a potential modification of an existing law, or passage of new legislation addressing the rise of AI-generated CSAM through Congress. And third, there is a possibility of action at the state level, with providing increased resources and support to the local law enforcement agencies in order to enable them to create better cooperation and collaboration programs with NGOs that specialize in ICAC investigations.

The ethical framework of Right-based ethics is applicable to the position of the first stakeholder group, as its founding principles align with this group's values and provide support for its goals. The Right-based ethical framework takes its origin from the Greek philosophers, particularly Socrates, and is focused on human rights-based approach, where a person maintains innate rights based on being human. The principles of this framework are further underlined in the Declaration of Independence that lists "inalienable rights" of each individual. Furthermore, the preamble of the Constitution states that one of its main purposes is to "establish Justice". The Rights based-ethical approach was further entrenched in the U.S. Department of Justice's Bill of Rights in code 18 U.S.C. § 3771, that was amended with the passage of Justice For All Act of 2004 (JFA), where the rights of victims had been assigned into law. The position of the first stakeholder groups is that the rights of the children subject to the AI-generated CSAM shall be upheld by ensuring that criminal justice system is adequately equipped with required resources, and through policy, in order to conduct identification and prosecution of CSAM distributors, and search and rescue operations for exploited children.

The main organization that receives and process reports from CyberTip Line in the United States is the National Center for Missing & Exploited Children (NCMEC). This organization was established in 1984 by the United States Congress and is primarily funded by the U.S. Department of Justice. The CyberTip Line is the nation's centralized reporting system for the online exploitation of children (NCMEC). Social media platforms like Instagram, TickTock and others are required by law (federal statute 18 U.S.C. § 2258), to report CSAM to CyberTip Line. In 2023 alone, "the CyberTipline received more than 36 million reports, nearly all from online platforms. Facebook, Instagram and Google were the companies that sent in the highest number of reports. The overall number has been dramatically increasing" (O'Brien, 2024). According to the research

conducted by the Stanford University Internet Observatory, concerning the resilience of the current reporting system against the effects of AI-generated CSAM, "in the years to come, the CyberTip Line will just be flooded with highly realistic-looking AI content, which is going to make it even harder for law enforcement to identify children who need to be rescued" (Grossman, 2024).

According the to the first stakeholder's group position, the course of action that constitutes strengthening the policies regarding distribution of AI-generated CSAM that uses images of real children, and providing more resources to state and federal law enforcement to combat child sexual exploitation online, is a correct course of action because it will strengthen the capabilities of criminal justice system in the face of advancing technology, and particularly technology-facilitated image-based abuses. This course of action seeks to alleviate the dangerous impediment that the overwhelming of reporting system creates, while helping law enforcement to maintain its focus on identifying and rescuing victims, prosecuting offenders and prioritizing the safety of the children.

The ethical framework of Individual relativism can be utilized in support of the second stakeholder's group position. The ethical framework of Individual relativism, also referred to as personal relativism, emphasizes the moral judgments of actions at the individual level, as guided by the individual's personal ideology. The origin of the relativism can be found in Greek philosopher Protagoras (Carter et al, 2022). The relativism infers that rather than a defined right or wrong, the morality is fluid and is depended on the individual's views and circumstances. The position of the second stakeholder group is expressed in their views of what they personally consider to be harmful and what they do not consider to be harmful. The second stakeholder group expresses their view that because the AI-generated CSAM does not incur physical abuse of children, even when the photographs of real children or real past victims being used in the process of its creation, that means that the material generated is deemed not harmful. The position of this

stakeholder group does not view mental harm to the children subject to AI-generated CSAM as a "harm" per se, or as a factor that needs to be considered, thus eliminating the need to address it as a part of their overall argument. The course of action that the second stakeholder group is arguing for consists of strengthening the rights of individual producers, distributors and consumers of AI-generated CSAM, by arguing against federal or state law enforcement measures in each individual criminal court case. This course of action is to be done with reliance on the interpretation of freedom of speech doctrine as supporting the creation and distribution of AI-generated CSAM. This course of action is seen as correct by the second stakeholder group as it will allow this group of stakeholders to continue creation of material that depicts realistic-looking images and videos of exploitation of minors with no legal consequences, thereby allowing the producers, distributors and consumers of AI-generated CSAM to obtain personal and financial gain from this material.

In accordance with the positions of the two groups of stakeholders described above, the position that the author gravitates towards is the position of law enforcement agencies in the first group of stakeholders. As the AI technology becomes more advanced and is ultimately on the path of becoming ubiquitous in human society, the ethical considerations of its effect on the safety of the community is of utmost importance. With the advent of technology-facilitated crimes, such as cybercrimes, that include threat actors utilizing computer devices to perpetuate harmful activities, there is an inherent risk that comes with new technology. This inherent risk represents an abuse of technology by the individuals that results in the harm of others. The author's position is in line with the notion that anytime a new technology comes about, there is a need for assessment of its potential for abuse and a need for discussion on policy implementations that mitigate that potential for abuse. Interestingly, this position is supported by the U.S. Supreme Court opinion in the very exact court case that the opposite side of the augment claims as its support - *Ashcroft v. Free*

Speech Coalition. Three justices on that case indicated that "rapidly advancing technology" may change the Court's holding that the Congress was required to "wait for the harm to occur before it can legislate against it" (Baughman, 2024). Indeed, this principle of harm prevention can be observed in many industries outside information technology, including the public transportation industry. The implementation of guardrails to mitigate human harm outcomes is common place. For instance, when it is known that if a human driver in a vehicle that carries children in it abuses the red light and goes onto the train tracks, the children will be harmed as a result. The constructors don't wait to for the harm to occur to erect the barricades that automatically go down to block the entrance of cars across the train track anytime the train is coming by.

When law enforcement rescue operations are not taking place due to the reporting systems being overwhelmed by the influx of AI-generated CSAM, the children are being harmed. When a child is being driven to self-harm or taking of one's own life by being threatened or extorted with the explicit AI-generated images of them being distributed to their family, a child is being harmed. The author is questioning how many children have to be harmed by the reckless proliferation of AI-generated CSAM in order for adequate guardrails to be implemented? As long as the lack of policy continues to allow the disruption of law enforcement and NMEC's CyberTip Line mission, by letting the AI-generated CSAM to flood the report lines, investigations and rescue operations will be delayed, resulting in irreversible outcomes. This is why the author supports the position of the first group of stakeholders and the mitigation measures that this group advocates for. Such mitigation measures include improving policy by strengthening laws, allocating more resources to federal and local law enforcement units that are a part of the ICAC Task Force, as well as providing funding to modernize the CyberTip Line systems to allow for a more streamlined collection, processing and analysis of the information coming from the online technology platforms.

References

Baughman, C. (2024). Deepfake culture and the 1st Amendment. *Daily Journal.* https://www.dailyjournal.com/articles/377089-will-scotus-reconsider

O'Brien, M., Ortutay, B. (2024, April 23). Child exploitation CyberTipline needs fixes before AI makes it worse. *Press Herald*. https://www.pressherald.com/2024/04/22/report-child-exploitation-cybertipline-needs-fixes-before-ai-makes-it-worse/

Carter, A., Baghramian, M. (2022). Relativism. *The Stanford Encyclopedia of Philosophy* https://plato.stanford.edu/archives/spr2022/entries/relativism/

Grossman, S. (2024). *Stanford Internet Observatory's CyberTipLine Report | Stanford Law School*. Stanford Law School. https://law.stanford.edu/podcasts/stanford-internet-observatorys-cybertipline-report/

Thiel, D., Stroebel, M., & Portnoff, R. (2023). Generative ML and CSAM: Implications and mitigations. *Stanford Digital Repository*. https://doi.org/10.25740/jv206yg3793

Thiel, D. (2023). Identifying and eliminating CSAM in generative ML training data and models. *Stanford Digital Repository*. https://doi.org/10.25740/kh752sm9123

Salvaggio, E. (2024, January 2). *LAION-5B, stable diffusion 1.5, and the original sin of generative AI*. Tech Policy Press. https://www.techpolicy.press/laion5b-stable-diffusion-and-the-original-sin-of-generative-ai/

Stryker, L . (2024, September 12). *What are Diffusion Models? | IBM*. (2024, September 12). https://www.ibm.com/think/topics/diffusion-models