A Study on the Ethics of AI Content Generation

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Gabriel Binning

Spring 2024

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Caitlin D. Wylie, Department of Engineering and Society

Since the origin of artificial intelligence (abbreviated as AI) in the 1950s with a maze-solving tool called Theseus, the computational power of AI tools has experienced exponential growth. Although, it is only recently that AI tools have started to outperform humans in a wide variety of tasks (Giattino, Mathieu, Samborska, & Roser, 2024). This has inspired organizations and individuals to experiment with these tools to solve problems and explore applications. The implementation of AI tools is fairly conspicuous for technical industries as it is commonly utilized to provide increased modeling and computational analysis capabilities (Giattino, Mathieu, Samborska, & Roser, 2024). However, for industries such as entertainment and media, the role of AI tools is more varied and obscure. AI for content creation is defined as using an AI tool to create text, images, or video with the intent to distribute the result for mass consumption. This can be content distributed on social media platforms, audio and video streaming services, and more traditional or physical media such as television or posters. AI for content generation is primarily done for harmless entertainment purposes but there is an abundance of ways these tools can be misused or mismanaged, creating ethical dilemmas between the creators of the AI tools and people who are affected by AI-generated media. These tools show significant promise for revolutionizing how people work, but these powerful tools lack significant regulation. This has left the creators of these AI tools to build preventive measures within their own products. With the lack of regulation, the creation of the generated content can create unethical situations. I will specifically focus on misuse, data sourcing, and the effect of taking the media creation processes out of the hands of creatives. My claim is that AI content generation can be ethical, however, significant considerations and restrictions must be made for the use of AI for content generation in media and entertainment to be ethical. I will do this within the frameworks of utilitarianism and deontology, where outcomes and intentions are

paramount. Utilitarianism is the ethical theory where an action should be analyzed by its effects with the goal of maximizing favorable outcomes and minimizing unfavorable ones (Santa Clara University, 2014). Deontology is the ethical theory where an action should be analyzed by its intentions with the justification of an action due to internal moral principles (Stahl, 2021).

The two primary ways that AI for content generation can be misused are through the creation of misinformation and generally harmful content. In 2020 a Denmark team created a "game" where participants were shown 10 images and texts and had to guess if they were created by a human or an AI (Partadiredja, Serrano, & Ljubenkov, 2020). 2383 participants played this game and the average correct score was 5/10 ~ 50%. This study demonstrated that people have a hard time, essentially equivalent to a coin flip, determining whether the media provided to them is AI or human-generated. They also found that "there seems to be neither positive nor negative correlation between the participants' time spent in playing the game with their score." This finding indicates that a more in-depth analysis of media does not directly equate to a better chance of determining whether the creators are human or AI. This particular study was done in 2020 and since then the quality of AI content generation has drastically increased, now being able to create photorealistic minute-long videos. These forms of generated media for the most part are harmless fun and can also be useful tools for content creators. However, as generated media becomes nearly indistinguishable from human-created media it can also allow the creation of believable misinformation.

Misinformation by itself is nothing new. One well-documented incident of misinformation was during the 2016 U.S. election where Russia spent over a million dollars a month to create and spread fake images and articles on social media to sway the election in their favor (Marcus, 2024). This is a relatively insignificant amount of money for a global superpower

but would be entirely impossible for someone working alone to achieve back then. Now the creation of believable misinformation can be offloaded to a content-generating AI which can be free to use. This is already being done, with the number of websites hosting AI-generated articles increasing from 49 in May of 2023 to 777 in March of 2024 according to NewsGuard, a website that tracks AI-enabled misinformation sources. The purposes of these websites vary from "intending to sway political beliefs or wreak havoc" to "draw clicks and capture ad revenue" (Verma, 2023). There is minimal regulation of these types of websites as the U.S. explicitly protects the freedoms of the press and speech and even if there were more extensive regulations put in place, it would take more time to shut down these websites than it takes for them to be created (Verma et al., 2023). Although the intentions might differ between websites, the mass production of AI-generated misinformation is inherently unethical as it causes people to be swayed or engrained in their beliefs due to false narratives.

Another unethical facet of the content produced by AI generative models is the creation of generally unsightly or horrific content, primarily categorized as pornography and gore. These forms of media have a history of being withheld from the general public without proper approval. This practice is evident in traditional media, where movies and television are given ratings based on the age group that the content is appropriate for, and in social media where posts that include these categories can result in bans or censoring. Most popular AI media models already have built-in filters to prevent any of these types of content from being created, however, certain AI models are being created for the distinct purpose of creating these forms of harmful content (Heikkilä, 2023; Wiggers & Silberling, 2022). Most of these AI models are taking the fully generative approach, allowing the unfiltered creation of disturbing images from text. A group called Unstable Diffusion created an AI based on Stability, an image-generating AI model,

to create a discord bot that could generate porn/gore images on request through a text command (Wiggers & Silberling et al., 2022). Other models are creating more niche applications, allowing users to input images of a specific person and generate harmful content with the provided person's likeness (Hao, 2022). A research company, Sensity AI, estimates, "between 90% and 95% of all online deepfake videos are nonconsensual porn, and around 90% of those feature women". Using AI to generate harmful content, especially when that content is supposed to include a real person is morally indefensible through the framework of utilitarianism. This is due to the vast negative societal consequences, such as a reduced sense of privacy and an increased exposure to harmful content, compared to the minimal positive outcomes of creeps getting their kicks. Governmental oversight is necessary to prevent the use of models that allow these types of content to be created, but as of March 2024, only ten U.S. states have banned their use (Davis, 2024). Further governmental oversight is required with steep punishments for any AI model that can create media that includes these forms of harmful content.

The sourcing of training data is another major ethical concern for content-creating AIs. AI models must be trained on an enormous amount of data to be able to produce relevant output. This can cost companies millions of dollars for just acquiring the training data, with the cost being even higher for the increased complexity and size of training data (Reilly, 2024). This is extremely relevant for AI for media generation as it is not just comprehensible language that these models generate, but also images and videos. This high upfront cost has inspired designers of AI models to use the abundance of free media found online to train their models without paying or even notifying the original creators (Reilly et al., 2024).

The creators of AI models, such as OpenAI, argue that this practice is ethical as the output of the models is transformative of the original works and is therefore legally protected

under federal fair use laws (Zirpoli, 2023). They also state that most AI models do not "regenerate… any unaltered data from any particular work in their training" which would defend their claim that the models are transformative as they are not built to directly copy other works. Furthermore, most content creation AI models, such as the DALL-E image generator, have built-in features that prevent the creation of works that are similar to the style of another artist. The combination of all these features and practices by the creators of these AI models should ideally prevent the models from outputting something that already has a distinct copyrighted style. However, content creators and management firms have filed multiple lawsuits claiming that by training models on their copyrighted works, the models produce unauthorized derivative works of the copyrighted training material (Kim, 2024).

In notable legal cases, authors, artists, and media companies claimed that generative AI models had illegally trained models on their works and produced output that was similar to their copyrighted style. In all relevant cases, judges have unanimously agreed that using copyrighted works to train models is decidedly fair use as the training images are not commercially released to the public (Kim et al., 2024). Furthermore, every attempt to prove the output of generative AI models was similar enough to copyrighted materials to be considered infringement has failed. These claims have failed primarily due to the specific nature of copyrights: while a specific work can be copyrighted, the general style of an artist cannot be. So despite a model being trained on copyrighted works, as long as it does not output the exact copyrighted work it is not legally considered copyright infringement. There are still many legal cases on this topic under review and with artists generally upset about the landscape of AI using their work to create similar material in their art style there will likely be even more legal cases started.

Although this practice might currently be legally protected, the creation of material based on an artist's work without their input is inherently unethical, especially if these AI models are used without the artist's wishes and provide output within the same creator space as the original artist. This can lead to artists losing money as they are inherently unable to keep up with the speed with which an AI model can quickly generate works in their style. This problem can be analyzed through the framework of deontology where the creators of these AI tools should recognize that using these artist's work without their permission is in line with stealing which is a morally wrong action. Further governmental oversight seems to be necessary for AI content generation to be ethical in how it pertains to artists being able to profit off their unique style and skill. However, with the relative newness of this issue, the Federal Copyright Office proposes that Congress "adopt a wait-and-see approach" before amending any federal copyright laws (Zirpoli, 2023). This would allow Congress the time to gather more information on the topic and allow the courts to gain experience before weighing in on the topic. It's unlikely that any legislation focusing on this particular issue with AI for content-generation purposes will be introduced anytime soon but for content-creating AIs to be ethical they should not be able to create similar works as a specific artist.

Another prominent issue with sourcing data for content-creating AIs is the potential bias behind the training data. A significant amount of media, especially easily found media online, contains bias/discrimination. With data collection being a tedious task and an exorbitant amount of data that needs to be processed, a lot of AI companies offload the classification of the training media to the people who originally posted the content (Reilly, 2024). By using user-defined data, the AI content generation models that are trained on this data can create harmful and stereotypical content (Ananya, 2024). The creation of biased/discriminatory images is not

necessarily illegal but it is unethical according to the framework of deontology. Through deontology, the principle of respecting others is not being adhered to as discriminatory output infringes upon people's right to fair and equal treatment. The creators of Generative AI models, such as OpenAI's DALL-E, have attempted to prevent discrimination in their models by utilizing techniques "to create more diversity from prompts that do not specify race or gender." According to DALL-E, "users were 12x more likely to say that DALL-E images included people of diverse backgrounds after the technique was applied" which is a vast improvement (OpenAI, 2022). However, implementing diversity algorithms can be tricky and can exacerbate other forms of bias. When Google released its image generator Gemini it was originally unable to produce pictures of white men, prompting them to temporarily prevent the model from creating images of people (Ananya, 2024). Although AI model creators do not seem bothered by the fact that their training data contains bias and discrimination, they are taking active steps to reduce the effect of that training data for its output.

With AI content creation being as easy as inputting a text prompt, many companies that historically hired creatives are considering replacing the work done by human artists with AI models. In 2023, CVL Economics, an LA-based equitability consulting company, surveyed 300 executives and managers from six different entertainment industries and found that 75% of those reported: "GenAI tools, software, and/or models had supported the elimination, reduction, or consolidation of jobs in their business division" (Wolters, 2024). This potentially affects nearly 204,000 creative U.S. jobs with likely more affected as these numbers do not reflect those who work for commissions. This study was commissioned by multiple artist unions that are hesitant about the adoption of AI for content creation. Therefore it is a useful study for determining the scope of the jobs that could be impacted, but not for jobs that are guaranteed to be impacted.

Examples of creative jobs lost to content-creating AIs are hard to find in the U.S., with most of the sources being commissioned workers who are receiving fewer commissions and predicting that it is due to AI (Wolters et al., 2024). This is likely due to strong union networks in these industries that are extremely hesitant about the adoption of AI. One of these unions, SAG-AFTRA, was able to reach a deal with record labels that "requires clear and conspicuous consent and minimum compensation requirements" for AI-generated music that includes an artist's likeness (Klar, 2024). This agreement secures artists' jobs against AI tools as they no longer have to compete against music generated in their likeness without their explicit consent. However, in China and Japan, the loss of creative jobs to content-creating AIs is already evident. In the Chinese gaming industry, illustrators are being encouraged to create content with AI tools, with the payment per illustration being drastically reduced (Zhou, 2023). This completely changes the illustrator's job, reducing their creative control to a glorified cleanup duty. In Japan, a Netflix animated short film was created using AI images. Netflix claims that they decided to use AI for image generation instead of people due to animator labor shortages (Cole, 2023). However, this is extremely unlikely as the Japanese animation industry is a massive market but despite the demand for production, "the industry has long been fraught with labor abuses and poor wages" (Cole et al., 2023). These situations demonstrate that without proper protections, it is entirely possible that artists can lose their jobs in favor of generative AI. This is in contrast to the U.S. where it appears that strong labor unions for creative industries are preventing AI from entirely replacing people's jobs but instead allowing artists to work with new tools to find a balance between genuine human-created art and technology that aids its creation.

One of the major unethical uses of AI content generation is the creation of stale media with the sole purpose of prioritizing views and thus profits. A significant portion of easily

identifiable AI media is found on YouTube in the form of kid's content. These channels, such as Yes! Neo, frequently imitate the simple 2D or 3D animation style used by popular kid's YouTube channels (Knibbs, 2024). This allows them to have content that is similar enough to popular, verified channels which trick children and parents into watching their videos by mistake thus increasing the reach of their channel. These AI-generated videos frequently contain bright colors and include songs to make their content even more entrancing. This content is designed to be easily digestible and visually interesting, however, it is rarely meaningful. The AI-generated videos from Yes! Neo frequently have brief counting or color lessons built into their videos but no true lesson or focus other than catchy colors and tunes. Well-respected children's YouTube channels such as PBS Kids teach morals, skills, and life lessons that are more applicable to the development and interests of a child. These include learning to read, being nice to siblings and friends, and even learning about other cultures. YouTube allows AI-generated kids content as long as the content is marked as AI. However, even if a video is marked as AI, YouTube moderators do not need to approve the post (Knibbs et al., 2024). This practice is harmful to creators on the platform as their thoughtful content is being imitated and mass-produced and it is harmful to children as the content they are being recommended is devoid of human creation. This is unethical through the framework of utilitarianism as the negative effects of this AI-generated content outweigh the minimal positives of providing a distraction for kids to look at.

The purpose of this paper is not to prove that using AI for content generation is inherently unethical but to consider the ethical implications of the content. There are many ways that AI can be used as a tool to create ethical and fascinating new content. Refik Anadol is a digital artist that uses massive collections of targeted datasets to train AI models and monitors their output in real time (Anadol, 2020). He then projects the model’s outputs on the walls, floor, and ceiling of a

space that people can walk around in. This creates fascinating imagery, described as “Machine Dreams”, that people can feel immersed in. He has completed many of these projects such as a timelapse through New York, the history of Disney, and Boston Airport wind data all through the lens of visualizing what the machine sees when it creates connections. It is an unorthodox way to create art from AI models but it is this type of art that epitomizes the necessity of human interaction with these models to create something unique and meaningful that wouldn’t be possible without this technology.

Overall, the current propulsion of AI content generation for media and entertainment purposes is not sustainable and will cause significantly more harm than good. By preventing deliberate misuse, training on ethical and ethically obtained data, and ensuring human creativity is at the forefront, AI models for content creation can be ethical. However, significant steps must be made for all these considerations to be true. These steps include significant governmental regulations on the output of AI models, preventing the creation of deliberately misleading and harmful content. It also requires a redefinition of copyright as it relates to AI, preventing AI models from explicitly copying the styles of living artists. Finally, AI-generated media should be legally required to be labeled as such. This would allow people to be aware of and choose the type of content they wish to view. With these legal changes in mind, AI for content generation can be an entirely ethical medium for widespread media. Unfortunately, these changes are likely to be slow, if at all, so until then, only very specific forms of AI-generated media can be truly ethical.

References

Anadol, R. (2020, January 16). How this guy uses A.I. to create art [YouTube]. Retrieved from https://www.youtube.com/watch?v=I-EIVlHvHRM

Ananya, R. (2024, March 19). Ai image generators often give racist and sexist results: Can they be fixed?. *Nature News*. Retrieved from https://www.nature.com/articles/d41586-024-00674-9

Cole, S. (2023, February 1). Netflix made an anime using AI due to a "labor shortage," and fans are pissed. VICE. Retrieved from https://www.vice.com/en/article/bvmqkv/netflix-anime-dog-and-the-boy-ai-generated-art

Davis, E. (2024, January 30). These States Have Banned the Type of Deepfakes That Targeted Taylor Swift. *U.S News*. Retrieved from https://www.usnews.com/news/best-states/articles/2024-01-30/these-states-have-banned-the-type-of-deepfake-porn-that-targeted-taylor-swift

Giattino, C., Mathieu, E., Samborska, V., & Roser, M. (2024, January 31). Artificial Intelligence. *Our World in Data*. Retrieved from https://ourworldindata.org/artificial-intelligence?insight=ai-systems-perform-better-than-humans-in-language-and-image-recognition-in-some-tests#key-insights

Hao, K. (2022, February 4). A horrifying new AI app swaps women into porn videos with a click. *MIT Technology Review.* Retrieved from https://www.technologyreview.com/2021/09/13/1035449/ai-deepfake-app-face-swaps-women-into-porn/

Heikkilä, M. (2023, February 27). Ai Image Generator Midjourney blocks porn by banning words about the human reproductive system. *MIT Technology Review*. Retrieved from https://www.technologyreview.com/2023/02/24/1069093/ai-image-generator-midjourney-blocks-porn-by-banning-words-about-the-human-reproductive-system/

Kim, R. (2024, February 21). Ai and copyright in 2023: In the courts. *Copyright Alliance*. Retrieved from https://copyrightalliance.org/ai-copyright-courts/

Klar, R. (2024). SAG-AFTRA, record labels reach deal over AI protections for artists. *The Hill*. Retrieved from https://thehill.com/policy/technology/4595485-sag-aftra-record-labels-reach-deal-over-ai-protections-for-artists

Knibbs, K. (2024, March 12). Your kid may already be watching ai-generated videos on YouTube. *Wired*. Retrieved from https://www.wired.com/story/your-kid-may-be-watching-ai-generated-videos-on-youtube/

Marcus, G. (2024, February 20). AI platforms like chatgpt are easy to use but also potentially dangerous. *Scientific American*. Retrieved from https://www.scientificamerican.com/article/ai-platforms-like-chatgpt-are-easy-to-use-but-also-potentially-dangerous/

OpenAI (2022, July 18). Reducing bias and improving safety in dall·e 2. Retrieved from https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2

R. A. Partadiredja, C. E. Serrano and D. Ljubenkov, "AI or Human: The Socio-ethical Implications of AI-Generated Media Content," 2020 13th CMI Conference on Cybersecurity and Privacy (CMI) - Digital Transformation - Potentials and Challenges(51275), Copenhagen, Denmark, 2020, pp. 1-6, doi: 10.1109/CMI51275.2020.9322673.

Reilly, J. (2024, January 25). Cost of AI in 2024: Estimating Development & Deployment expenses. *Akkio*. Retrieved from https://www.akkio.com/post/cost-of-ai

Santa Clara University. (2014, August 1). Calculating consequences: the utilitarian approach to ethics. *Markkula Center for Applied Ethics*. Retrieved from https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/calculating-consequences-the-utilitarian-approach/

Stahl B. C. (2021). Concepts of Ethics and Their Application to AI. Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies, 19–33. https://doi.org/10.1007/978-3-030-69978-9_3

Verma, P. (2023, December 17). The rise of AI fake news is creating a 'misinformation superspreader.' *The Washington Pos*t. Retrieved from https://www.washingtonpost.com/technology/2023/12/17/ai-fake-news-misinformation/

Wiggers, K., & Silberling, A. (2022, November 17). Meet Unstable Diffusion, the group trying to monetize ai porn generators. *TechCrunch*. Retrieved from https://techcrunch.com/2022/11/17/meet-unstable-diffusion-the-group-trying-to-monetize-ai-porn-generators/

Wolters, J. (2024, February 1). The Animation Guild: Future unscripted: The impact of generative AI on Entertainment Industry Jobs. *CVL Economics*. Retrieved from https://animationguild.org/wp-content/uploads/2024/01/Future-Unscripted-The-Impact-of-Generative-Artificial-Intelligence-on-Entertainment-Industry-Jobs-pages-1.pdf

Zirpoli, C. T. (2023, September 29) Generative Artificial Intelligence and Copyright Law. Congressional Research Service. Retrieved from https://crsreports.congress.gov/product/pdf/LSB/LSB10922

Zhou, V. (2023, April 11). Ai is already taking video game illustrators' jobs in China. *Rest of World*. Retrieved from https://restofworld.org/2023/ai-china-video-game-layoffs-illustrators/