Received: 2025-01-06
Accepted: 2025-02-02

(HOW) DOES AI MAKE ART?

Nora Miškulin (nm1128@rit.edu)

Abstract

It has been taken for granted that AI "makes art" now, but how it does it and what that art really is, remains a black box for many people not familiar with this rather advanced technology. This essay examines the central features of generative AI models by comparing the underlying architectures, training methods, and generative processes of DALL-E 3, Midjourney v7, and Stable Diffusion. I delve into the differences between these three models in order to provide a comprehensive yet simple and digestible explanation of their technological frameworks and the way in which they produce creative outputs. This includes a detailed comparison of each model's strengths, limitations, and unique capabilities in relation to the general field of generative AI. The potential creative role it assumes and problems that arise with it are another point I consider, by exploring the issue of data ownership and copyright. Finally, I discuss decentralized AI as a possible solution, through a succinct comparison between centralized and decentralized AI models.

Keywords*:* generative AI, AI model architecture, generative art, AI model training process, AI

1. (How) Does AI Make Art?

A significant number of people refer to "A" without fully understanding what it stands for, which leads to widespread misuse of the term and the creation of baseless assumptions. Yes, AI stands for artificial intelligence, but there are different types of AI technology. It is important we are able to distinguish between them by engaging in meaningful discussions and think critically, without falling into misinformation or misleading others. While generative AI, a specialized area of artificial intelligence known for producing original output, has shown impressive capabilities in generating content, it is important to recognize its framework, and limitations, as well as understand that it still operates within the boundaries of human input, on human-made models. We made a powerful tool. And we all know that with great power comes great responsibility.

AI's bad reputation is partially rooted in pop culture, with films like *Terminator* fueling fears of machines taking over. Misinformation spreads easily when people lack a deeper understanding of technology, causing unnecessary unease and adding to the sense of mystery surrounding artificial intelligence today. What many call "AI" is generative AI, but before generative AI (later referred to as gen AI), we were met with traditional AI. This AI was the one that beat Garry Kasparov at chess back in 2017. On move 37, the Grandmaster conceded, marking the first time a computer program had won in official tournament play (Waters, 2023). This kind of AI does not invent new ways of playing chess but makes

decisions from preprogrammed strategies. It refers to systems built to respond to specific inputs in specific ways, with the ability to learn from data and make predictions based on that information, all very precisely and fast. But that's all.

On the other hand, generative AI was built to create new content. Gen AI creates new content from the information it receives; it can generate not just text, but also images, music, and computer code. These models are trained on data and learning patterns to produce outputs similar to their training set (Marr, 2023). The distinction is often overlooked, leading to confusion about the technology's capabilities.

Gen AI has caused an uproar because it appears to outperform humans in many areas, including those we view as subjective and inherently human, such as creativity. It first gained attention as an image-generating tool, producing what is now referred to as “AI art.” Some of the first images we encountered were of avocado chairs. They were created with the model called DALL-E. This model was the first to use a single transformer-based model to generate images directly from text prompts, which enabled it to create an image of an avocado-shaped chair, without having one in its training set (Ramesh et al., 2021).

Now, there are many great AI art generators on the market, among which the best and most popular ones are DALL-E 3, Midjourney v7, and Stable Diffusion. One could say that they all achieve the same goal of generating images from text, but different generative models are built on different AI technologies and specialize in different things. Still, it is important to understand several shared fundamental characteristics that enable them to produce desired results.

The life of every AI generative model starts with the training process (Gcore, 2023). In the case of gen AI which has a goal to create images from text, training begins exactly with that; the model is trained on a large dataset of pairs that assume an image and its text description. This allows it to find and learn patterns, as well as relationships between text prompts and visual elements in the pictures. Training like that provides the AI model with a way of generating relevant images based on the user’s inputs.

Inference is the final stage of the lifecycle of an AI model. After a model has completed its training, it enters the inference phase, where it uses the learned patterns to generate outputs. Inference refers to the process of applying a trained model to real-time user input. During inference, the model does not re-learn or adjust its underlying knowledge; instead, it generates outputs based on what it already knows from training (Martineau, 2023). In other words, inference is the action phase of a generative model, where it transforms learned information into unique outputs.

The generative process itself however is rather convoluted. Each model applies to a collection of complex algorithms that create new images, leveraging the training data in order to produce new, unique outputs. While the specific algorithms that are used for this process differ, the core idea remains the same: generating new content based on learned patterns. All generative AI models also refine their output over time, which means they refine their initial results repeatedly, with the goal of improving both the quality and accuracy of generated images.

Denoising plays a crucial part in this process, especially in diffusion-based models like Stable Diffusion and DALL-E 3 (Kanerika, 2024). This process starts with random noise, and

the model gradually 'denoises' the image, refining it through multiple stages to remove unwanted noise and artifacts, revealing a clearer and more detailed image. In essence, the model starts with an unclear image and progressively works to transform it into something that aligns more closely with the user's input, a key aspect of the image generation process. While denoising is most prominent in diffusion-based models, even models that do not explicitly use a denoising technique still handle noise and artifacts in ways that help maintain the visual coherence and clarity of the result.

Diffusion models are a type of generative model that starts with random noise (essentially a blurry, pixelated image) and gradually transforms it into a coherent image by systematically removing the noise and refining the details (Connor, 2022). This process helps ensure that the final image aligns more closely with the user's input, as it becomes clearer and more detailed with each step. Denoising refers to this technique of removing unwanted "noise" or random pixels during the image generation, improving its quality over time. While denoising is most prominent in diffusion models, all generative AI models must deal with noise and artifacts in some capacity, if not for image generation, then for refinement.

Moreover, all models mentioned focus on the interaction with the user; they rely on user input to guide the generation process. This implies that the quality level and relevance of the output is significantly influenced by the preciseness and value of the user input. Another important characteristic is the diversity of that output. All models aim to produce outputs of high quality and variety, allowing them to be a better match to the user's prompts. This capability ensures a wider field of possible application of the model.

Last, but maybe the most important characteristic that is shared by all generative AI models is the integration of natural language processing (NLP). This is fundamental to all mentioned models because it allows them to interpret and respond to human language input, but it also significantly improves the relevance of the generated outputs. NLP is a branch of AI focused on helping computers recognize, understand, and generate human language. It is not something mysterious or dangerously new. In fact, we encounter NLP daily in digital assistants like Amazon's Alexa or Apple's Siri, which both use it to comprehend and respond to our requests (Stryker & Holdsworth, 2024). Simply put, NLP makes it possible for generative models to take detailed text prompts and transform them into images.

Since most gen AI models follow the same previously described flow, it is easy to assume that there cannot be too big of a difference between them after all. This is, however, false. The trick is that the difference is in the parts most people do not bother to explore and learn - what kind of architecture the model has, model, what training approach it utilizes, and which algorithms it uses for generation - which heavily define how models generate new content (A comprehensive comparison of AI image generation architectures, 2023).

Model architecture refers to the design of a generative AI model, such as transformers, GANs, or diffusion models. It is the model's internal structure; it determines how data is processed and how outputs are generated. These outputs can include text, images, or audio. Each model's architecture affects its ability to understand context, interpret prompts, and generate content with desired characteristics. This capability allows processing inputs where order matters (e.g. NLP). By maintaining this contextual

relationship between elements in a sequence, transformers can generate coherent and relevant outputs.

DALL-E, the model that generated those avocado chairs images, is a transformer-based model; DALL-E 3 (the newest available version of DALL-E) relies on OpenAI's GPT (generative pretrained transformer) for image creation (yes, that is what GPT in ChatGPT stands for). Its ability to understand human input allows it to create images that closely match the intended results, directly from it (The Upwork Team, 2024). This gives it a great amount of control allowing for a highly specific and detailed generative output, as well as image alterations. In this way, it also overcomes mode collapse; a state where the generator produces repetitive or limited images (Chruściński, 2023).

On the other hand, Midjourney v7 focuses on artistic styles and aesthetics. While specific details about its architecture are less public than those for DALL-E 3, it draws inspiration from Generative Adversarial Networks (GANs) and other machine-learning techniques (CyberVenturer, 2023). GANs, rely on two networks in the creation process — a generator that creates images, and a discriminator that evaluates whether those images are real (from the training set) or fake (created by the generator), which often results in issues like mode collapse. Because of that, Midjourney is particularly good at imitating various art forms and creating visually stunning and stylistically specific artwork, but its output often lacks variety (Boesch, 2023).

Stable Diffusion is different from both models mentioned in multiple ways. It has a similar image generation approach, but it uses specific architecture which makes it extremely efficient in producing high-quality images in a short amount of time. Stable Diffusion relies on a model called latent diffusion model. This is a type of image generation approach that creates images based on simplified, compressed data instead of working with all of the image details (Mishra, 2023). It does this by focusing on essential features of an image, which makes it much faster, while still letting it produce high-quality images. This is called a variational autoencoder (VAE) framework. VAEs encode images into a compressed representation called 'latent space,' enabling the model to generate images from simplified data rather than reconstructing them pixel by pixel. By working with compressed data, VAEs make diverse images similar to those in the training set without directly duplicating any image (Boesch, 2024). Another thing that separates Stable Diffusion from Midjourney and DALL-E, is that it is an open-source model, meaning that anybody can access the code of the program and create their own instances of the model. This is by no means an easy thing to do and requires a decent amount of informatical knowledge, but it is possible, it is done, and unfortunately, often misused.

However, that is only one aspect of much broader concerns about generative AI. Regardless of the technology these AI art generators utilize, they wouldn’t be possible without the previously mentioned foundational step: training. While what ‘training’ means for AI may not seem important at a glance, it’s actually tied to many ethical and copyright problems surrounding generative AI. For the generative model to function, it needs to be trained on massive datasets, which typically consist of hundreds of millions of images. The model cycles through this data, repeatedly, imitating patterns it finds. It then reevaluates its outputs and tries again to improve. This repetitive process is the basic principle behind the way gen AI ‘learns’.

Training is unique for every generative AI model; there are many types of training models, each impacting the AI's outputs in specific ways. GANs, for example, involve a competitive training process where one part of the model creates images while another critiques them, pushing the generator to improve (Boesch, 2023). Diffusion models, on the other hand, start with random noise and gradually learn to remove it; functionally transforming random pixels into coherent images, but both methods rely heavily on vast datasets of real images used during training, where the model studies patterns, textures, and compositions (Scale, 2024).

This dependence on large datasets raises more ethical questions. For most of the models, it was proven that training data includes publicly available or scraped images, very often used without explicit permission. This sparked debates about copyright, consent, and the rights of artists whose work may have been used without acknowledgment. As generative AI develops, understanding these training methods is crucial for addressing concerns about originality, ownership, and biases in its dataset. Therefore, it is important to understand that the training approach not only defines the model's technical abilities but also shapes the ethical surroundings in which AI art generators operate.

Revelations about AI models using images from the internet without consent have raised significant concerns about privacy and copyright. In December 2022, David Holz, the founder of Midjourney, acknowledged that the platform trained on hundreds of millions of online images without obtaining permission from creators, citing the lack of available infrastructure for tracking and gaining consent for such a massive number of images (Salkowitz, 2022). Holz's statements roused a huge public backlash, especially among photographers and artists whose work was included in these datasets (Salkowitz, 2022).

Recently, a new solution that deals with most of the problems mentioned has been in development, and that is decentralized AI. Most of the existing generative AI models, including, the previously mentioned DALL-E 3 and Midjournej v7, are centralized. Centralized AI refers to a system where data and model processing are controlled by a single entity or centralized infrastructure (Restack, 2024). These models dominate the landscape because they are built and managed by large corporations with significant computational power and data resources, such as OpenAI, Google, and Meta.

The challenges they face span from limited data access to the lack of transparency and accountability, but they all stem from the limitations inherent to its architecture. These issues restrict the potential of AI in areas that deal with sensitive information, like personalized healthcare, and can lead to inaccurate or biased outcomes. The closed nature of centralized systems also brings about trust issues and hinders innovation (MIT Media Lab, n.d.).

Conversely, decentralized AI is a vision of artificial intelligence that operates on an open and collaborative network, rather than being controlled by a single, centralized organization or entity. It aims to use blockchain and related technologies to address issues like transparency, data privacy, and inclusivity, by distributing control and infrastructure across multiple stakeholders (Protocol, 2024). Decentralized AI models seek to democratize the AI landscape, making it more accessible for individuals and smaller organizations to develop, train, and deploy their own AI models. This comes as a stark contrast to centralized AI which relies on isolated data and operates under limited transparency. By decentralizing

these processes, AI could become a more inclusive tool, addressing issues of bias, data ownership, and control (Li, 2024). Unfortunately, despite their potential advantages, decentralized systems currently face challenges in scaling and competing with the efficiency and resources of centralized AI models.

Moreover, debates continue over whether these creations can truly be considered art. Gen AI is often accused of imitating creativity, as it mimics the human process of learning from examples. Humans subconsciously absorb vast amounts of information from their environment and cultural context, and AI does the same with the data it's fed. However, while AI simply processes patterns and produces results based on those inputs, humans add layers of meaning, emotion, and intention to their creations. The question that arises is whether AI is truly imitating creativity, or is simply reflecting the data we give it. Gen AI's creative outputs, while impressive, are (for now) limited by the human data it processes and lack the emotional and intentional depth that defines human art. To put it simply; we may be just overthinking it.

Generative AI is still in its early stage of development, yet it is already blurring the line between human creativity and machine output. By relying on vast, often publicly sourced datasets, models like DALL-E 3, Midjourney v7, and Stable Diffusion can produce diverse, detailed, and highly specialized outputs. The reliance on human-made data, as well as the abstract way in which it creates outputs, brings numerous questions about originality, ethics, authorship, and the 'creative' role of AI today. These concerns have sparked the need for more transparent, equitable, and inclusive infrastructures for generative AI. As these models grow increasingly more sophisticated, understanding the principles of their development becomes essential—not to prevent a hypothetical AI apocalypse, but to responsibly integrate a powerful tool into our cultural and creative practices.

2. References

Boesch, G. (2023, April 2). *A comprehensive comparison of AI image generation architectures*. Uni Matrix Zero. https://unimatrixz.com/blog/latent-space-comparing-ai-image-generation-architectures/

Boesch, G. (2024, July 9). *Generative AI: A guide to generative models*. Viso.Ai. https://viso.ai/deep-learning/generative-ai/

Chruściński, M. (2023, April 5). *A brief history of AI-powered image generation*. Blogersii. https://sii.pl/blog/en/a-brief-history-of-ai-powered-image-generation/

Connor. (2022, May 12). *Introduction to Diffusion Models for Machine Learning*. News, Tutorials, AI Research. https://www.assemblyai.com/blog/diffusion-models-for-machine-learning-introduction/

CyberVenturer. (2023, July 16). *MidJourney: The AI art tool that's taking the internet by storm*. Medium. https://medium.com/@business_money/midjourney-the-ai-art-tool-thats-taking-the-internet-by-storm-714ca94f3f11

Gcore. (2023, December 13). *What is AI inference and how does it work?* Gcore. https://gcore.com/learning/what-is-ai-inference/

Kanerika. (2024, October 31). *The power of diffusion models in AI: A comprehensive guide*. Kanerika. https://kanerika.com/blogs/diffusion-models/

Li, M. (chong). (2024, November 12). *Watch decentralized AI in 2025: The convergence of AI and crypto*. Forbes. https://www.forbes.com/sites/digital-assets/2024/11/12/watch-decentralized-ai-in-2025-the-convergence-of-ai-and-crypto/

Marr, B. (2023, July 24). *The difference between generative AI and traditional AI: An easy explanation for anyone*. Forbes. https://www.forbes.com/sites/bernardmarr/2023/07/24/the-difference-between-generative-ai-and-traditional-ai-an-easy-explanation-for-anyone/

Martineau, K. (2023, October 5). *What is AI inferencing?* IBM Research; IBM. https://research.ibm.com/blog/AI-inference-explained

Mishra, O. (2023, June 8). *Stable diffusion explained*. Medium. https://medium.com/@onkarmishra/stable-diffusion-explained-1f101284484d

MIT Media Lab. (n.d.). *Decentralized AI*. MIT Media Lab. Retrieved November 17, 2024, from https://www.media.mit.edu/projects/decentralized-ai/overview/

Protocol, L. (2024, November 14). *Building a democratic AI ecosystem: A vision for a decentralized, inclusive future*. Lumerin Blog. https://medium.com/lumerin-blog/building-a-democratic-ai-ecosystem-a-vision-for-a-decentralized-inclusive-future-9609e3d6b710

Ramesh, A., Pavlov, M., Goh, G. & Gray, S. (2021, January 5). *DALL·E: Creating images from text*. Openai.com. https://openai.com/index/dall-e/

Restack. (2024, August 11). *Centralized vs decentralized vs distributed AI*. Restack.Io. https://www.restack.io/p/ai-in-blockchain-knowledge-centralized-vs-decentralized-vs-distributed-cat-ai

Salkowitz, R. (2022, September 16). *Midjourney founder David Holz on the impact of AI on art, imagination and the creative economy*. Forbes. https://www.forbes.com/sites/robsalkowitz/2022/09/16/midjourney-founder-david-holz-on-the-impact-of-ai-on-art-imagination-and-the-creative-economy/?sh=571764c32d2b

Scale, A. I. (2024). *Diffusion models: A practical guide*. Scale.com. https://scale.com/guides/diffusion-models-guide

Stryker, C. & Holdsworth, J. (2024, August 11). What is NLP (natural language processing)? *Ibm.com*. https://www.ibm.com/topics/natural-language-processing

The Upwork Team. (2024, May 13). *Midjourney vs. DALL-E: Differences, Examples, and Which Is Better*. Upwork.com. https://www.upwork.com/resources/midjourney-vs-dall-e

Waters, D. (2023, March 22). *The historic chess showdown between man and AI, decades before ChatGPT*. Washingtonpost.com. https://www.washingtonpost.com/history/2023/05/22/garry-kasparov-chess-deep-blue-ibm/