# AI, ETHICS, AND LEGAL LOOPHOLES

## THE STEVEN ANDEREGG CASE AS A WAKE-UP CALL

Zhouchi Wang |29/3/2025

# Introduction

The rapid advancement of artificial intelligence has changed many fields. These fields include creative industries, medicine, and security. However, it has also brought up new ethical and legal problems, especially in the area of generative AI. This paper looks into the misuse of AI-generated imagery in the case of Steven Anderegg. He is a software engineer. He used the Stable Diffusion model to create over 13,000 extremely realistic images of child sexual abuse material (CSAM) (Jonathan Limehouse, 2024).

This case is a really important turning point in the discussion about AI governance. It shows the risks that publicly available generative models can bring when they are used for unethical and illegal purposes. The investigation focuses on three main concerns. First, there are technical vulnerabilities in the moderation of AI-generated content. Second, there are ethical implications of synthetic CSAM. Third, there are legal loopholes that make prosecution difficult.

The significance of this topic is really important for computing professionals, law enforcement, and policymakers. AI researchers and developers need to understand that they have an ethical responsibility. They should design and deploy generative models that can't be easily abused. For example, there was an open - source AI model called Stable Diffusion. It was manipulated to get around safety filters and create explicit images. This shows there are serious problems in AI system governance. (James Liddell, 2024)

Unlike traditional image manipulation techniques, diffusion models let users generate new, photo - realistic content according to textual prompts. This creates significant challenges for content moderation (Matthias Schneider & Thilo Hagendorff, 2024). There are no embedded protection mechanisms. Because of this, individuals like Anderegg can use AI for criminal activities. This raises urgent concerns about the responsibility of developers and the need for strong safeguards.

Moreover, the legal framework around synthetic CSAM is still underdeveloped. Traditional CSAM laws, like the ones in 18 U.S.C. §2256 in the United States, mainly focus on prosecuting crimes involving real children (Chad M.S. Steel, 2024). However, AI - generated content is in a legal gray area. This is because it doesn't show actual victims, and this makes prosecution more complicated (Wael Badawy, 2025).

The case of Anderegg shows that synthetic CSAM can still cause harm. It normalizes exploitation, provides material for offenders, and makes it harder for law enforcement to identify and protect real victims. Research shows that even purely artificial CSAM causes psychological harm. It creates environments that allow child predators to act (Chad M. S. Steel, 2024; Wael Badawy, 2025). This emphasizes the need for legislative updates.

These updates should explicitly criminalize AI-generated CSAM to make sure the legal system can keep up with technological advancements.

Beyond legal concerns, the case of AI-generated child sexual abuse material (CSAM) has broader societal implications. The democratization of AI image generation tools has made it easier for individuals with bad intentions to create CSAM.

You can produce highly realistic explicit content with minimal technical expertise. This erodes traditional safeguards against exploitation (Amy Trivison, 2024). Studies show that it's very challenging to monitor and detect synthetic CSAM. Law enforcement agencies have a hard time telling the difference between real and artificially generated content (Thiel, Stroebel & Portnoff, 2023). This makes the heavy burden on cybersecurity and child protection organizations even worse. We need innovative detection techniques and proactive intervention measures.

The structure of this paper uses a systematic way to look at the intersection of AI ethics, legal accountability, and technological vulnerabilities. In Section 1, I will give a detailed analysis of the Anderegg incident. I'll explore how Stable Diffusion was used in a wrong way to create CSAM. In Section 2, I'll dig into the technical parts of generative AI. I'll explain how diffusion models work and where their weak points are.

Section 3 links these technical factors to ethical concerns. Specifically, it focuses on accessibility, synthetic realism, and the use of unregulated training data. Section 4 deals with the legal challenges related to prosecuting AI - generated CSAM. It discusses jurisdictional gaps and suggests potential policy recommendations. Finally, the conclusion sums up the key findings and puts forward actionable steps to reduce the risks of generative AI misuse.

As AI capabilities keep getting better, it's really important that computing professionals, policymakers, and technology firms take proactive steps to prevent the misuse of AI. The Anderegg case is a warning. It shows that if technological innovation is not controlled, it can lead to unexpected ethical and legal problems. We can reduce the risks of AI-generated CSAM by putting in stricter safeguards, updating the laws, and promoting interdisciplinary collaboration. This way, we can make sure that AI development is in line with the well-being of society and ethical responsibility.

# Literature review

**Case Study Analysis: AI-Generated CSAM Through Stable Diffusion**

## 1. Case Study Description and Ethical Issues

Steven Anderegg is a 42 - year - old software engineer from Wisconsin. His arrest is a landmark case about the misuse of generative AI to produce hyper - realistic child sexual abuse material (CSAM). It is alleged that Anderegg used Stable Diffusion, which is an open - source AI image generator. He created over 13,000 explicit images depicting...

Prepubescent minors are shown in sexually explicit scenarios (Jonathan Limehouse, 2024). The images were made with tailored prompts. These prompts were detailed descriptions of abuse. Specialized add - ons were used to show genitalia (James Liddell, 2024). This case shows the ethical problem of trying to balance AI innovation and protecting people from exploitation.

The ethical issue here involves technical accountability, legal ambiguity, and societal harm. Even though no real children were directly abused, the realistic nature of the content makes victimization continue. It does this by making exploitation seem normal and might even help groomers. (Chad M. S. Steel, 2024) Also, the current U.S. laws are meant to protect real victims. But they have a hard time dealing with fully synthetic content. This creates gaps in jurisdiction. (Chad M. S. Steel, 2024)

## 2. AI Techniques Used in Case Study

StableDiffusion uses a latent diffusion model. This is a machine - learning technique. It refines noise into coherent images bit by bit based on text prompts. Earlier generative adversarial networks (GANs) are different. Diffusion models are really good at creating high-resolution images with detailed features. This makes them perfect for creating really realistic pictures.

Anderegg took advantage of this situation. He didn't just use negative prompting, which means giving clear instructions to remove clothing. He also used fine - tuning tools, which are custom add - ons that can bypass safety filters and generate anatomically accurate features (Jonathan Limehouse, 2024; Chad M.S. Steel, 2024).

The model is open - source, so Anderegg could manipulate it without any supervision. Later, Stability AI introduced some safeguards, like restricted access and content filters. But in this case, he managed to get around these safeguards (James Liddell, 2024).

## 3. Technical-Ethical Linkages

3.1 Accessibility versus Misuse

StableDiffusion is accessible because it's free, open - source, and you don't need much technical skill to use it. This makes it easy for people to misuse it. The fact that AI has become more accessible to the public means that it's easier for offenders to create child sexual abuse material (CSAM) on a large scale. For example, Anderegg produced 13,000 images on an industrial scale (Jonathan Limehouse, 2024). This also brings up ethical questions about the responsibility of developers. Should AI providers be held responsible for how people misuse their products later on? Or should only the users be accountable?

3.2 Synthetic Realism and Harm Amplification

The model can generate “hyper - realistic” images. This blurs the line between real and synthetic abuse. Research shows that even artificial CSAM causes harm. It makes pedophilic fantasies seem normal and makes it hard for law enforcement to find real victims (Chad M.S. Steel, 2024; Wael Badawy, 2025). Ethical frameworks need to figure out if synthetic content should have the same legal penalties as material with real children.

3.3 Consent and Data Ethics

While the training data of StableDiffusion is still unknown, its possible use of publicly scraped images, including photos of minors, goes against the principles of informed consent. A study by Wael Badawy in 2024 points out that children's images in AI datasets are at risk of being exploited if they aren't properly anonymized or removed. Anderegg's case indirectly shows the chain of harm: unregulated training data leads to tools that can be exploited, which then causes the spread of CSAM.

**4. Key Messages from Analysis**

Regulatory lag is still a very important problem. The legal frameworks that make child sexual abuse material (CSAM) a crime need to be updated. They should include synthetic content and get rid of the loopholes that protect offenders like Anderegg. This is noted in "Artificial intelligence and CSEM - A research agenda" by Chad M.S. Steel in 2024 and "The ethical implications of using children’s photographs in artificial intelligence: challenges and recommendations" by Wael Badawy in 2025.

Developer accountability is really important. AI firms need to take proactive steps. They should put in place things like built - in content filters and stricter access controls. They shouldn't just rely on reacting after abuse happens. A news story called "Wisconsin man is charged with making child pornography images using AI in first case in US" (James Liddell, 2024) points this out.

We must also prioritize ethical design principles in AI development. We should ensure privacy by design approaches. And we need to avoid using public datasets that contain images of minors. As The ethical implications of using children's photographs in artificial intelligence: challenges and recommendations (Wael Badawy, 2025) mentions. This case shows that if we don't control technological advancement, it may go faster than ethical governance.

Mitigating AI-driven CSAM needs collaboration among developers, legislators, and child protection agencies. They should align innovation with societal well - being.

# Problem statement – How, What, Why

**Problem Statement: The Ethical Collapse of AI Image Generation in the Steven Anderegg CSAM Case**

Steven Anderegg misused StableDiffusion to create hyper - realistic child sexual abuse material (CSAM). As a result, he faced federal prosecution (James Liddell, 2024). This case exposes the fundamental failures in AI ethics frameworks and technical safeguards.

This case shows that when three critical limitations come together, AI tools made for creative purposes can be used as weapons against vulnerable groups. These limitations are: not good enough content moderation mechanisms, open - source architectures that can be exploited, and legal frameworks that are out of date and were designed for analogera crimes against children (Amy Trivison, 2024).

**1. How Should AI Image Generation Work Ideally?**

Ethical AI image generators should balance creative freedom and strong protections against harmful outputs. As Wael Badawy proposed in the 2025 paper "The ethical implications of using children's photographs in artificial intelligence: challenges and recommendations," responsible systems must:

To effectively fight against AI-generated CSAM, we need a multi-layered approach. First, we should use classifiers to automatically filter CSAM-related prompts. These classifiers can analyze content, context, and user behavior. Second, we must set technical constraints to limit the anatomical realism of depictions involving minors. For example, we can implement latent-space blurring for genitalia. Third, we should put watermarks on AI-generated outputs so that we can trace them forensically.

Finally, the ability to customize things, like the LoRA add-ons that bad guys like Anderegg used, should be limited. We can do this by controlling API access. Jonathan Limehouse (2024) pointed this out.

There was a big difference between the ideal protections and what actually happened in practice. This allowed Anderegg to use StableDiffusion's open - source architecture to get around the commercial safeguards. He used specialized add - ons to create over 13,000 explicit images (Liddell, 2024).

**2. Ethical Issues Caused by Technical Limitations**

2.1. Moderation Failures:

Even though StableDiffusion can cause harm, none of the ten popular implementations that Matthias Schneider and Thilo Hagendorff studied in 2024 had NSFW filters. These systems perfectly carried out prompts like "naked young girls" (Amy Trivison, 2024) and "prepubescent children engaging in sexual acts" (Jonathan Limehouse, 2024). This goes against the ethical principle of non - maleficence (Wael Badawy, 2025).

2.2. Architectural Vulnerabilities:

StableDiffusion has an open - source design. This allowed Anderegg to get around safety measures in two important ways. First, he could remove safety classifiers. For example, he removed LAION’s NSFW filters. Second, he installed LoRA adapters. These adapters were specifically trained on child anatomy (Jonathan Limehouse, 2024). These changes show a big problem. Chad M.S. Steel (2024) pointed it out in “Artificial Intelligence and CSEM - A Research Agenda”. AI systems that are made for photorealism often don't have built - in child protection mechanisms. So, they can be easily misused.

2.3. Bias Amplification:

Anderegg's creations show broader AI bias patterns. In these patterns, 73% of violent AI - generated images unfairly depict minorities (Matthias Schneider & Thilo Hagendorff, 2024). Even if these are synthetic, they strengthen harmful stereotypes that affect real - world children.

**3. Why AI Techniques Generate Ethical Risks?**

The technical aspect shows that Stable Diffusion uses latent diffusion models to learn anatomical features from training data on its own. If there are no strict controls, users can take advantage of this. They can overfit body proportions through Dreambooth training

(Understanding AI's Line). Or they can combine multiple unsafe ideas, like "prepubescent" and "basement chains," through compositional generation.

The legal aspect poses challenges. Right now, U.S. law demands proof of real child harm for CSAM charges. This is based on the case of Ashcroft v. ACLU in 2004. Even so, 81% of citizens are in favor of banning synthetic CSAM, according to Chad M. S. Steel in 2024. Meanwhile, law enforcement officers have a hard time telling the difference between synthetic and real evidence under the current legal standards, as Chad M. S. Steel also mentioned in 2024.

The societal impact is quite big. When abuse becomes normal, it gives offenders more chances to keep doing bad things (Amy Trivison, 2024). Also, it's dangerous for minors because there are unauthorized "de-aging" of legal photos (Chad M.S. Steel, 2024).

4. Critical Need for Resolution

If we don't stop the production of synthetic CSAM, it will cause several serious threats. Monitoring systems might get overwhelmed. For example, there were over 20,000 AI - generated CSAM images.

According to Amy Trivison (2024), it is posted monthly on single forums. Victim identification becomes more complicated. As James Liddell (2024) noted, distinguishing between real and synthetic evidence can undermine prosecutions. Public trust is also at risk. According to a study in "The ethical implications of using children's photographs in artificial intelligence: challenges and recommendations" (Wael Badawy, 2025), 62% of parents express concerns about AI being misused to target children.

As Wael Badawy concludes, "ethical guidelines must evolve faster than generative capabilities." This requirement was clearly not met in the Anderegg case. To solve this problem, we need to reengineer AI architectures with built - in child safeguards instead of doing post-hoc content moderation.

# Conclusion and Discussion

### 1. Concise Summary of the Topic and Significance

The investigation is centered on the case of Steven Anderegg. He took advantage of Stable Diffusion's AI abilities to create hyper - realistic child sexual abuse material

(CSAM). This material included more than 13,000 explicit images of prepubescent minors (Jonathan Limehouse, 2024, James Liddell, 2024). This case shows how advanced generative AI systems, ethical problems, and legal weaknesses in regulating synthetic abuse content come together.

The incident directly points out three important issues. First, it's technically easy to bypass AI safeguards to create anatomically explicit material (Chad M.S. Steel, 2024). Second, the legal framework for prosecuting completely synthetic CSAM is unclear. Currently, it depends on proving harm to real children (Wael Badawy, 2025). Third, normalized exploitation causes harm to society and puts pressure on law enforcement (David Thiel, Melissa Stroebel & Rebecca Portnoff, 2023).

**2. Significance of Findings**

This case shows how urgent it is to deal with the gaps in AI governance. This is especially true when diffusion models like Stable Diffusion make it easy for people to create harmful content. The super - realistic synthetic CSAM makes it hard to identify victims. It also encourages grooming practices and may overload monitoring systems. (Chad M. S. Steel, 2024; James Liddell, 2024)

Technically, Stable Diffusion is an open - source thing. This allowed Anderegg to bypass safety filters. He used specialized add - ons, like LoRAs, and fine - tuned prompts. He used these to refine genitalia. (Jonathan Limehouse, 2024; Matthias Schneider & Thilo Hagendorff, 2024).

Ethically, public AI models lack refusal mechanisms. In the study "When Image Generation Goes Wrong: A Safety Analysis of Stable Diffusion Models" (Matthias Schneider, & Thilo Hagendorff, 2024), they observed 10 variants of Stable Diffusion. These models violate the non - maleficence principles. They allow harmful outputs to be produced without any resistance.

Legally, the prosecution of Anderegg under existing U.S. statutes shows a reactive way to deal with synthetic abuse. Laws like 18 U.S.C. §2256 focus on materials involving real children, not AI - generated content (Chad M.S. Steel, 2024). This difference shows that there is a need for legislative reforms. These reforms should make synthetic CSAM a crime, whether there are direct victims or not. This is a stance that advocacy groups and scholarly analyses support (Wael Badawy, 2025; Chad M.S. Steel, 2024).

**3. Implications for Future Action**

To mitigate risks, collaborative efforts should prioritize proactive safeguards. We need to embed multi - layered content filters into AI models. These filters include textual, contextual, and anatomical filters. By doing this, we can block CSAM - related prompts (Chad M. S. Steel, 2024). Also, legislative reform is essential. We should expand flawed laws. These expanded laws should explicitly criminalize synthetic CSAM and align with emerging global standards (Wael Badawy, 2025).

Additionally, we should adopt ethical design practices. We can do this by incorporating privacy - by - design principles. These principles help us to exclude minors' images from training datasets. Also, we need to implement forensic watermarking to ensure traceability (David Thiel, Melissa Stroebel & Rebecca Portnoff, 2023).

**4. Conclusion**

The Anderegg case ultimately acts as a clear call for rebalancing AI innovation with protections focused on children. As Wael Badawy said in the article "The ethical implications of using children's photographs in artificial intelligence: challenges and recommendations" in 2025, "ethical guidelines must develop faster than generative capabilities". This requirement was not met in this important misuse of Stable Diffusion.

# References

Jonathan Limehouse (2024, May 22). Man indicted after creating thousands of AI-generated child sex abuse images, prosecutors say. Retrieved from https://www.usatoday.com/story/news/nation/2024/05/22/steven-anderegg-arrest/73801269007/

James Liddell (2024, May 22). Wisconsin man is charged with making child pornography images using AI in first case in US. Retrieved from https://www.independent.co.uk/news/world/americas/ai-child-sex-wisconsin-anderegg-b2549615.html

Chad M.S. Steel (2024). Artificial intelligence and CSEM - A research agenda. *Child Protection and Practice*, 1–7. https://doi.org/10.1016/j.chipro.2024.100043.

Wael Badawy (2025). The ethical implications of using children's photographs in artificial intelligence: challenges and recommendations. *ORIGINAL RESEARCH*, 1–12. https://doi.org/10.1007/s43681-024-00615-2.

Amy Trivison (2024). Understanding the Line Between Art and Abuse: How Generative AI Changes the Landscape of Child Sexual Abuse Materials. *Catholic University Journal of Law and Technology, 33*, 88–114. https://scholarship.law.edu/jlt/vol33/iss1/6.

JUSTIA (2004). Ashcroft v. ACLU, 542 U.S. 656 (2004). Retrieved from https://supreme.justia.com/cases/federal/us/542/656/.

Matthias Schneider & Thilo Hagendorff. (2024). When Image Generation Goes Wrong: A Safety Analysis of Stable Diffusion Models. arxiv, 1–12. https://doi.org/10.48550/arXiv.2411.15516.

David Thiel, Melissa Stroebel & Rebecca Portnoff. (2023). Generative ML and CSAM: Implications and Mitigations, THORN Stanford Internet Observatory Cyber Policy Center, 1–20. https://stacks.stanford.edu/file/druid:jv206yg3793/20230624-sio-cg-csam-report.pdf.