UNMASKING THE ILLUSION: THE TECH BEHIND DEEPFAKES

MICHELE BRIENZA,[1] DOMENICO DANIELE BLOISI,[2] DANIELE NARDI[1]

[1] Sapienza University of Rome, Department of Computer, Control and Management Engineering "A. Ruberti", Rome, Italy
brienza@diag.uniroma1.it, nardi@diag.uniroma1.it
[2] International University of Rome - UNINT, Department of International Humanities and Social Sciences, Rome, Italy
domenico.bloisi@unint.eu

The chapter provides an overview of the technology behind deepfakes, describing what a deepfake is and how it is created. The chapter is structured around three sections: (i) theoretical foundations of artificial intelligence, machine learning, and deep learning, (ii) generative models and synthetic data, and (iii) the synthetic media toolkit. Firstly, it describes AI evolution, starting from the early stages leading up to the latest models that can generate data. The latest models are then described, highlighting their capabilities and explaining how these models open a wide range of opportunities, as well as the concerns regarding the generation of highly realistic data that can deceive users, as is the case with deepfakes. Finally, knowledge of how the machines learn from the data helps in using these tools. A clear understanding of the process behind the technology leads to unmasking the illusion and understanding how the technology works, enabling informed use.

DOI
https://doi.org/10.18690/um.feri.2.2026.1

ISBN
978-961-299-109-8

Keywords:
Generative AI, GANs, deepfakes, machine learning, AI history

University of Maribor Press

1 Theoretical Foundations: Artificial Intelligence, Machine Learning, Deep Learning

A "deepfake" is a digital content (i.e., an image, a video, or an audio) modified or generated by using artificial intelligence (AI) tools. The term combines two concepts: "deep" refers to deep learning, a branch of AI, while "fake" indicates that the content has been manipulated or altered in some way. So, to understand deepfakes, it is important to first examine the theoretical foundations of AI.

The objective of AI is to create machines capable of performing tasks that typically require human intelligence. The journey begins in the 1950s, when pioneers such as Alan Turing posed the fundamental question of machine intelligence: "Can machines think?" (Epstein et al. 2009). This chapter provides a brief technical overview of how the technology born in 1950 has evolved to today's capability of creating highly realistic synthetic data content.

The path from Turing's question to today's deepfakes not only regards technological advancement, but it is also a fundamental transition in how machines process and generate information. In scientific evolution over the last decades, we have seen a progression from rule-based systems to machine learning and, finally, to deep neural networks. This evolution has led nowadays to models that can generate realistic human faces, synthesize speech in any voice, and produce entirely fictional yet photorealistic scenarios.

The formal birth of the term "artificial intelligence" took place in 1956, when, during a conference, John McCarthy, one of the pioneers of AI, coined the term (McCarthy et al. 2006). However, the conceptual foundations were laid earlier by Alan Turing, whose 1950 paper "Computing Machinery and Intelligence" introduced the famous Turing Test as a benchmark for machine intelligence, posing the question, "Can machines think?"

Developing a machine capable of "thinking" is a challenge that extends beyond mere technical complexity. It requires sophisticated models, advanced computational architectures and, crucially, a profound sense of responsibility. The aim is to create systems that can produce high-quality content while adhering to fundamental ethical

principles; a balance that is becoming increasingly important as these technologies become more powerful and widespread.

To understand how machines learn to replicate human intelligence, Tom Mitchell provides a foundational definition of the learning process: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E" (Mitchell 1997).

Early AI development focused on symbolic AI or knowledge-based systems. These systems operated on the principle that intelligence could be replicated through explicit rules and logical reasoning. Engineers would interview domain experts, codify their knowledge into if-then rules, and create systems that could make decisions within structured domains. Due to their rule-based architecture, small changes in input could lead to system failures; to operate in a human-like fashion, a machine requires learning from experience or adapting to new situations not explicitly programmed. Consider image recognition: the exclusive use of standard geometric features such as size, colour, and pose is not sufficient for this task, which requires a more robust approach. Machine learning has emerged as a solution to these limitations, representing a significant advance in the field of artificial intelligence. Instead of programming explicit rules, ML enables systems to learn patterns from data. The fundamental idea is that human behaviours can be learned through data without being specified by a set of rules.

Machine learning comprises three main paradigms:

- **Supervised Learning**: a machine learning paradigm that involves training algorithms on labelled datasets, where each training example consists of an input paired with its corresponding correct output (label). The system learns to map inputs to desired outputs by analysing these input-output pairs during the training phase. Through this process, the algorithm identifies patterns and relationships within the data that enable it to make accurate predictions. Classification tasks (determining whether an email is spam) and regression problems (predicting house prices) are examples of supervised learning. The mathematical foundation rests on finding functions that minimize prediction errors across training data while generalizing well to unseen examples.

- **Unsupervised Learning**: a machine learning paradigm that addresses the challenge of discovering hidden patterns, structures, and relationships within data without the guidance of explicit labels or target outputs. Unlike supervised learning, these algorithms must identify meaningful patterns just using the input data itself, making this approach particularly valuable for exploratory data analysis and knowledge discovery. These algorithms might cluster customers into market segments, reduce data dimensionality for visualization, or discover anomalies in network traffic. Principal Component Analysis (PCA) and k-means clustering represent classical unsupervised techniques that remain widely used today.
- **Reinforcement Learning**: a machine learning paradigm that draws inspiration from behavioural psychology, where agents learn through trial and error in interactive environments. The agent receives rewards or penalties for actions, gradually learning optimal strategies to reach a specified goal. This approach has produced remarkable successes in game-playing AI, from chess programs to AlphaGo's historic victory over human champions.

The early 2000s marked a pivotal transformation in artificial intelligence, driven by exponential growth in computational resources and unprecedented access to large datasets. This technological convergence enabled the emergence of Deep Learning methodologies that fundamentally changed how machines process information. Complex neural architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) demonstrated remarkable capabilities in pattern recognition, language understanding, and cross-modal translation tasks that had previously been challenging. These networks are made of layers that learn patterns from input data to predict outputs.

A salient moment arrived in 2014, with Ian Goodfellow's introduction of Generative Adversarial Networks (GANs), establishing the foundation for modern generative artificial intelligence (Goodfellow et al. 2014). This breakthrough represented more than incremental progress; it introduced an entirely new paradigm for synthetic data creation.

The GAN framework operates through an adversarial training process involving two competing neural networks. The generator network learns to transform random noise into increasingly synthetic content, whether images, audio, or text. Meanwhile,

the discriminator network distinguishes whether the content in the input is generated or not, acting as a digital detective. This competitive process creates a feedback loop where both networks continuously improve: the generator becomes more skilled at creating convincing fakes, while the discriminator becomes better at detecting them. Eventually, this adversarial process reaches an equilibrium where generated content achieves a high level of realism.

However, the most revolutionary advancement in generative AI came with the 2017 introduction of the Transformer architecture (Vaswani et al. 2017), which transformed both natural language processing and artificial intelligence. Transformers introduced the attention mechanism, enabling models to dynamically focus on relevant input segments during processing. This architecture consists of an encoder that builds rich contextual representations through self-attention, and a decoder that generates outputs autoregressively by considering both encoded input and previously generated tokens. This innovation directly enabled the development of GPT models by OpenAI, which demonstrated unprecedented natural language capabilities through large-scale pre-training on diverse text corpora, fundamentally changing how machines understand and generate human language

The impact of GANs has led to advantages across numerous applications. Beyond generating photorealistic synthetic faces of non-existent individuals, these models have revolutionized creative industries, enhanced image super-resolution techniques, and provided synthetic training data solutions when real datasets are limited or prohibitively expensive. GANs essentially established the conceptual groundwork for the generative AI revolution, inspiring subsequent innovations including diffusion models and transformer-based architectures.

While GANs dominated early generative AI development, diffusion models (Ho et al. 2020). emerged as an alternative, particularly for image synthesis tasks. These models employ a fundamentally different approach: rather than direct generation, they learn to progressively remove noise from random input through iterative refinement steps. This denoising process offers greater training stability and finer control over the generation process. Contemporary models like DALL-E (Ramesh et al. 2021) and Stable Diffusion (Rombach et al. 2022) exemplify how diffusion approaches can produce both artistic and photorealistic imagery with unprecedented precision and creative flexibility.

2 Generative Models and Synthetic Data

The original GAN, called Vanilla GAN (Goodfellow et al. 2014), marked the beginning of adversarial networks. Training GANs is challenging due to technical issues such as mode collapse and unstable gradients that cause not high-quality output generation, but this model laid the groundwork for the future development of GAN architectures. Since the introduction of Vanilla GANs, numerous improvements and variations have been proposed to address limitations and expand the applicability of GANs. Some of the major advancements include:

- **Conditional GANs (cGANs)** (Mirza and Osindero 2014) introduced the ability to condition the generation process on additional information, such as class labels. By conditioning both the generator and discriminator on a desired output class, cGANs allowed for greater control over the generated outputs. This was a critical advancement in applications like image-to-image translation and text-to-image generation.
- **Deep Convolutional GAN (DCGAN)** (Radford et al. 2016) leveraged convolutional neural networks (CNNs) to improve the stability and quality of generated images. DCGANs enabled the generation of more detailed and higher-resolution images by using convolutional layers in both the generator and discriminator. This model became a benchmark for image synthesis tasks and paved the way for many subsequent GAN models.
- **Wasserstein GAN (WGAN)** (Arjovsky et al. 2017) addressed the training instability issue by using the Wasserstein distance to measure the difference between real and generated data distributions. This led to more stable training and better convergence.
- **Progressive GAN (PGAN)** (Karras et al. 2018) improved the generation of high-resolution images by starting with a low-resolution image and progressively increasing the resolution during training. This method helped generate more stable and realistic images.
- **CycleGAN** (Zhu et al. 2020) introduced the concept of cycle consistency, enabling unpaired image-to-image translation. This meant that CycleGANs could learn to transform images from one domain to another without needing paired training data, making it highly versatile for applications like photo enhancement and style transfer.

- **StyleGAN** (Karras et al. 2019) introduced a new way to control the generation process by manipulating latent space features to produce human faces. StyleGAN's ability to disentangle features like pose, facial expression, and hairstyle in the generation process resulted in highly realistic images. StyleGAN2 (Karras et al. 2020) and StyleGAN3 (Karras et al. 2021) were followed by improvements, including better handling of textures and the ability to create more coherent images across different resolutions.
- **BigGAN** (Brock et al. 2019) enhanced the performance of GANs by training on large-scale datasets with higher computational power. BigGANs demonstrated the capability of GANs to generate incredibly detailed and high-quality images, pushing the boundaries of what GANs could achieve.
- **Self-Attention GAN (SAGAN)** (Zhang et al. 2018) introduced self-attention mechanisms that allowed the network to focus on relevant parts of an image while generating it. This enabled the generation of images with greater structural and spatial consistency.
- **DragGAN** (Pan et al. 2023) represents a novel approach to interactive image manipulation. It enables users to control specific points in an image and drag them to target positions, providing precise control over shape, pose, and expression. This interactive manipulation method enhances user control in generating and editing images.

The evolution of GANs from simple image generation to sophisticated manipulation tools marks a critical turning point in synthetic media creation. As these models became more powerful and accessible, they enabled a new phenomenon that would capture attention: deepfakes. Thanks to their ability to generate and modify existing multimedia content with increasingly realistic results, the phenomenon of deepfakes has spread widely. Deepfakes are synthetic content created through sophisticated artificial intelligence models, particularly GANs and diffusion models, which allow for convincing manipulation of videos, images, and audio recordings.

This technological capability, while impressive from an engineering perspective, has introduced unprecedented challenges. The same algorithms that can enhance medical imaging or create innovative art can also fabricate convincing videos of public figures saying things they never said, or place individuals in scenarios they

never experienced. The democratization of these tools has made synthetic media creation accessible beyond academic laboratories, raising fundamental questions about truth, authenticity, and the nature of evidence in our digital age.

3 Beyond GANs: The New Wave of Generative Models

While GANs have dominated the field of image synthesis and related tasks for almost a decade, other approaches have shown promising results in producing highly detailed and controllable outputs. These alternative models have provided new perspectives on how generative AI can be approached.

Variational Autoencoders, introduced by Kingma and Welling in 2013 (Kingma and Welling 2013), represent one of the earliest and most influential alternatives to GANs in the generative modelling landscape. VAEs combine the concepts of autoencoders with variational inference, creating a probabilistic framework for learning latent representations of data. The architecture consists of two main components: an encoder network that maps input data to a probabilistic latent space, and a decoder network that reconstructs the original data from latent representations. Unlike traditional autoencoders that learn deterministic mappings, VAEs learn to encode data into probability distributions in the latent space, typically Gaussian distributions characterized by mean and variance parameters.

The key advantage of VAEs lies in their stable training process, which leads to more stable and predictable training compared to the adversarial training of GANs. The probabilistic nature of VAE latent spaces enables smooth interpolations between different data points, making them particularly useful for tasks requiring controlled generation and data exploration. However, VAEs also have notable limitations, particularly in generating sharp, high-resolution images, where they tend to produce somewhat blurry outputs compared to GANs.

The introduction of the Transformer architecture (Vaswani et al. 2017) revolutionized natural language processing and opened new possibilities for generative modelling across multiple modalities. Originally designed for sequence-to-sequence tasks, Transformers have proven to be remarkably versatile and powerful for generative applications. The self-attention mechanism allows Transformers to model long-range dependencies effectively, making them

particularly suitable for sequential data generation. Unlike classic methods such as RNNs or CNNs, transformers can process entire sequences in parallel and capture complex relationships between distant elements.

Most transformer-based generative models work autoregressively, predicting the next token in a sequence given all previous tokens. This approach has proven highly effective for text, code, and even image generation when images are treated as sequences of tokens. The Generative Pre-Trained Transformer (GPT models) are language models. These models showed that scaling up transformers with massive datasets could lead to emergent capabilities in text generation, reasoning, and even multimodal understanding.

Building upon the foundations laid by VAEs and the architectural innovations of Transformers, Diffusion Models have emerged as perhaps the most significant advancement in generative AI in recent years. These models have achieved unprecedented quality in image generation and are rapidly expanding to other modalities. Diffusion models work by modelling the process of data generation as the reverse of a diffusion process, starting with random noise and gradually denoising it through a series of iterative steps to generate data that resembles the original distribution.

The generation quality of diffusion models has been demonstrated through their exceptional capability in generating highly detailed, realistic images that often surpass the quality of both GAN and VAE-generated content. Unlike GANs, diffusion models do not suffer from adversarial training instabilities, and unlike early transformer approaches, they do not require massive computational resources for basic functionality. These models can cover the data distribution more accurately than GANs, allowing them to generate a wider variety of high-quality content.

4 The Deepfake Pipeline: Tools and Techniques for Synthetic Content Creation

Deepfakes, a specific application of GANs, have become a key technology for generating hyper-realistic fake videos and audio. Deepfakes allow for the alteration of visual and auditory content in a manner that is nearly indistinguishable from real media. While the foundational models and applications mentioned here represent

the pioneering technologies that first brought deepfakes to mainstream attention, the field is currently experiencing unprecedented rapid evolution. Early tools like **FakeApp**, **FaceSwap**, and **Face2Face** established the fundamental principles, but today's landscape features increasingly sophisticated architectures, real-time processing capabilities, and improved accessibility. The focus has shifted from experimental proof-of-concepts to robust, production-ready toolkits that can deliver professional-quality results with minimal technical expertise required from users. Below are the major techniques used in deepfake generation and their current implementation frameworks:

- **Face-swap technology** is the most recognized form of deepfake. It involves replacing a person's face in a target video with the face of another individual by training an AI model, one of the ones seen in the previous section, on two sets of facial images: the source face and the target face. The model learns to encode the distinctive features of the source person's face into a latent representation, then reprojects these encoded features onto the target person's facial structure. The face swap process allows for automatic swapping of facial features while maintaining the context and integrity of the original video's environment, adjusting for different face shapes, angles, lighting conditions, and camera perspectives.
- **Lip-syncing deepfakes** manipulate the movement of the lips in a video to match a specific audio input. This technique takes the target video frames and the desired audio as inputs, analysing the phonetic structure and temporal patterns of the speech to generate corresponding visual mouth shapes and facial muscle movements. The model encodes the audio features and learns the correlation between speech sounds and their visual representations, generating a video where the target person's mouth movements are synchronized with arbitrary speech audio. Advanced models employ GenAI to generate real-time lip movements, eye-blinking, and facial expressions that naturally accompany speech, resulting in highly realistic video content.
- **Face reenactment** is a deepfake technique where a source actor's facial expressions, gestures, and head movements are transferred to a target video. The system takes video input from both a source performer and a target person, encoding the source's facial dynamics, including expression parameters, head pose, and micro-movements, into a control representation. This encoded motion

data is then reprojected onto the target's facial structure and appearance, enabling the modification of a person's facial expressions in real-time or recorded videos. This makes it appear as if they are displaying emotions or gestures, they never performed, while preserving the target person's identity and the original video context.

- **Voice cloning** generates synthetic speech that mimics a target person's voice characteristics and speaking patterns. This technique utilizes pre-trained models that can clone voices with minimal input requirements, requiring only a few seconds of audio samples from the target speaker. The system takes the reference audio sample and the desired text as inputs, encodes the vocal characteristics from the audio sample, and generates new speech content that maintains the original speaker's voice while saying the provided text. These models can reproduce acoustic properties, speaking mannerisms, and natural speech variations, enabling the generation of convincing audio content where the target appears to be saying words or phrases they never actually spoke.

The evolution of deepfake technology from academic research to practical applications has followed a predictable pattern of democratization. Initially dominated by complex open-source frameworks requiring substantial technical expertise, the field has progressively become more accessible through lighter models and, most recently, commercial platforms that abstract away all technical complexity.

The early deepfake ecosystem was built around open-source projects. DeepFaceLab emerged as the dominant force, responsible for creating the majority of professional deepfake content. This comprehensive framework provided end-to-end functionality for face extraction, training, and merging, but demanded significant technical knowledge and powerful hardware. Users needed to understand neural network architectures, manage training parameters, and navigate complex file structures. FaceSwap offered a similarly powerful alternative with better cross-platform support and multi-GPU capabilities, built on TensorFlow and Python.

This technical and financial barrier initially limited the creation of deepfakes mainly to academic researchers and industry professionals who used the technology for academic studies or private work. However, demonstrations of the technology's capabilities quickly attracted the interest of a wider community, including people with malicious intentions. Online forums and GitHub communities began sharing

code, tutorials, and pre-trained models. The availability of detailed guides and community support made the technology increasingly accessible, although it still required considerable technical expertise and a significant hardware investment. Cloud platforms played a crucial role in this democratization. Google Colab, launched in 2017 as a tool for research and training in the field of machine learning, offered free access to powerful GPUs. This allowed anyone to experiment with deep learning models without having to purchase expensive hardware by following online guides. Although many have used these resources for legitimate projects, unfortunately, some uses have involved misuse for the creation of non-consensual deepfakes. To circumvent the problem in 2022, Google implemented specific restrictions in its terms of service, explicitly prohibiting the use of Colab for the generation of deepfakes.

The landscape began shifting with the development of lighter, more efficient models that could run on consumer hardware. These optimized architectures reduced training times and memory requirements, making deepfake creation feasible on standard gaming PCs. Real-time applications like DeepFaceLive demonstrated that face swapping could be performed live during video calls or streaming, achieving real-time performances by using a common laptop.

Today's synthetic media landscape has been transformed by commercial platforms that have eliminated virtually all technical barriers. It is possible to generate multimedia content instantly through services offered to users on websites. Services such as FakeYou and DeepSwap offer professional-quality results with a few simple clicks. These platforms handle all the computational complexity in the cloud, allowing users to create convincing multimedia content simply by uploading images and videos. Subscription models typically start at tens of dollars per month, making the technology accessible to anyone willing to pay the cost of the service to generate content that can be used for entertainment purposes.

Even large companies have released their own models for using video generation models for benevolent purposes. HeyGen, Veo3, and Runway are some of the models used to generate videos useful for advertising campaigns or simply for fun.

The above-mentioned models were then developed by design to be used ethically to avoid the generation of malicious content. The landscape of video generation has been transformed by models such as Runway and Google's Veo 3, which represent significant advances in both quality and control. Runway Gen-4 introduces consistent character generation across scenes, allowing filmmakers to maintain visual continuity while generating content from multiple perspectives. The model excels at realistic physical simulation and can use visual references combined with textual instructions without requiring fine-tuning. Google's Veo 3 goes a step further by generating videos with native audio integration, including synchronized dialogue, ambient sounds, and music, creating 8-second clips that achieve cinematic-level realism while maintaining built-in security protocols.

The development of these platforms has opened up a world of opportunities for advertising campaigns, corporate training, multilingual customer support, and educational content thanks to software such as HeyGen, which focuses on creating AI avatars for business applications such as The platform can create highly realistic digital twins from a single photo, with advanced understanding of the script that regulates facial expressions, body language, and voice inflections to match the meaning of the content, while maintaining built-in protections and human moderation to prevent misuse.

The synthetic media toolkit has been further expanded by advanced image manipulation capabilities. Google's Nano Banana, integrated into Gemini, represents a breakthrough in natural language-based image editing that transforms how synthetic content is created. Unlike traditional photo editing software requiring technical skills, Nano Banana allows users to modify images through simple conversational prompts. Users can seamlessly blend multiple photos, change backgrounds, alter clothing and appearance, or place subjects in entirely new environments while maintaining photorealistic consistency. The model excels at character preservation, ensuring that people and animals retain their distinctive features across edits, making it particularly powerful for creating convincing synthetic scenarios.

This capability transforms content creation workflows by eliminating the traditional barrier between imagination and execution. Content creators can now generate complex composite images by describing desired changes rather than mastering

complex editing techniques. The model's ability to maintain spatial coherence and lighting consistency across edits makes it possible to create highly believable synthetic content that would previously require professional photography and post-production skills.

These advanced platforms incorporate sophisticated content moderation systems designed to prevent harmful applications. The model includes embedded content filters that restrict inappropriate content generation, visual watermarking with SynthID for authenticity verification, and metadata identification to maintain provenance tracking. This "safe-by-design" philosophy implements restrictions during the development process rather than relying solely on post-generation filtering.

5 Concluding Remarks

This chapter examined the technological foundations of artificial intelligence that have taken this discipline from rule-based systems designed to mimic human intelligence to modern generative models capable of generating synthetic media, representing a fundamental change in the way multimedia content is generated.

The generative artificial intelligence ecosystem has spread very quickly, starting with GANs, which, with their variants, have led to the development of models such as StyleGAN, capable of generating non-existent human faces with a very high level of quality, and transformers for textual and multimodal applications. An example of the use of these models is GPT and all the large language models currently in use, which allow the generation of very high-quality text and are capable of generating text documents with human-like "reasoning" capabilities. Then there are diffusion models, which achieve image generation and editing with unprecedented image quality through iterative denoising.

It is important to note that the development of these models requires huge data sets, high-performance GPU clusters, and weeks of training. The technology has evolved from resource-intensive research projects to cloud-based services that hide all the technical complexity, transforming the creation of synthetic content from an exclusive research domain to a tool. The technology has evolved from resource-intensive research projects to cloud-based services that hide all the technical

complexity, transforming the creation of synthetic content from an exclusive research domain to a tool accessible to everyone. This is a positive aspect in terms of the democratization of technology, but it raises ethical risks to consider regarding the spread of inauthentic content that can mislead users or defame individuals.

For this reason, the models developed by companies have built-in security measures that include content filters that prevent the generation of harmful content, training data curation, watermarking technologies such as SynthID, and API-level restrictions. Although imperfect, these represent an improvement over the early unrestricted models, which, in the early stages of generative artificial intelligence, allowed non-consensual or defamatory content to be spread online.

The fundamental challenge is that algorithms that enable legitimate applications in entertainment and education can equally serve harmful purposes, such as disinformation and fraud.

End notes

Michele Brienza is the main author of this chapter. He wrote all the sections that briefly describe the history of AI, from its birth to new generative AI models, and the tools and processes that enable the creation of synthetic media. Domenico Daniele Bloisi and Daniele Nardi collaborated in the organization of the content and final review of the chapter.

References

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein Generative Adversarial Networks. *Proceedings of the 34th International Conference on Machine Learning*,214–223. https://proceedings.mlr.press/v70/arjovsky17a.html

Brock, A., Donahue, J., & Simonyan, K. (2018, September 27). Large Scale GAN Training for High Fidelity Natural Image Synthesis. *International Conference on Learning Representations*. https://openreview.net/forum?id=B1xsqj09Fm

Epstein, R., Roberts, G., & Beber, G. (Eds.). (2009). Parsing the Turing Test: Philosophical and Methodological Issues, *Quest for the Thinking Computer*. Springer Netherlands. https://doi.org/10.1007/978-1-4020-6710-5

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. Advances, *Neural Information Processing Systems*, 27. https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. Advances, *Neural Information Processing Systems*, 33, 6840–6851. https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive Growing of GANs for Improved Quality, Stability, and Variation, *arXiv*. https://doi.org/10.48550/arXiv.1710.10196

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., & Aila, T. (2021). Alias-Free Generative Adversarial Networks. Advances, *Neural Information Processing Systems*, 34, 852–863. https://proceedings.neurips.cc/paper/2021/hash/076ccd93ad68be51f23707988e934906-Abstract.html

Karras, T., Laine, S., & Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 4396–4405. https://doi.org/10.1109/CVPR.2019.00453

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and Improving the Image Quality of StyleGAN, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 8110–8119. https://openaccess.thecvf.com/content_CVPR_2020/papers/Karras_Analyzing_and_Improving_the_Image_Quality_of_StyleGAN_CVPR_2020_paper.pdf

Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. https://openreview.net/forum?id=33X9fd2-9FyZd

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazin*e, 27(4), 12–12. https://doi.org/10.1609/aimag.v27i4.1904

Mirza, M., & Osindero, S. (2014). Conditional Generative Adversarial Nets, *arXiv*. https://doi.org/10.48550/arXiv.1411.1784

Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.

Pan, X., Tewari, A., Leimkühler, T., Liu, L., Meka, A., & Theobalt, C. (2023). Drag Your GAN: Interactive Point-based Manipulation on the Generative Image Manifold. A*CM SIGGRAPH 2023 Conference Proceedings*, 1–11.https://doi.org/10.1145/3588432.3591500

Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,arXiv. https://doi.org/10.48550/arXiv.1511.06434

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

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, 30. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2019). Self-Attention Generative Adversarial Networks. *Proceedings of the 36th International Conference on Machine Learning*, 7354–7363. https://proceedings.mlr.press/v97/zhang19d.html

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2242–2251. https://doi.org/10.1109/ICCV.2017.244