AI Art and Data Ethics

Ryan Lunas and Pieper Smith

CS395-001

Dr. Harris

February 25, 2024

Abstract

In this paper, we discuss the rapid growth of generative AI models and their impact on the art industry. We outline the importance of moral, ethical, and legal considerations to the development of these models. We begin by defining and tracing the roots of contemporary generative AI models. We then profile some of the key players in the market and discuss recent developments regarding their products. We cover data ethics, and relate it to how these technologies have caused ethical and legal problems using concrete examples such as the Anderson v. Stability AI lawsuit. Tools such as Nightshade, which were developed to combat unethical data gathering, and their impact on the models are discussed. We finish with a discussion that emphasizes the need for further research in the field of AI and data ethics.

Introduction

The use of Generative AI models have seen explosive growth recently due to advancements in the technologies that they are built on. Players such as OpenAI, Google, Apple, Midjourney, and Stability AI are creating new AI models and competing with each other in what has been called an AI arms race. This rapid advancement has created a giant and growing worldwide industry, “The market size in the Artificial Intelligence market is projected to reach US$305.90bn in 2024. …[it] is expected to show an annual growth rate (CAGR 2024-2030) of 15.83%, resulting in a market volume of US$738.80bn by 2030.” (Statista, 2024)

Since this market is growing so fast and is increasingly becoming a part of people’s daily lives, it is important to understand the ethical considerations surrounding the development and use of these AI products. We will provide an overview of the recent history of AI image generation technology, the ethical and legal problems related to it, and the remedies suggested by researchers. We will approach these issues with a focus on the art industry as it has been sharply affected by the recent rise in AI image generation. In this industry, there is a significant amount of vocal concern regarding the ethics of current web scraping technologies and other data sourcing methods, as well as the economic impact on artists (Chatterjee, 2022).

The sudden arrival of AI technology in this space has caused some artists and researchers to use and develop technologies intended to protect artwork and art styles from being co-opted into the models. These technologies include poisoning tools such as Nightshade, which add data to images to confuse the AI algorithms, and services such as Artshield and Mist, which scan the datasets used by AI in an attempt to find specific pieces of art, so that artists can request that they be removed.

AI Image Generation

The easy definition is that AI image generators are computer programs that take a prompt and generate an image that attempts to represent that prompt. More formally, AI image generation is the use of computerized generative models trained by machine learning techniques to synthesize images that match a certain prompt. Generative models are statistical models of the joint probability distribution between observable variables (in the feature space) and target variables (in the output space). Machine learning is a field of study concerned with enabling machines (computers) to “learn” from inputs and to imitate intelligence. The prompt given to the model can be any data that can be mapped to the output space such as text (text-to-image) or an image (image-to-image) (Jiang et al., 2023).

The origins of AI image generation were some very early experiments before the 2000s which used computers to generate algorithmic and generative art. Early advances in the field of machine learning, such as the development of the “Perceptron” in 1958, set the stage for the more advanced neural networks that would later underpin contemporary AI image generation. The Perceptron was an early probabilistic model that attempted to mimic and model the storage and organization of information in the brain (Rosenblatt, 1958). Another one of the most notable examples that came around in the early 1970s was Harold Cohen’s AARON, which was a computer program written in C which “was intended to identify the functional primitives and differentiations used in the building of mental images and, consequently, in the making of drawings and paintings.” (Garcia, 2016) The AARON system is different from the contemporary applications of AI because the forms that it was programmed with were explicitly written in code, as opposed to the inference of forms from large-scale data which define the techniques of the deep learning era.

The next important development in this field was texture synthesis. In the early through late 2000s, computer vision research was being done to take images of textures as inputs to algorithms designed to extend or generate new images based on the source texture (Efros & Freeman, 2001). After texture synthesis came more recent techniques developed in the deep learning era. The deep learning era started in 2012, and is still ongoing today. One of the techniques developed during it was convolutional neural networks. These repeatedly break up images into smaller and smaller pieces and apply various techniques including convolution to create many copies of these pieces that bring out the features of each piece. The processed pieces are then passed to the input layer of a neural network which, through a series of weighted nodes, will associate the inputs with a single output node which represents the most probable item in the output space (Krizhevsky et al., 2012).

Variational autoencoders were also developed during the deep learning era. This is a machine learning architecture that uses a statistical method called a SGVB (Stochastic Gradient Variational Bayes) estimation which reparameterizes neural networks to improve their inference capabilities (Kingma & Welling, 2013). The neural networks are mirrored around a latent space. One half is the encoder which processes inputs and passes them to the latent space where information about the similarity of items is embedded, the latent space passes the inputs to the decoder which produces a similar, but unique item (Rocca & Rocca, 2019).

Another important kind of generative model is generative adversarial networks. One of the first papers about these has a good explanation:

> The generative model can be thought of as analogous to a team of counterfeiters, trying to produce fake currency and use it without detection, while the discriminative model is analogous to the police, trying to detect the counterfeit currency. Competition in this game drives both teams to improve their methods until the counterfeits are indistinguishable from the genuine articles. (Goodfellow et al., 2014)

Researchers began to experiment with the use of NLP (Natural Language Processing) algorithms to enhance or perform image synthesis. This led to the creation of NLP transformers, which use the transformer architecture (Vaswani et al., 2017) in conjunction with NLP algorithms to create high performance models that can coherently deal with long streams of data (that is, with a large context window/long-range interactions) (Esser et al., 2020). These advances led to research that was used to create DALL-E, an image generator built using transformers and the GPT-3 LLM (Large Language Model) (Ramesh et al., 2021).

Diffusion models are probabilistic models that use a method of generating images inspired by thermodynamic diffusion in physics. Images are generated through repeated denoising of an image that is originally 100% noise. The models are specially trained using images that have partial noise applied to them. The model "learns" to undo this noise to make coherent images. These models were used alongside NLP transformers and a set of contrastive pretrained autoencoders known as CLIP to create Stable Diffusion (Rombach et al., 2021). CLIP provides a neural network architecture designed to map large datasets of image-text pairs to allow it to associate text prompts with images. This essentially makes it so that the text of "a photo of an apple" and an actual photo of an apple are seen as similar to the network. This same structure is used by DALL-E 2 to perform its image generation (Radford et al., 2021). Other than CLIP, the dataset used by DALL-E 2 is not publicly available, whereas the dataset for Stable Diffusion is.

Two important image-text datasets used to train image generation models were JFT-300M and LAION-5B. These two giant image-text datasets gathered their images through web scraping and contain millions to billions of image-text pairs (Schuhmann et al., 2022). It should be noted that the scale of these datasets makes it impossible to manually verify that the images contained within are wanted in the set. To address this they applied various filtering methods and algorithms to judge the "unsafeness" of images and then reject the ones that are deemed "unsafe". This has not been entirely successful however, and the LAION-5B dataset is currently unavailable (since December 2023) after a Stanford Cyber Policy Center paper had found known illegal abuse images within the dataset (Thiel, 2023).

Major Image Generation Companies

Looking at the most recent part of the deep learning era, 2023, the current state of the art models are contrastively pretrained diffusion models trained on image-text pair datasets (Jiang et al., 2023). Some of the major companies that created these models include OpenAI, Stability AI, Midjourney, and Google.

OpenAI is the creator of ChatGPT and DALL-E. They are responsible for the GPT series of LLMs as well as CLIP. Their most current image generator model is DALL-E 3 which improves upon DALL-E 2 by algorithmically recaptioning their dataset to include significantly more text detail about the contents of the images. This makes its generation of images more closely follow the prompt, making it less likely to leave out requested details. It is a contrastively pretrained diffusion model.

Stability AI is the creator of Stable Diffusion and Dream Studio. These also use a contrastively pretrained diffusion model. Stable Diffusion is open-source and is trained on publicly available datasets. Midjourney is the creator of the Midjourney service. Users connect to a discord server and ask a bot to generate their images from prompts. It is also a contrastively pretrained diffusion model.

Lastly, there is Google, the creator of Bard and Gemini AI. Their most recent model is Imagen 2, which is also a contrastively pretrained diffusion model. Imagen 2 was only released for public use very recently, and was accessed by prompting Bard AI (now Gemini) to generate an image. It was also available in a Google Labs experiment called ImageFX. However, Google removed the people image generation capabilities of the models because they were generating offensively historically inaccurate images. As people were using the model, they noticed that the model would generate diverse groups of people in situations where it was inappropriate, such as generating images of Black Nazi soldiers (Grant, 2024). Additionally, some users allege that the generator had a pattern of incorrectly seeing some types of prompts as "sensitive" and refusing to generate them (O'Brien, 2024). Google acknowledged that issues exist with the model and is currently working on tweaking their model before rereleasing its people generation capabilities (Raghavan, 2024).

Data Ethics: AI Copyright Lawsuits

As previously mentioned, these image generating models are built off of billions of images scraped from the web. In addition to inappropriate or illegal content, such as some of the images in the LAION-5B dataset, some of the images used as training data were copyrighted. One important lawsuit related to the data ethics of AI image generators is *Andersen v. Stability AI Ltd*., a lawsuit from October of 2023 against Stability AI, Midjourney, and DeviantArt accusing them of copyright infringement. The plaintiffs of the lawsuit claimed that Stable Diffusion was trained on the plaintiffs' artworks, which allows users to create images in the style of specific artists. They argue that this harms the livelihood of those artists, as the images created through these generators "compete in the marketplace with the original images" (*Andersen*, 2023a). Instead of having to commission artists for art, or license their creations, the potential customer can instead generate a work of art in their style for free instead.

This is also a problem because images created via image generators are not able to be copyrighted in the United States, due to the lack of human authorship. This was first decided in 2022 in a rejection of copyright for an artwork created entirely by AI with no human assistance.

The decision was affirmed by a court decision in 2023 (Brittain, 2024). Since AI generated images can not be copyrighted, they are public domain, and can be used for commercial purposes by anyone. This means that someone could create works in the style of an artist, and take away money from them that way, they could also sell these AI created artworks and make a profit from it. As the case says, "[t]he harm to artists is not hypothetical—works generated by AI Image Products "in the style" of a particular artist are already sold on the internet, siphoning commissions from the artists themselves" (*Andersen*, 2023a).

One of the lawsuit's primary claims was that "Defendants, by and through the use of their AI Image Products, benefit commercially and profit richly from the use of copyrighted images" (*Andersen*, 2023a). This case was dismissed, but some of the key claims were allowed to move forward. The claims not allowed to move forward were "copyright infringement, right of publicity, unfair competition and breach of contract claims against DeviantArt and Midjourney" (Cho, 2024). The allegations were said to be "defective in numerous respects" (*Andersen,* 2023b), but the case was not completely dismissed, and the plaintiffs can still amend their claims and try again.

The most recent development in this case, on February 8th of this year, denied DeviantArt's motion to strike the case for good. The case is currently about the use of the plaintiff's names or styles when the companies commercially promote their products (Cho, 2024). The motion to strike the case was based on California's anti-SLAPP laws, which stop lawsuits that try to silence criticism and infringe upon the rights of free speech. DeviantArt's motion was denied because having a ruling on the case would be in the public interest of the artists of California. Additionally, the plaintiffs have claimed, but have not been able to prove that their names/art styles were used to advertise Midjourney products. If their claims are true, they will not fall under the anti-SLAPP laws (Cho, 2024). This means that in order to determine if anti-SLAPP laws apply, a decision on the case must be made, so it can not be dismissed.

Nightshade

Another recent development in the field of AI image generation was the release of Nightshade. Nightshade is a prompt-specific image poisoner. This means that it transforms images into something unsuitable for training image generation models. Nightshade poisons images by changing what AI image generators see whenever they look at the image (e.g. a handbag instead of a cow) (“What is”, 2024). Nightshade is called ‘prompt-specific’ because it does not attempt to poison the entire image generation model, but rather a single specific prompt. Because the data pool for each individual prompt is much smaller than that of the entire model, Nightshade is capable of corrupting the prompt using a more reasonable amount of poisoned images. It is also capable of corrupting related prompts (Shan et al., 2023). Additionally, the effects of Nightshade will remain even after changes are made to the image. “You can crop [the image], resample it, compress it, smooth out pixels, or add noise, and the effects of the poison will remain. You can take screenshots, or even photos of an image displayed on a monitor, and the shade effects remain” (“What is”, 2024). In particular, pixel smoothing, also known as blurring the image, does not break Nightshade because it changes the majority of the pixels in an image. The visible artifacts are only a small portion of the image changes, so smoothing them out does not break the effects of Nightshade (“Frequently”, 2024). Both of these facts combined, Nightshade prompt-specific approach and resistance to the poison being removed, mean that Nightshade is in theory a highly effective way of poisoning image generation models.

The purpose of Nightshade was to “help deter model trainers who disregard copyrights, opt-out lists, and do-not-scrape/robots.txt directives” (“What is”, 2024). Some companies such as Stable Diffusion have no ‘opt-out’ for artists (“Stable”, n.d.), but some such as DALL·E 3 do, and let artists opt-out from their work being used to train Dall-E’s future image generation models (“DALL·E 3”, n.d.). Publicly released in late January, Nightshade got over 250,000 downloads in 5 days, showing that many artists are concerned about their art work being used as training data without their consent (Franzen, 2024). However, while the original purpose of Nightshade was to protect artists, some worry it may be used maliciously instead, and that some of those downloads were from people hoping to harm image generation models.

There are many who dislike AI image generators. A more subjective negative take on image generation models is that AI art is missing human emotion, and isn’t ‘real art’, or creative. An alternative perspective on this says that image generators can be called creative because they can combine ideas together so quickly in so many ways that they can come up with concepts humans may never have. However, “...they entirely lack the capacity to anchor these outputs in the world. Unlike the kind of creativity valued in humans both in and beyond the arts, ML has little scope for contextualisation” (Ploin et al., 2022). This paper argues that the value of art comes from its connections to the world and to individuals, which AI art lacks without additional human interference. Additionally, once people find out art was AI generated, they see it as less creative. One study found that “people devalue art labeled as AI-made across a variety of dimensions, even when they report it as indistinguishable from human-made art, and even when they believe it was produced collaboratively with a human” (Horton et al., 2023). However, the same study found that AI art could potentially help people appreciate human creativity more. AI image generation can also help non-artists visualize their ideas, and provide inspiration to artists, serving as a sort of concept artist for them. Some artists are even using smaller, self-built image generation models trained off their own work to create art (Ploin et al., 2022).

Implications and Conclusions

AI Image generation models are models trained by machine learning that create images based on prompts, both written and images. They have seen a burst of progress since the beginning of the Deep Learning Era in 2012 which has increasingly enabled them to generate more advanced images which are now difficult to differentiate between traditional art images. This progress has created a significant impact on artists, as the products of these models directly compete with and are sometimes direct imitations of their artwork.

The rapidly improving quality of the images generated by these models have caused them to explode in popularity in the past couple of years. In the race to obtain a share of this growing market, some large companies are having issues with data ethics, and are using copyrighted materials as training data. Many artists are using Nightshade to discourage companies from using their art as training data. There are a lot of pros and cons to AI image generators, and overall, they can’t be called inherently good or bad. This paper focused a lot on the negatives, due to our focus on data ethics, but AI image generation isn’t all bad, and it’s not our intention to say AI art is good or bad, or judge whether or not it’s art at all. Those are things you can hopefully start to decide for yourself, after reading this paper.

References

Andersen v. Stability AI Ltd., 3:23-cv-00201-WHO (N.D. Cal. Jan. 13, 2023a). https://fingfx.thomsonreuters.com/gfx/legaldocs/myvmogjdxvr/IP%20AI%20COPYRIGHT%20complaint.pdf

Andersen v. Stability AI Ltd., 23-cv-00201-WHO (N.D. Cal. Oct. 30, 2023b). https://casetext.com/case/andersen-v-stability-ai-ltd/case-details.

Asperti, A. (2023). Generative models and their latent space. *The Academic*. https://theacademic.com/generative-models-and-their-latent-space/

Brittain, B. (2024, January 23). *Computer scientist makes case for AI-generated copyrights in US appeal*. Reuters. https://www.reuters.com/legal/litigation/computer-scientist-makes-case-ai-generated-copyrights-us-appeal-2024-01-23/

Chatterjee A. (2022) *Art in an age of artificial intelligence.* Front Psychol. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.1024449/full

Cho, W. (2024, February 9). *AI Companies Take Hit as Judge Says Artists Have "Public Interest" In Pursuing Lawsuits*. The Hollywood Reporter. https://www.hollywoodreporter.com/business/business-news/artist-lawsuit-ai-midjourney-art-1235821096/

DALL·E 3. Retrieved February 10, 2024, from https://web.archive.org/web/20240210101847/https://openai.com/dall-e-3

Despois, J. (2017). *Latent space visualization — Deep Learning bits #2*. Hackernoon. https://hackernoon.com/latent-space-visualization-deep-learning-bits-2-bd09a46920df

Efros, A.A., and Freeman, W.T. (2001). *Image quilting for texture synthesis and transfer*. Proceedings of the 28th annual conference on Computer graphics and interactive techniques (SIGGRAPH '01). Association for Computing Machinery, New York, NY, USA, 341–346. https://doi.org/10.1145/383259.383296

Esser, P., Rombach, R., and Ommer, B.. (2020). *Taming Transformers for High-Resolution Image Synthesis.* 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. https://doi.org/10.48550/arXiv.2012.09841

Feller, W. (1957). “An introduction to probability theory and its applications”, Vol. 1, 3rd edition. pp. 217–218.

Franzen, C. (2024, January 23). *AI poisoning tool Nightshade received 250,000 downloads in 5 days: ‘beyond anything we imagined’*. Venture Beat. https://venturebeat.com/ai/ai-poisoning-tool-nightshade-received-250000-downloads-in-5-days-beyond-anything-we-imagined/

Frequently Asked Questions (FAQ). (January 31, 2024). Retrieved February 24th from https://web.archive.org/web/20240213122802/https://nightshade.cs.uchicago.edu/faq.html.

Garcia, C. (2016). *Harold Cohen And Aaron - A 40-Year Collaboration.* Computer History Museum. https://computerhistory.org/blog/harold-cohen-and-aaron-a-40-year-collaboration/

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, D., Bengio, Y. (2014). Generative Adversarial Networks. *arXiv e-prints*. https://doi.org/10.48550/arXiv.1406.2661

Grant, N. (2024) *Google Chatbot's A.I. Images Put People of Color in Nazi-Era Uniforms.* The New York Times. https://www.nytimes.com/2024/02/22/technology/google-gemini-german-uniforms.html

Horton, C. B., Jr, White, M. W., & Iyengar, S. S. (2023). Bias against AI art can enhance perceptions of human creativity. *Scientific reports*, *13*(1), 19001. https://doi.org/10.1038/s41598-023-45202-3

Jiang, H.H., Brown, L., Cheng, J., Khan, M., Gupta, A., Workman, D., Hanna, A., Flowers, J., and Gebru, T. (2023). *AI Art and its Impact on Artists*. Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23). Association for Computing Machinery, New York, NY, USA, 363–374. https://doi.org/10.1145/3600211.3604681

Kingma, D.P., and Welling, M. Submitted 2013. Revised 2022. *Auto-Encoding Variational Bayes*. https://doi.org/10.48550/arXiv.1312.6114

Krizhevsky, A., Sutskever, I.and Hinton, G.E. (2012). *Image Net Classification with Deep Convolutional Neural Networks*. Advances in Neural Information Processing Systems, F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger (Eds.), Vol. 25. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

O'Brien, M. (2024) *Google says its AI image-generator would sometimes 'overcompensate' for diversity.* The Associated Press. https://apnews.com/article/google-gemini-ai-chatbot-imagegenerator-race-c7e14de837aa65dd84f6e7ed6cfc4f4b

Ploin, A., Eynon, R., Hjorth I. & Osborne, M.A. (2022). *AI and the Arts: How Machine Learning is Changing Artistic Work.* Report from the Creative Algorithmic Intelligence Research Project. Oxford Internet Institute, University of Oxford, UK. https://www.oii.ox.ac.uk/news-events/reports/ai-the-arts/

Radford, A., Kim, J.W., Hallacy, C., Ramesh, D., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger G., and Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *International Conference on Machine Learning*. https://doi.org/10.48550/arXiv.2103.00020

Raghavan, P. (2024). *Gemini image generation got it wrong. We'll do better.* Google Blog. https://blog.google/products/gemini/gemini-image-generation-issue/

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021). Zero-Shot Text-to-Image Generation. *arXiv* https://doi.org/10.48550/arXiv.2102.12092

Rocca, J., and Rocca, B. (2019). *Understanding Variational Autoencoders (VAEs)*. Towards Data Science. https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2021). High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.48550/arXiv.2112.10752

Rosenblatt, F. (1958). *The perceptron: A probabilistic model for information storage and organization in the brain.* Psychological Review, 65(6), 386–408. https://doi-org.unco.idm.oclc.org/10.1037/h0042519

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, R., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk R., and Jitsev, J. (2022). LAION-5B: An open large-scale dataset for training next generation image-text models. *arXiv*. https://doi.org/10.48550/arXiv.2210.08402

Shan, S., Ding, W., Passananti, J., Zheng, H., Zhao, B.Y. (2023). Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models. *arXiv*. https://doi.org/10.48550/arXiv.2310.13828

Stable Diffusion Online. Retrieved February 10, 2024, from https://web.archive.org/web/20240210223457/https://stablediffusionweb.com/

Statista. (2024). *Market Insights - Artificial Intelligence - Worldwide.* Statista Market Insights. https://www.statista.com/outlook/tmo/artificial-intelligence/worldwide

Thiel, D. (2023). *Identifying and Eliminating CSAM in Generative ML Training Data and Models*. Stanford Digital Repository. https://doi.org/10.25740/kh752sm9123

Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. (2017). Attention is All you Need. *Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.1706.03762

What is Nightshade? Retrieved February 10, 2024, from https://web.archive.org/web/20240209154004/https://nightshade.cs.uchicago.edu/whatis.html