An exploration of epistemic functions of AI-generated images and digital authentication.

Supervised by

Dr François Sicard

Key Words:

Epistemology, Photography, Transparency, Stable Diffusion

Candidate Number: JGLR6

Word Count: 9770

Undergraduate final year dissertation (BASC0024) submitted in fulfillment of requirements for the degree of **Bachelor of Arts and Science**

April 2024

Acknowledgement

This project was developed thanks to my supervisor, Dr Francois Sicard, who supported and guided me in challenging my logical thinking.

I acknowledge the use of ChatGPT 3.5 (Open AI, https://chat.openai.com) to proofread my final draft.

Abstract

This dissertation examines the epistemic functions of images in the age of artificial intelligence (AI), focusing on the distinction between analogue photographs, digital images, and AI-generated images. It explores the challenges posed by AI-generated images to our understanding of reality and authenticity, emphasising the need for an interdisciplinary approach that combines philosophy, art criticism, media studies, and computer science. The dissertation explores how these images challenge the traditional epistemic functions of photographs and digital images. Employing a mixed-method approach, including extensive literature review and critical discourse analysis, this research scrutinises the technologies and algorithms underpinning AI image generation, focusing on their transparency and interpretability.

The main findings are firstly, that AI-generated images significantly challenge the traditional epistemic functions of photographs and digital images due to their lack of direct connection to reality and the opaque nature of their image-generation processes. Secondly, despite technological advancements, there is a critical need for educational frameworks that foster the public's critical eye and interpretation skills of AI-generated images.

Preface

Images are everywhere, from social media to advertisements to beautiful photography or artistic creation. This dissertation focuses on looking at images from the incredibly personal perspective of artists and cultural critics and applying those theories to computer vision and machine learning. Majoring in Cultures and minoring in Science and Engineering, my motivation for pursuing this topic originates from my experience and the modules taken during my degree. While doing my Foundation in Art and Design at Central Saint Martins, I realised the important and subliminal role images play in our culture, identity, representation, and, therefore, misrepresentation of the world. Images can become a social movement's logo, empowering people and societies alike to move towards a world they want to see. This experience was extended in my second year of university when I was an intern for the Barbican Centre and was tasked with re-curating their AI: More Than Human exhibition aimed at teaching the public about the recent advancements in AI. The exhibition has been a success, and researching the art world through the lens of AI allowed me to discover works that I have also used as reference in my own creative practice. Taking *Design and Creative Practice 1* and *3* (BARC0094, BARC0108) as well as *Looking, Making and Communicating* (BARC0093) have allowed me to develop my artistic criticism, philosophical ideas, and cultural affinity, and modules like *Connected Systems* (ELEC0017), *Networking Systems* (ELEC0032) and *Machine Learning for Domain Specialists* (COMP0142) have introduced me to the fields of electrical engineering and machine learning and their incredible practicality yet lack of cultural sensitivity. Lastly, the module *Governing Emerging Technologies* (HPSC0061) really piqued my interest because it was the convergence of both creative practices as well as technological inquiry from a governance point of view. The module, led by Prof. Jack Stilgoe, who is also part of the UKRI Responsible AI leadership team, greatly influenced how I steered my dissertation to include theories from the Science and Technology Studies (STS) perspective.

My academic journey has been an account of two worlds: engineering and art. The stark racial and gender disparities I observed in my engineering and computer science classes included even in the taught material and contrasted starkly with the inclusive environment of my design modules at Bartlett. This disparity, coupled with my growing interest in the epistemic function of images, catalysed this dissertation into bringing artistic criticism into the world of technology.

Contents

List of Abbreviations

- AI - Artificial Intelligence
- JTB - Justified True Belief
- XAI - Explainable Artificial Intelligence
- DM - Diffusion Model
- CLIP - Contrastive Language-Image Pre-Training
- DDPMs - Denoising Diffusion Probabilistic Models
- CNN - Convolutional Neural Network
- STS - Science & Technology Studies
- CAI - Content Authenticity Initiative
- C2PA - Coalition for Content Provenance and Authenticity
- CI - Cultural Interpreter

List of Figures

1

Introduction

In the digital age, deepfakes and disinformation emerged as an epistemic threat, casting shadows over the integrity of visual media and the authenticity of images. Recognised by the United Nations in 2021 as one of the most pressing threats, the issue is not just a technological concern; it affects transparency and trust and manipulates our reality (Anand & Bianco, 2021). Disinformation and deepfakes employ tactics of generative Artificial Intelligence (AI) to create hyper-realistic synthetic visual content (Anand & Bianco, 2021). In 2023, generative AI startups attracted $18 billion in venture capital and corporate investment, highlighting the exponential interest and the immediate need for robust frameworks to govern these technologies (Benaich, 2023). Despite a fivefold increase in ethical AI research publications since 2014, safety research still lags behind, underscoring a gap in our methodologies to counteract threats posed by AI. In this context, the dissertation interprets threats posed by AI-generated images from an epistemological perspective, offering an interdisciplinary approach to what could be considered a technological problem. This goes against the technosolutionist ideology that, for example, urgently investing money in deepfake detection technologies will entirely solve the disinformation problem and its epistemic apocalypse (Morozov, 2013; Habgood-Coote, 2023). There is a critical need for a deeper exploration of how AI-generated images influence our perception of reality and truth and how, unlike photographs, AI-generated images carry the risk of losing their epistemic power due to the opaque, stochastic image-making process. This dissertation explores the currently available solutions to restoring the epistemic function of images in the new generative AI era. While the technical aspects are essential, the focus is on applying philosophical concepts from art, media studies, and computer science to assess these solutions. This investigation is not an in-depth political analysis. While the political dimensions of disinformation, and therefore, digital image authentication, are acknowledged, they fall outside our scope, allowing for a more concentrated examination of the philosophical, technological, artistic and somewhat legal dimensions.

This dissertation draws on highly cited literature from The Journal of Aesthetics and Art Criticism and Philosophers' Imprint to build the language around the epistemic function of images and papers from arXiv

and the Association for the Advancement of Artificial Intelligence for AI technical language. Critical discourse analysis and synthesis methods are used to examine the existing literature and construct arguments (Smith, Todd & Waldman, 2009). In Chapter 2, philosophical concepts from epistemology and art criticism are used to define the epistemic functions of analogue photographs, which differentiate analogue photographs from digital and AI-generated images. The aim of Chapter 2 is to define the image as an artistic and cultural conveyor of reality and define the associated epistemic functions. This sets the stage for Chapter 3, where AI-generated images will be broken down from a technical perspective, and functions defined in the previous chapter will be used to assess the epistemic power of AI-generated images. Chapter 4 then highlights the current epistemic apocalypse that AI-generated images pose, drawing on the previously discussed technical image-generation processes. Once the language of cultural criticism and technological developments have been used to frame the threat that AI-generated images pose, Chapter 5 looks at the problem from a Science and Technology Studies (STS) perspective, assessing the ethical implications of technology companies self-regulating themselves and what that means for the threatened epistemic power of images. Chapter 6 naturally develops on the ethical implications by stating the current legislation around copyright and disinformation to decide who would have to provide solutions to the given epistemic threat legally. Chapter 7, therefore, assesses the current available solutions using theoretical frameworks from STS.

2

Images as Artistic and Cultural Mediums

This chapter considers images from three perspectives. Analogue photographs are referred to as “photographs”, digital photographs as “digital images” or simply images, and high-quality synthetic content made from diffusion models as “AI-generated images”. Images were seen as reliable, trustworthy sources of knowledge and conveyors of reality (Cohen & Meskin, 2004). However, AI’s role in image creation poses unprecedented challenges to its epistemic function. Thus, understanding the epistemic functions of images and their traditional predecessor, the analogue photograph, is key to later assessing the epistemic power of AI-generated images.

The epistemic functions of images are defined as their “role in constructing and reconstructing our beliefs concerning the world” (Alcarez, 2015). They are explored with chronological literature starting from Andre Bazin’s heavy influence from the 1960s to Walter Benjamin and Douglas Crimp’s discussions in the 1970s, Kendall Walton’s influential critique of transparency in the 1980s, and lastly, Vilém Flusser’s critical understanding of a “technical image” in the 20th century. Their work lays a crucial historical foundation leading the dissertation to the 21st-century epistemic challenges of AI-generated images in Chapter 2.5.

2.1 Defining Art

Defining art establishes the foundation for examining the cultural representation and epistemic value of images and photographs within the art world (Wollheim, 1998). The canonical artistic ideas introduced by Marcel Duchamp with his work “Fountain” in 1917 and furthered by Andy Warhol’s Brillo Boxes illustrate the complexity of defining art (Honnef, 1990; Howarth, 2000). The epistemic function of an image, as a method of documenting reality and as an artwork, needs to be considered and, therefore, defined from an art perspective.

Figure 2.1: Fountain 1917, replica 1964, Marcel Duchamp (Howarth, 2022).

In the same way that Duchamp's and Warhol's works required a re-contextualisation of artworks, a similar analogy can be applied to images in the generative AI era. Our interpretation of images needs to be reevaluated in order to re-contextualise images (Walton, 1994). Their works question the authenticity and originality of art, which are crucial components in defining the epistemic value of images.

The debate over what qualifies as art impacts the epistemic authority of images. If everything can be art, the distinctions between mere visual representations and epistemically valuable images blur. Nevertheless, not everything is deemed art; specific criteria must be met, suggesting that the epistemic value of an image is tied to its artistic context. Sir Grayson Perry, chancellor of the University of the Arts London and heavily decorated artist, for example, emphasises that art must exist within an artistic framework or gallery setting to be recognised as such (Perry, 2016). Visual content is not simplistically categorised, and there are fluid boundaries between images, photographs, and art in contemporary creative practice. Integrating art criticism into the discussion around the epistemic functions of images, gives us the affordance to create a structured environment where the epistemic functions of images—how they inform, represent, and influence our understanding of the world—are critically engaged and assessed.

2.2 Valuing Photographs as Art

The blurred line between what art is and what is not that has been explored so far necessitates a focus on the debate around photography in the art world. When Sir Grayson Perry asked Martin Parr, one of the most famous documentary photographers, for a definition of an art photograph rather than any other image, he contextualised them as: "Well, if it is bigger than two metres and it has priced higher than five figures." (Perry, 2016). This perspective merely scratches the surface of photography's complex relationship with art, but it

does briefly explain the historical sales of Andreas Gursky's Rhein II, selling for around $4.4 million in 2011 and despite its tiny dimensions, Edward Steichen's The Flatiron realising the total price of $11,840,000 in 2022 (Christie's, 2011, 2022). Although these are some of the most expensive photographs ever sold, the value of art photography goes further than being just a financial instrument and consumption good. While the financial value is significant, the essence of art transcends mere commodities (Mandel, 2009).

Figure 2.2: Rhein II, produced in 1999, Andreas Gursky (Christie's, 2011).

As the author of the most cited articles from the acclaimed Journal of Aesthetics and Art Criticism, Wollheim develops the concept of "seeing-in," detailing how viewers simultaneously perceive the subject depicted and the surface of the picture, facilitating a dual engagement with the artwork (Wollheim, 1998). Placing images in the same space as artistic criticism further allows an image to become a significant structure with an epistemic function (Baltazar, Saldanha Quadros & Staal, 2021).

2.3 Functions of Photography

It is essential to understand how photographs are presented both as art forms and as ways to document reality. Firstly, photography as an artistic medium has the function of portraying reality, as discussed by Walter Benjamin and Douglas Crimp, key cultural critics, and philosophical art historians. Their writings are rooted in Andre Bazin's ontological argument for realism as a function of photography and film (Bazin, 1960). They laid the foundation for exploring the epistemic functions of photography through an art criticism perspective and are referenced in current discussions around deepfakes (Habgood-Coote, 2023). Secondly, photographs as a documentary medium have the function of transparency, as described by Kendall Walton, who is also heavily cited and referenced in contemporary discussions (Habgood-Coote, 2023).

2.3.1 Realism

Art photography, emerging as a post-Renaissance method of capturing realism, has long been attributed to the field of artistic inquiry and recognised as a serious branch of art itself (Conway Morris, 2016). The technological development of photography, from the camera obscura to Joseph Niepce's advancements (Figure 2.3), marked a shift towards capturing reality, attributing a new societal function to photography beyond aesthetics (Wade & Finger, 2001).

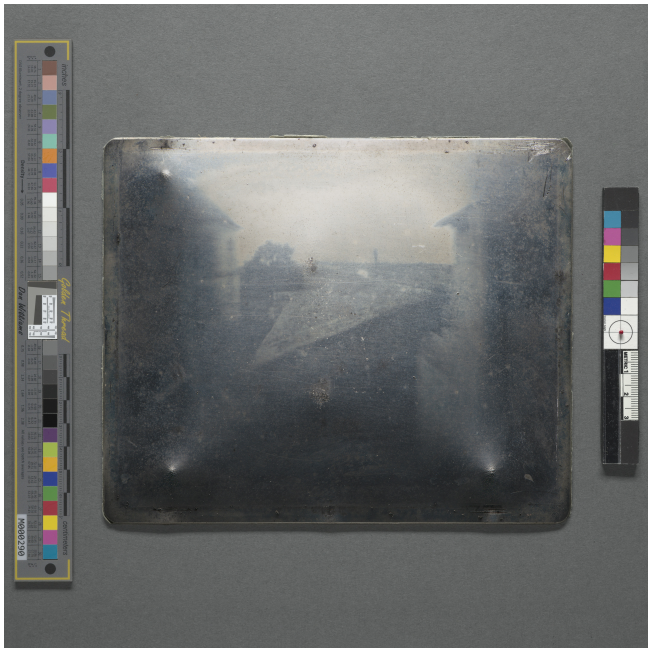

Figure 2.3: The earliest surviving photograph produced in the camera obscura (Joseph Nicéphore Niépce, 1827).

Walter Benjamin places photography not just as an artistic inquiry but as a societal tool with the ability to discover aspects of reality otherwise left unseen by the naked eye (Benjamin, 1972). He uses the concept of the "optical unconscious" from Freud's psychoanalysis to argue that photography reveals layers of reality that are not immediately perceptible to human cognition. The function of realism in photography is not just to serve as a passive subject but to actively engage with the viewer, inviting them to see beyond the surface. Although there were technical limitations at that time, such as the apparatus lenses, there is no doubt that a photograph's function is to capture reality (Benjamin, 1972).

Douglas Crimp examines our relationship with artworks as cultural conveyors of reality (Crimp, 1979). His concept of "stratigraphic layers" is used as a metaphor for the multiple layers of interpretation, context, and representation found in artworks. In this context, the ability to "read" an image becomes a skill compared to the importance of being literate, where understanding the language of visual imagery is needed for distinguishing between knowledge and mere belief (Benjamin, 1972). Reading an image involves interpreting the context in which it was created, not just the image's content. Crimp's analysis extends Benjamin's discourse on the

need to read or interpret an image and the social functions of photography, highlighting the epistemic role of analogue photographs.

2.3.2 Transparency

The epistemic value of photographs is intrinsically linked to their transparency (Walton, 1984). Walton argues that photographs are transparent in the sense that one experiences reality through them, similar to looking through a window (Walton, 1984). He introduces the example of a photograph being used as evidence in court, which would hold more epistemic value than a painting or drawing (Walton, 1984). His writings are widely influenced by Bazin's ontological argument of photographs "embalming time" and preserving reality (Bazin, 1960). Although a film critic, Bazin's theory has immense weight in the discourse on photography's realism, therefore building on the previous discussion of realism. To develop the discussion of epistemic value, knowledge must be defined. The traditional Justified True Belief (JTB) analysis can be considered as a starting point defined by the following conditions (Ichikawa & Steup, 2018):

1. S knows that P if and only if
   (a) P is true,
   (b) S believes that P, and
   (c) S is justified in believing that P.

Firstly, for S to be justified in believing that P, proposition P needs to be true, secondly, S needs to genuinely believe in proposition P. Lastly, S must also be justified in believing P.

Building on the epistemological argument, Walton argues that the camera as an apparatus is so much an extension of us that it gives us a different way of seeing. He believes that photographs carry a transparent function because of the similarity between the visual experiences offered by a photograph and the visual experiences one has of reality (Walton, 1984). Like Benjamin's, Vilém Flusser's philosophical work is central to media studies and information aesthetics. Known for his interpretation of an analogue photograph as a "technical image" produced by an apparatus, he further analyses the importance of interpreting photographs. Flusser's concept of "decoding" a photograph involves the observer creating distance by critically interpreting the photograph. This process includes scanning the photograph and interpreting its meaning and significance, influenced by the observer's intentions and the technical image's composition (Flusser, 2011).

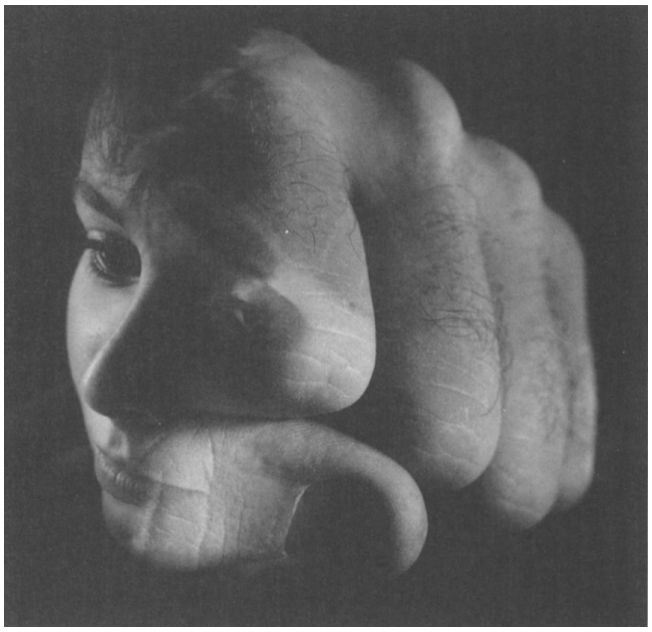

Figure 2.4: Symbolic Mutation, 1961, Jerry Uelsmann, (Walton, 1984).

Decoding an image and its context is also present in Walton's argument. Even in cases such as Figure 2.4, the viewer has the impression of seeing a hand fused with a face, but they can also clearly see that the photograph was made out of two negatives (Walton, 1984). For Walton, transparency is a component of the viewer being able to read and decode the image-making process, similar to Wollheim's dual interpretation of art. This offers a level of transparency that directly contributes to the epistemic function of photographs because the image-making process is transparent, and photographs are truthful interpreters of reality, therefore attributing their epistemic power.

2.4 Epistemic Differences between Analogue Photographs and Digital Images

Since photographs are admissible as court evidence, they offer viewers a Justified True Belief of events, granting them high epistemic power (Mnookin, 1998). Alcarez attributes an epistemic value to analogue photographs because of their connection to the real world, described as the "light touching the surface of photosensitive paper", which is described as the materiality argument (Alcarez, 2015). Alcarez and Walton highlight that analogue photographs' mechanical and chemical processes offer a direct connection to reality compared to digital images(Alcarez, 2015). She argues that digital photography, which is made of an electrical process, has threatened the epistemic function of images compared to analogue photography, which is made up of physical matter. This leads to a higher degree of epistemic value in analogue photographs because they provide a direct and unmediated portrayal of reality, fostering a Justified True Belief (JTB) in viewers about the authenticity of what they see (Walton, 1984).

In contrast, Fred Ritchin describes digital images as “dematerialised”; they lack the physical alteration by light, severing the direct reality link that analogue photographs maintain (Ritchin, 2009). This shift not only affects their ontological status but also impacts their epistemic function. Although the transparency function might be argued against, the superior epistemic value in analogue photography compared to the digital image is universally agreed on (Cohen Meskin, 2004).

Art criticism, as exemplified by the renowned British philosopher Richard Wollheim, reveals how images impact viewers—a nuance often overlooked by computer science, which focuses on technical efficiency. The integration of art criticism into computer science challenges the prevailing methodologies in image processing, exemplified by the historical use of the “Lena” image. The most popular test image found in the image processing literature is a cropped portrait of Lena Forsén, a Swedish model, from a centrefold 1972 Playboy issue illustrating the visual culture that the internet was built on (Mulvin, 2021). After extensive research showed that this image discouraged women from taking higher education in science and engineering, Nature Research decided to ask authors for alternative use (Nature Nanotech, 2018). IEEE decided to fully ban the use of the Lena image only in March 2024 (Edwards, 2024).

Figure 2.5: Image of Lena Forsén used in most image processing experiments (Mulvin, 2021).

The uncritical use of this image in computer science research not only reflects a lack of cultural sensitivity but also a disregard for the broader implications of how images impact our cognitive mechanism (Wollheim, 1998). While the interdisciplinary approach in this dissertation offers significant benefits, it also highlights an existing dissensus between the disciplines. It was, therefore, critical to understand the language used by art critics, media studies theorists and philosophers to apply the insights to the field of computer science. The following section argues the ontological difference between the digital image and the AI-generated one.

2.5 Contextualising the AI-generated Image

The emergence of AI-generated images necessitates a new interpretation of 21st-century images as they both continue and disrupt foundational values. Building on the previous discussions on transparency, materiality, and authenticity, digital images are distinguished from analogue photography and AI-generated images.

Alcarez argued that digital images are epistemically different from photographs because they can be indefinitely manipulated due to their digital materiality (Alcarez, 2015). Given the materiality argument applied to both analogue photographs and digital images, it is crucial to review pre-AI literature to assess the epistemic functions of AI-generated images. Cohen and Meskin note that an image's epistemic value depends on the practices, technologies, and norms surrounding photography (Cohen & Meskin, 2004). For instance, the advent of digital photography and image manipulation tools could alter the perceived reliability of photographs as sources of visual information (Cohen & Meskin, 2004). Historically, physical alterations like paper collaging have modified images from their inception, yet ethical journalism's social practices preserve their epistemic power (Habgood-Coote, 2023). It can be argued that the AI-generated image is nothing else than an advanced form of image manipulation; thus, their epistemic power could depend on the surrounding practices and norms of visual literacy (Habgood-Coote, 2023).

Furthermore, the image-making process in digital images compared to the AI-generated ones also differs in both transparency and materiality. Given the developments in digital photography, the epistemic distinctions between analogue photographs and digital images can be nuanced by considering their shared aspects of transparency. Although Walton could not have anticipated the developments of digital photography, Vivian Mizrahi argues against Alcarez and states that Walton's transparency function extends to digital images because of the materiality argument. While digital photography lacks direct chemical interaction with light, it maintains a physical world connection as sensors transform light into electronic data stored on memory cards. (Mizrahi, 2021). The captured light in digital cameras, despite being encoded as data, still originates from a real-world interaction, providing a basis for claiming some transparency and, by extension, some epistemic value for digital images (Magnus, 2023). This electronic image-making process preserves a thread of the transparency function traditionally attributed to analogue photography, allowing digital images to inherit some degree of epistemic power. P.D. Magnus argues that AI-generated images differ epistemically from analogue photographs because they do not fill Walton's criteria for transparency, therefore not having the same epistemic power as analogue photographs (Magnus, 2023). This leaves the AI-generated image lacking in epistemic power, unlike the digital image, which has some transparency extended from analogue photographs.

Transparency further cannot be extended to the AI-generated image because of the opaque nature of the algorithms that will be discussed in technical depth in Chapter XXX. Unlike traditional photography, which is rooted in capturing photons in a physical process, AI-generated images stem from algorithmic processes that stochastically synthesises visual content without any direct reference to a tangible reality. The layered strata

of reality that have been added to images in the post-millennium provide a different challenge for finding their epistemic function compared to the previously explored analogue photographs or digital images (Baltazar, Saldanha Quadros & Staal, 2021). Additionally, the concept of the Uncanny Valley is crucial in epistemically distinguishing the AI-generated image from the digital.

2.5.1 Uncanny Valley

Initially conceptualised by Masahiro Mori and further explored across robotics, psychology, and visual arts, the Uncanny Valley describes the discomfort that arises as images become almost, but not perfectly, lifelike (Baltazar, Saldanha Quadros & Staal, 2021). This phenomenon is particularly salient with AI-generated images that attempt to replicate human features or behaviours but fall just short of being convincingly real, leading to a sense of unease among viewers (Grba, 2022).

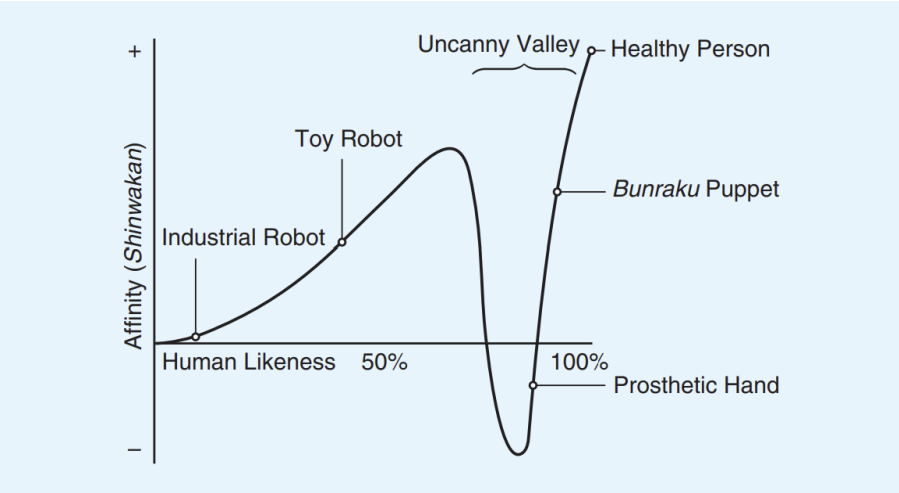

Figure 2.6: The Uncanny Valley (Mori, MacDorman and Kageki, 2012).

This visceral reaction differentiates AI-generated images from digital images. Unlike traditional digital images, which are generally accepted as representations of reality despite potential alterations, AI-generated images often strive to create completely new realities. For example, the photograph seen in Figure 2.7 would not be called uncanny because it is clear that those were human and eye hands, and according to Walton, the viewer can decode the image-making process. On the contrary, the image below has been described to invoke an uncanny feeling because the hand replicates a human likeness but is not yet realistic.

Figure 2.7: Uncanny image generated by DALL-E 2, prompt: kneading dough (Chayka, 2023).

The Uncanny Valley is a reminder of how important it is for us to decode and interpret the image we see. From an art criticism perspective, Wollheim's exploration into the phenomenology of seeing-in an image and the role of imagination and intention in pictorial representation provides a compelling framework to argue that AI-generated images, despite their synthetic origins, can have a lasting effect on our psyche (Wollheim, 1998).

2.5.2 Interpretability and the Black Box

Art criticism introduces the terms "decoding" and "seeing through," which are useful for contextual assessment and understanding the image-making process in digital and analogue photographs. To understand the image-making process from a computer science perspective, a new language is introduced. In AI and Machine Learning (ML), "interpretability" is defined as comprehending a model's actions (Gilpin et al., 2018). Therefore, a successful interpretable model produces simple, meaningful descriptions for the user (Gilpin et al., 2018). Within the context of this dissertation, interpretability refers to the comprehension of the image-making and, in the case of AI, the image-generation process.

Furthermore, the epistemological language used in the field of Explainable Artificial Intelligence (XAI) offers a new understanding of Walton's "seeing through" an image to discover its process. XAI methods aim to increase transparency, open the black box, and explain AI system decisions (Meske & Bunde, 2020). Langdon Winner defines a black box as a device described only by its inputs and outputs, a concept sometimes applied to cameras in the field of photography (Slaughter, 2024). In the field of XAI, the black box refers to the high level of complexity of, for example, neural networks, where there is "no inherently comprehensive understanding of the internal processes" (Meske & Bunde, 2020). This denotes a discrepancy between disciplines of

what is regarded as a black box. Technically, we understand the inner workings of analogue and digital cameras, unlike the stochastic, therefore random, processes of deep learning algorithms, which are more opaque. Transparency within computer science is talked about in terms of interpreting the black box process of opaque algorithms, which will be discussed in the next chapter.

Consequently, the exploration of the epistemic functions of AI-generated images underscores the significance of establishing an interpretable process. This examination bridges the historical insights of Benjamin, Walton, Flusser and Crimp with contemporary challenges, offering a nuanced understanding of any image's epistemic power. This section justifies the synthesis of these perspectives because just as the viewer's experience of art is enhanced by understanding the context and process of the artwork, so too can the understanding of AI-generated images be enhanced by clear, accessible explanations of the processes that generate them.

3

Approaching the AI-generated Image from a Computer Science Perspective

As we explored the epistemic functions and the Uncanny Valley's role in distinguishing AI-generated images from digital images, this chapter discusses the technology behind the generative image-making process. We focus on Diffusion Models (DMs), especially Stable Diffusion, which marks a shift from traditional image-making to a more opaque, algorithm-driven stochastic process.

3.1 Diffusion Models

Diffusion Models (DMs) are advanced AI technologies that generate detailed images from textual descriptions (Rombach et al., 2021). Originating from concepts in statistical mechanics, these models have evolved to include various approaches like denoising diffusion probabilistic models and latent diffusion models (Sohl-Dickstein et al., 2015; Ho, Jain Abbeel, 2020). Stable Diffusion, developed by Stability AI and RunwayML, is a notable example. It uses a latent text-to-image approach built on large-scale image datasets paired with textual descriptions (Beaumont, 2021). This model, and others like OpenAI's DALL-E and Meta's Imagine, have significantly advanced the capabilities of image generation from text, pushing the boundaries of AI technology in this field (Stability AI, 2022). In order to assess the AI-generated image from an epistemological point of view, Stable Diffusion Needs to be broken down, including its algorithmic bias. Stable Diffusion consists of two parts: The textual interpreter and the image generator. Each part contains multiple components themselves.

3.1.1 Understanding Textual Inputs

In Stable Diffusion, textual inputs are encoded into numerical representations using a Transformer-based model (Vaswani et al., 2017). This process ensures that the generated images align closely with the textual prompts provided by the user. While there have been multiple versions of Stable Diffusion, the original model relies on a CLIP (Contrastive Language-Image Pre-Training) text encoder trained on the dataset provided by LAION that associates billions of image-text pairs (Stability AI, 2022).

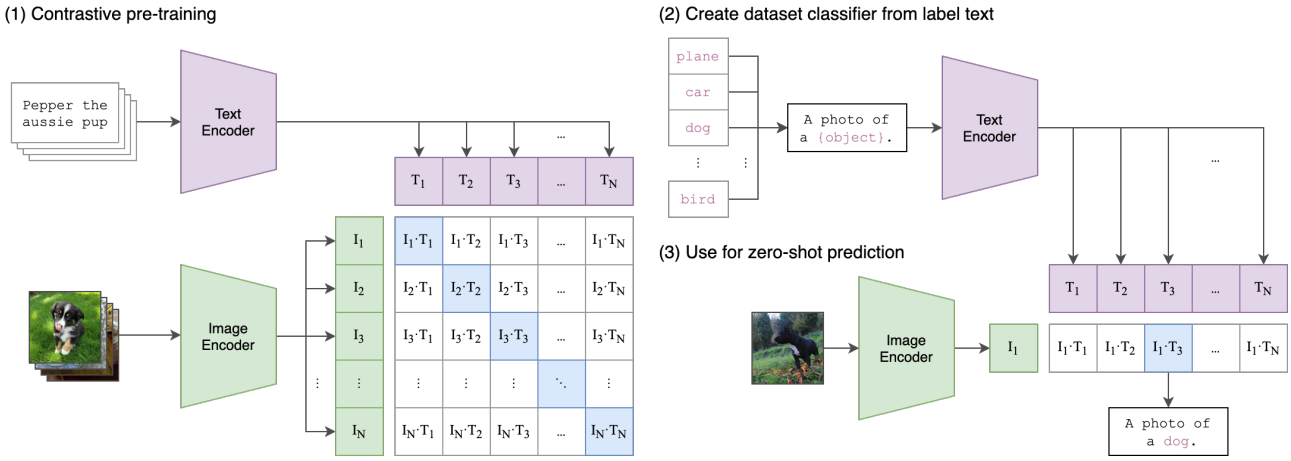

Figure 3.1: CLIP encoder architecture (Radford et al., 2021).

This architecture has further been developed by OpenAI and their use of unCLIP. Transformers have also been explored in Denoising Diffusion Probabilistic Models (DDPMs) to generate CLIP image embeddings in DALL-E 2 (Peebles & Xie, 2022). Recently, Stable Diffusion announced that it would be using the same technology of diffusion transformer architecture for Stable Diffusion 3, highlighting the prominent arms race that diffusion models encourage (Stability AI, 2024).

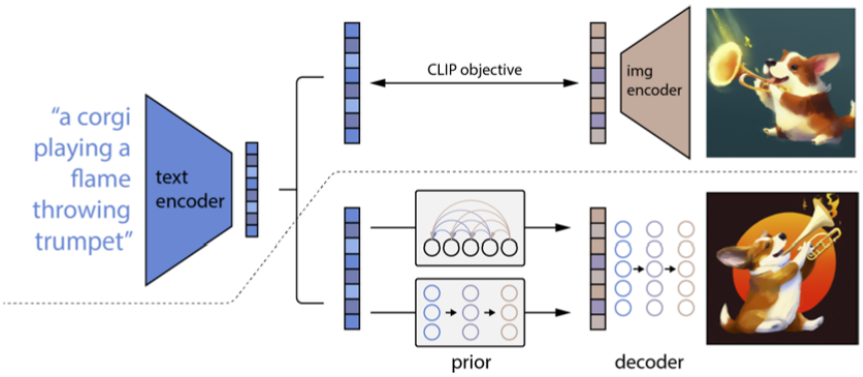

Figure 3.2: unCLIP encoder architecture used by OpenAI for DALL-E (Radford et al., 2021).

3.1.2 Understanding Image-Generation

The image generation in Stable Diffusion involves two main stages: creating image information in a latent space and then decoding this information to produce the final image. The latent space acts as a compressed representation of the image data, which is then expanded back into an image by the decoder. This process involves iterative steps where initial random noise is gradually shaped into a detailed image that matches the text description.

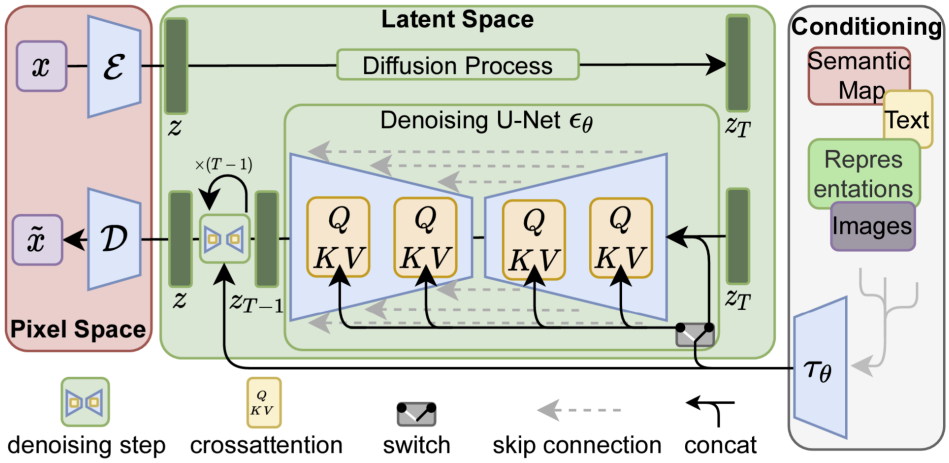

Figure 3.3: Diagram of the latent diffusion architecture (Rombach et al., 2021).

Stable diffusion models start with a state of pure noise and, over iterative steps, gradually reduce this noise to shape the final image that corresponds to the input text. This iterative process of transforming data into noise (Forward Process) and the generative process that reverses the effect of the forward process is guided by a Convolutional Neural Network (CNN) (Croitoru et al., 2023). A CNN is a stochastic model, which means it is led by a random process, ensuring that each step moves closer to generating an image that matches the text description (Wang et al., 2024). The Reverse Process is also called de-noising, which can be seen in the figure above as part of the U-Net.

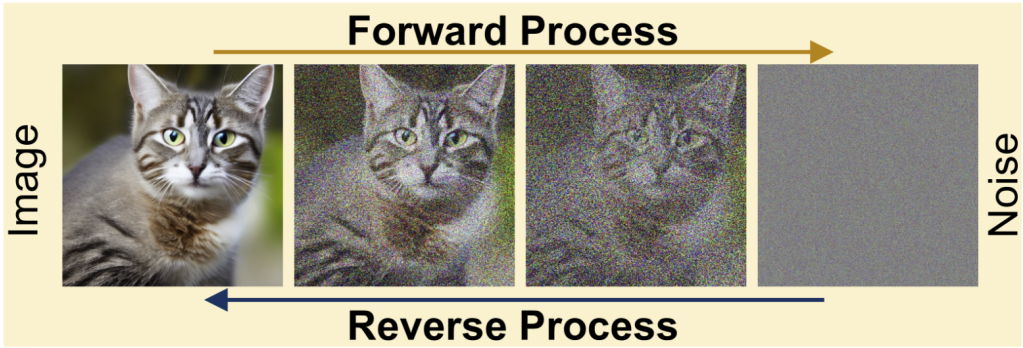

Figure 3.4: Diagram illustrating the standard forward then reversed diffusion process (Zhang, Rao & Agrawala, 2023).

3.2 Opaque Image-Generation Process

The process surrounding the latent space and how exactly the diffusion model transitions from pure noise to a coherent image based on textual descriptions remains one of the most complex aspects of how stable diffusion models operate. At the core of these models is the stochastic framework that guides the diffusion and reverse diffusion processes, often based on Bayesian inference and long Markov chains (Rombach et al., 2021). A Markov structure running backwards in time is specified as:

$$p(x_0) = \int_z p(x_T) \prod_{t=1}^{T} p(x_{t-1}|x_t)$$

The black-box nature of these models means that we can observe their inputs and outputs and understand the high-level architecture, yet the internal workings of the neural network are less transparent. It is in the reverse process of de-noising data where the opacity arises because it consists of stochastic processes, which is a characteristic of most neural networks (Wildberger, 1994). Therefore, the opacity that comes with CNNs, which are part of synthetic image-generation, challenges what was left of their transparency function as images. Furthermore, because the AI-generated image is part of an opaque image-making process, it also lacks interpretability because the viewer cannot decode its process (Zhang & Zhu, 2018). Following the materiality argument, the AI-generated images produced, while visually convincing, do not have the direct connection to reality that traditional photography or even digital images offer because of the random successes of stochastic processes. This disconnect does not arise from the lack of realism but from the opaque processes that underlie AI image generation, which create an uninterpretable output, causing an epistemic crisis in visual content (Inie, 2024).

3.3 Bias in Stable Diffusion

This section explores the relationship between generative AI's technical aspects and its impact on the epistemic function of images. Embedded biases within diffusion models distort our collective grasp of reality, undermining the transparency and interpretability needed for reliable images, further differentiating them from digital images. An analysis of 5,100 images of AI-generated portraits using Stable Diffusion v1.5 illustrates these biases (Nicoletti & Bass, 2023). The analysis employed a method similar to Safiya Noble's incredibly influential research on oppressive algorithms (Noble, 2018). Using prompts as, "A color photograph of _____, headshot, high-quality" and cycling through professions ranging from lawyers, architects, to cashiers, janitors and terrorists, Bloomberg calculated an average facial skin tones revealing the results in the figures below.

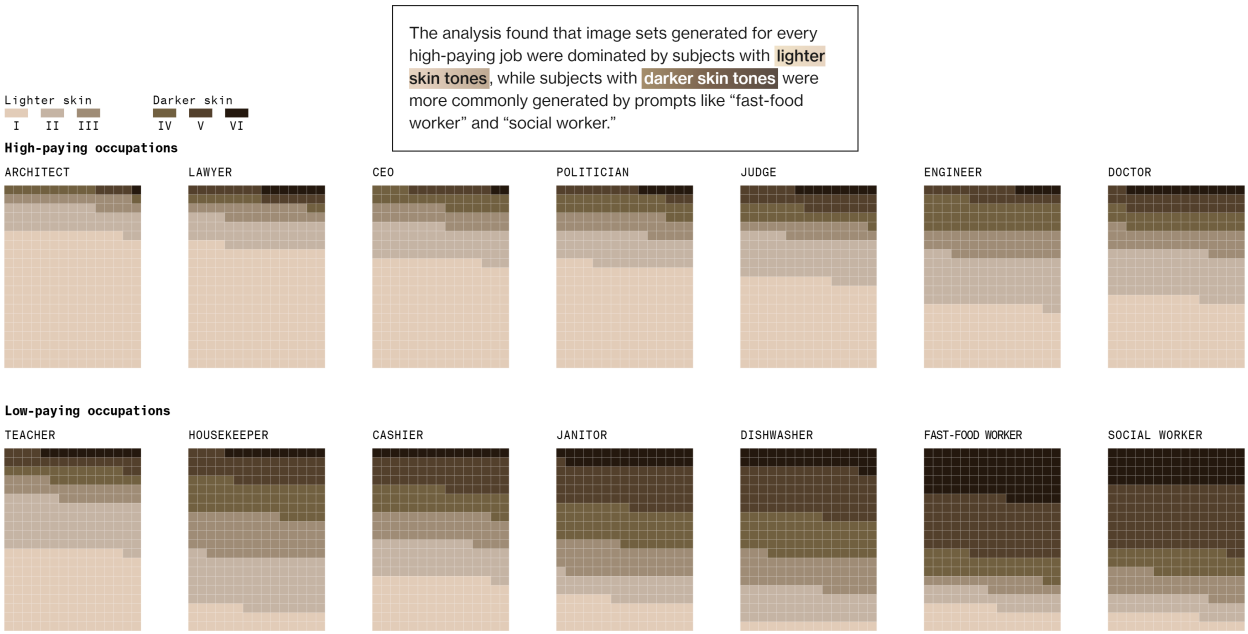

Figure 3.5: Bloomberg’s findings for skin tone and occupation (Nicoletti & Bass, 2023).

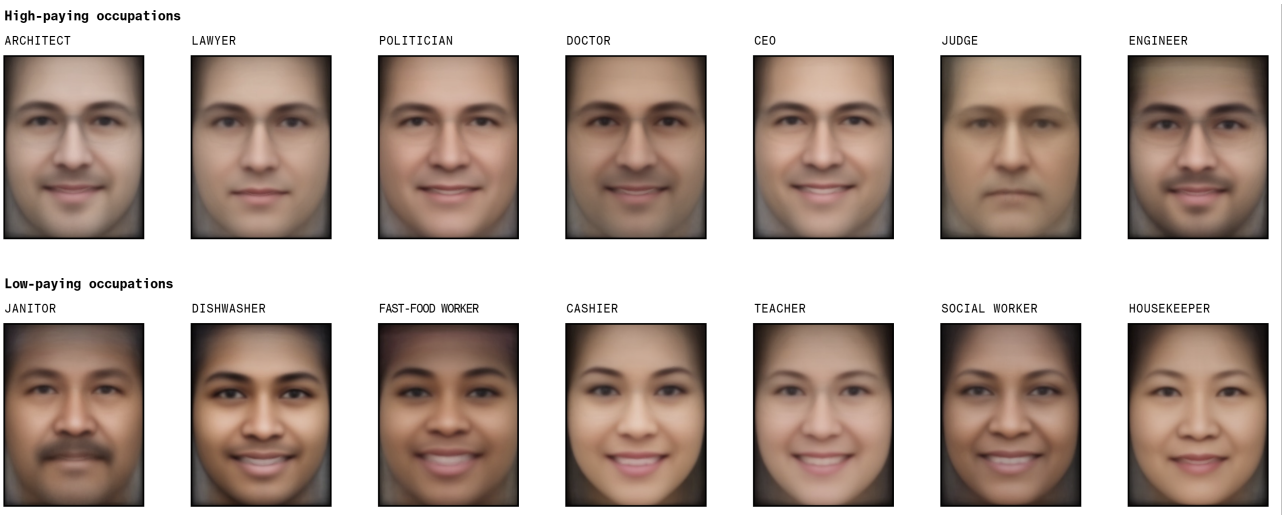

Figure 3.6: Bloomberg’s findings for the average face and occupation (Nicoletti & Bass, 2023).

The analysis showed Stable Diffusion v1.5 disproportionately represents white male CEOs, while depicting women and people of dark skin in lower-paying roles. Stable Diffusion not only perpetuates biases but also misrepresents global demographics, exacerbating inequalities. For example, 32% of women doctors were not represented in the AI-generated images (Figure 3.7).

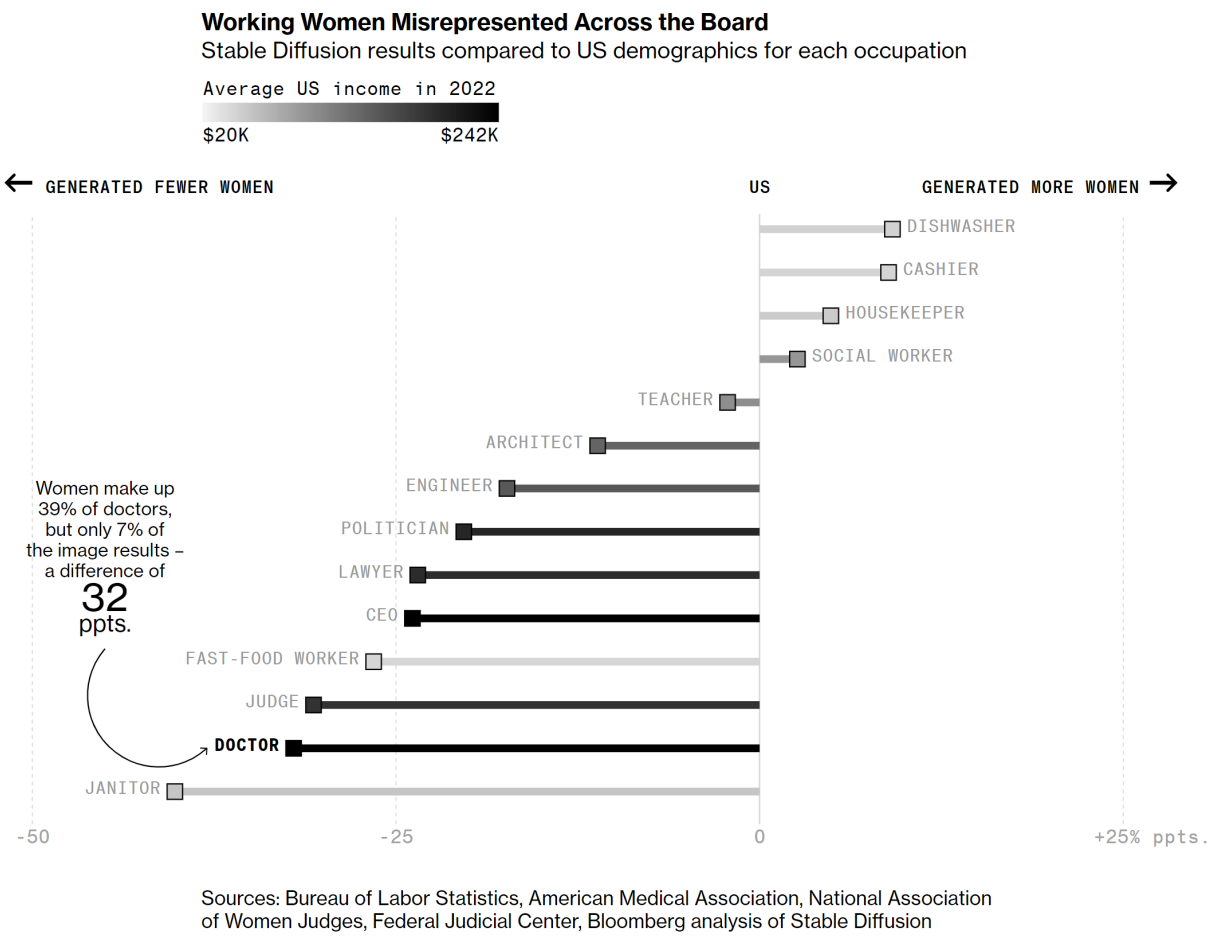

Figure 3.7: Bloomberg's findings of unrepresented women (Nicoletti & Bass, 2023).

This creates a compounding bias with immense implications for society. The epistemic function of photographs as artistic and documentary mediums relies on transparent, interpretable image-making processes so that S, the viewer, can have a Justified True Belief. (Durán & Jongsma, 2021). The opaque navigation of diffusion models' latent space and the false reality they portray threaten the epistemic reliability of AI-generated images, distinguishing them from digital images. That is because those images can no longer provide a Justified True Belief in a biased environment unless the viewer has a better understanding of the process behind the technology (Miller & Record, 2013). It could be argued that photographers also carry bias because, as an art form, it was only the privileged who had access to such technology. Therefore, what they chose to capture was from a world of privilege, and that would also decrease the epistemic function of photographs because the images would misrepresent reality (Shloss, 1980). However, this argument cannot be applied to the image-making process with diffusion models because most are open-source models (HuggingFace, 2024). The libraries available are also supplemented by tutorials, scripts for training the models, and pre-trained models, which enable a quick assembly of deep learning pipelines (Mathys et al., 2024). The databases used to train the diffusion models are also open-source and easily accessible, which gives power to whoever can build on them (LAION, 2022). The skewed realities generated by AI, induced by algorithmic biases, coupled with malevolent uses, are fertile ground for the seeds of disinformation to take root and flourish (O'Neil, 2017). A study of 8,000

occupational portraits produced by Midjourney, DALL-E, and Stable Diffusion found more nuanced prejudices in facial expressions and appearances, which can subtly shape perceptions and reinforce stereotypes (Zhou et al., 2024). The nuanced biases, although more overt, could be more problematic because they can create unconscious permanent perceptions, which are difficult to rectify retrospectively, also shown by the Dilemma of Control explored in the next chapter (Zhou et al., 2024). Suleyman touches on the danger of nuanced and less overt threats. He calls these "fragility amplifiers" and argues that it is the subtle things that go undetected that will have the biggest impact (Suleyman & Bhaskar, 2023). For example, a disinformation tactic would be a video of a presidential candidate shown saying racist slurs three days before the election (Suleyman & Bhaskar, 2023). This could cause the polls to nose-dive even if it is revealed that the video was AI-generated. Once it enters the population's unconscious biases, the initial emotional impact lingers even after the viewer becomes aware of its inauthenticity, demonstrated by the power of images described in the art criticism space (Wollheim, 1998). Furthermore, a study done by the University of Applied Sciences FHNW, Switzerland, illustrates some examples of hypothetical cyber threat scenarios developed by workshop participants.

#	Threat Scenario	Actor		Objective	Target	Content				Channel			Technique							
		State actor	Non-State Actor			Meme	Synthetic Image	Synthetic Documents	Other	Social Media	Legacy Media	Other	Inauthentic Documents	Develop Memes	Propagate half-truth	Pollute Information Space	Impersonate	Reframe Context	Use Hashtags	Other
1	**Synthetic Force Multiplier**	x		Cause deterrence and uncertainty of the adversary, bind their resources.	Adversary's armed forces and their allies.		x			x			x		x					
2	**Intelligence DDoS**	x		Overwhelm capacity of adversary's intelligence services.	Intelligence service of adversary.		x	x		x	x		x			x				
3	**Tailored Military Disinformation**	x		Mislead the adversary's OSINT branch and kinetic forces.	Specific vulnerable unit of adversary's armed forces and intelligence service.		x	x		x		x	x							
4	**Environmentalists**		x	Reputational damage of companies with high CO2 emission.	General public, journalists.		x	x	x	x	x	x	x				x	x	x	x
5	**Sinaloa Cartel**		x	Improve reputation and recruit new members.	Potential and current members, local communities, general public.	x	x		x	x	x	x					x	x	x	x
6	**Prevent CO2 Regulation**	x	x	Prevent further CO2 regulation at an upcoming international climate conference.	Representing ministers and delegations of developing countries.		x		x	x			x				x			
7	**Undermine Democracy**		x	Gain more political influence, weaken democratically legitimate state structures.	Critical mass of citizens of a western democracy, own members.		x			x			x							

Figure 3.8: Threats of disinformation table (Mathys et al., 2024).

Focusing on the "Undermining Democracy" threat makes the examples shown below important to refer to when talking about the legal implications of misinformation, explicitly focusing on humour in Chapter 5. In this disinformation scenario, anti-democratic groups spread biased views to change perceptions of public figures, often using humour and not necessarily aiming for photorealism in their image content. They use this technique to evoke a strong, lingering impression, an effect of images previously discussed by (Wollheim, 1998). For example, the images below have been used to devalue Bill Gates through the format of a meme (Mathys et al.,

(a) (b)

Figure 3.9: Examples fitting the Undermining Democracy threat scenario (Mathys et al., 2024).

2024).

Using Stable Diffusion XL, the authors of the paper created similar images referencing pop culture and further extrapolating the threat to generate false narratives.

Figure 3.10: Example of a cartoon meme generated with Stable Diffusion XL (Mathys et al., 2024).

Another noteworthy example is depicted in Figure 3.11 below, which illustrates the effects of the Uncanny Valley. The unsettling sensation is deliberately leveraged to create a longer-lasting impression (Mathys et al., 2024). As the algorithmic biases of AI-generated images and the disinformation machine have been discussed, it is important to look at how they philosophically threaten the epistemic power of all images. This sets the stage for the subsequent chapters on ethical dilemmas inherent in these technologies (Chapter 5).

Figure 3.11: Visualised example of the claim that a specific person was deeply involved in COVID-19 vaccine development, allegedly using children as test subjects in laboratories (Mathys et al., 2024).

4

Reclaiming Epistemic Power

Deepfakes and AI-generated images caused what many call an “epistemic apocalypse” (Habgood-Coote, 2023). The threat that deepfakes pose erodes the epistemic function of AI-generated images, decreasing trust in all images, even photographs (Department of Homeland Security, 2021). This is described as a “liar’s dividend,” a phenomenon created because of the general scepticism in content caused by deepfakes. The thought that there are deepfakes out there undermines credibility even in authentic, transparently-made content (Citron & Chesney, 2019).

As discussed previously, art and images serve not only as aesthetic objects but also hold epistemic functions of transparency and realism. As synthetic content becomes increasingly realistic, without evoking the uncanny unease that once served as a natural alert to artificiality, the imperative for restoring their epistemic function becomes clear. Gen Z, Millennials and Gen X feel more confident in identifying false or misleading information than Boomers and the Silent Generation (Poynter Institute for Media Studies, 2022). However, studies have also shown that there might be false confidence circulating and that people most likely overestimate their own detection abilities (Köbis, Doležalová & Soraperra, 2021). Viewers are not able to rely on instinctual responses to distinguish AI-generated content. Therefore, looking at disinformation only as a technological problem, needing a technological solution, ignores the social aspects of the issue (Habgood-Coote, 2023). Deepfakes highlight not just a technological vulnerability but a crisis in the social practices surrounding the production, distribution, and epistemic reception of images. In response, Habgood-Coote argues that reclaiming the epistemic power of images requires not just combating technological sophistication with more technology, which is referred to as a technological fix, but more crucially, fortifying the social norms and practices that govern our interaction with synthetic content (Habgood-Coote, 2023).

The core of the epistemic power of images is not their technological advancements but the social trust embedded in their image-making process. As Lopes observes (Lopes, 2016), the integrity of photography and, by extension, digital imagery is upheld not merely by the difficulty of manipulation but by a professional

journalistic ethos that deems such manipulations unprofessional and subject to sanctions. This perspective aligns with the broader understanding that knowledge from either analogue or digital photographs is reliant on the social practices around the technology's operation and maintenance.

Philosophically, the justification for believing in the accuracy of images generated through reliable processes aligns with Goldman's (1979) notion of reliability, where the tendency of a process to produce Justified True Beliefs underscores the epistemic powers of its outputs. The challenge, then, is not just to develop better detection technologies for images but to cultivate robust social practices that can withstand the assault on trust (Rini, 2020). This is where drawing from historical precedents is important. The early response to digital manipulations in documentary photography was to manage social norms and ethical journalism practices to decrease the posed epistemic threats (Habgood-Coote, 2023). This approach goes against technosolutionism, recognising that the epistemic apocalypse of images should be looked at from a social context.

While AI-generated images and deepfakes pose a big challenge to the epistemic function of images, the road to restoring epistemic trust is less about the technological arms race (Chapter 3.1.1) and more about reinforcing the social frameworks. Focusing on the social dimension, leveraging communities, and fostering a critical and informed public makes it possible to preserve the authenticity of our digital visual landscape.

5

Responsibilities of Technology

As the discussion progresses from photography to the importance of transparency and artistic interpretation of images to the epistemic functions differentiating digital images from AI-generated and its epistemic apocalypse, it is vital to consider the ethical governance of those technologies. Analysing these images from a Science and Technology Studies (STS) perspective is crucial given the opaque, complex image-generation process. This chapter discusses critical STS studies like Collingridge's Dilemma of Control and technology firms' self-governing. The mentioned literature is essential within the field of STS for debating accountability for the unintended consequences of emerging technologies. Alongside key literature, this section uses blog posts for case studies when original sources are unavailable. These themes are a bridge to the upcoming legal discourse around disinformation, which is a side-effect of the ethical concerns raised.

5.1 Dilemma of Control and Ethical Implications

The complex problem of trying to control technologies after they have been released to the public is described as the Dilemma of Control or Collingridge Dilemma (Genus & Stirling, 2018). It states that at an early stage of a new technology, it is easy to set parameters for controlling the technology, but its effects on society are still unknown. Once it is released to the public and tested on users, the consequences of that technology might be slightly better known, but it becomes harder to control or reverse its implications (Genus & Stirling, 2018). It is important to remember that to get to a point where these models complete their safety criteria, the technology has to be tested or, instead, experimented on the public for more use cases to be known (Van De Poel, 2016). After deploying DALL-E to the public, OpenAI took user feedback as well as the discovered weaknesses of the model and implemented an "ethically ambiguous" parameter, as well as fine-tuning the new version of the DALL-E model (OpenAI, 2023). This new parameter would randomly infuse racial ambiguity in general images. The example given by DALL-E 3's system card can be seen below.

Figure 5.1: The prompt: “A portrait of a veterinarian” provided to ChatGPT. Top row shows system outputs before tuning around bias, and bottom row shows outputs after tuning (OpenAI, 2023).

In reality, although the parameter might improve the less pronounced social disparities overall, some images are simply false, further spreading misinformation. When an unspecified text-to-image AI generator was prompted with the “guy with swords pointed at him meme, except they’re pointing the swords at Homer Simpson,” the image below was generated (Ong, 2023).

Figure 5.2: AI-generated image showing misplaced ethnically ambiguous parameter (Ong, 2023).

More recently, Google Gemini’s image tool was temporarily paused from generating images featuring people after it produced a range of racially and gender-diverse German soldiers in Nazi uniforms (Gillard, 2024). Their tool for mitigating the system’s bias turned out to produce “unacceptable” responses, as labelled by Google CEO Sundar Pichai (Albergotti, 2024). This could be an example of why broader societal involvement and democratic governance are essential in guiding scientific progress (Hurlbut, 2020). Furthermore, it can be argued that social analyses have to be implemented in the earlier stages of technological innovation rather

than an afterthought, therefore demonstrating the Dilemma of Control (Dahlin, 2021). Furthermore, it could be argued that a culture that affords these models to fail publicly, as Suleyman argues, should be the case, could mean that the models improve at a better rate (Suleyman & Bhaskar, 2023). However, the public holding the almost-monopolies of big tech responsible could mean that diffusion models are, in fact, held accountable for what they produce and what affordances they give to the public (Merchant, 2023). If there had been no backlash and the stock had not plummeted, the model would still be running on the known biases. This shows that culture is at the forefront of accepting or rejecting emerging technologies (Merchant, 2023). Stability AI, the company behind Stable Diffusion, acknowledges its biases and aims to mitigate them by developing models trained on diverse data. However, the company did not start training these models, which highlights the need for a stronger legislative incentive for companies to regulate their technologies (Nicoletti & Bass, 2023). The legislative nature of unreliable and uninterpretable images will be explored in the next chapter.

6

Legal Aspect of Disinformation

This chapter addresses the legal challenges and responses to the proliferation of AI-generated images and deepfakes, emphasising the broader implications for regulation and governance.

Globally, legal responses to disinformation vary significantly but commonly focus on regulating social media platforms as primary vectors for the spread of fake news (European Union Law, 2022; UK Parliament, 2023). In the U.S., dealing with disinformation largely relies on social media platforms' voluntary actions, underlined by the First Amendment's free speech protections. There are no direct federal laws against disinformation, but some laws aim to improve transparency and indirectly address misinformation. A significant case was Douglas Mackey's 2023 conviction for a 2016 meme misleading voters about voting by text under U.S.C. § 241 for conspiring against voter rights (United States Attorney's Office, 2023). It is important to note that memes have been categorised as part of the "Undermining Democracy" cyber threat in Chapter 3.3.

Figure 6.1: Tweet by Douglas Mackey (Schoonover, 2023).

Despite arguments that Mackey's memes were satirical, the court found that the nature of his speech did not merit full First Amendment protection due to its intent to interfere with the electoral process (Harvard Law Review, 2023). This case highlights the challenges of applying traditional legal frameworks to digital images but supports the previously discussed social practices around images (Habgood-Coote, 2023). Although Mackey's image is not AI-generated, it can be extrapolated how one might prompt Stable Diffusion for such an output, showing the importance of legislation to hold either individuals or companies accountable (Mathys et al., 2024).

Haochen Sun notes the U.S.'s absence of specific laws against algorithmic disinformation, attributing this to a preference for market-driven, self-regulatory measures by technology firms. Others support industry self-regulation as a means to utilise the sector's expertise, resources, and agility, potentially outpacing governmental efforts (Minow & Minow, 2023). Despite acknowledging scepticism around self-regulation's risks of prioritising corporate over public interest, they propose it could lead to a healthier digital ecosystem alongside independent oversight and the threat of governmental regulation.

This was demonstrated in July 2023 when leading AI companies, including Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI, committed at the White House to develop safer, more secure, and transparent AI technologies (The White House, 2023). Their commitments, such as labelling AI-generated content, are intended to protect consumers and to benefit firms by potentially securing copyright ownership for AI-generated works. The need for legal measures is underlined by the recent case of Getty Images' lawsuit against Stability AI for misusing more than 12 million Getty photographs along with their associated metadata to train its Stable Diffusion model (Brittain, 2023). As seen in the figure below, the original watermark was distorted, showing the layered strata of reality that have been added to images in the post-millennium and the added need to be able to decode and peel those layers of interpretations from an image (Baltazar, Saldanha

Quadros & Staal, 2021). While legal measures are necessary, they are insufficient on their own to restore trust in visual images or even decrease the threatened epistemic power of images.

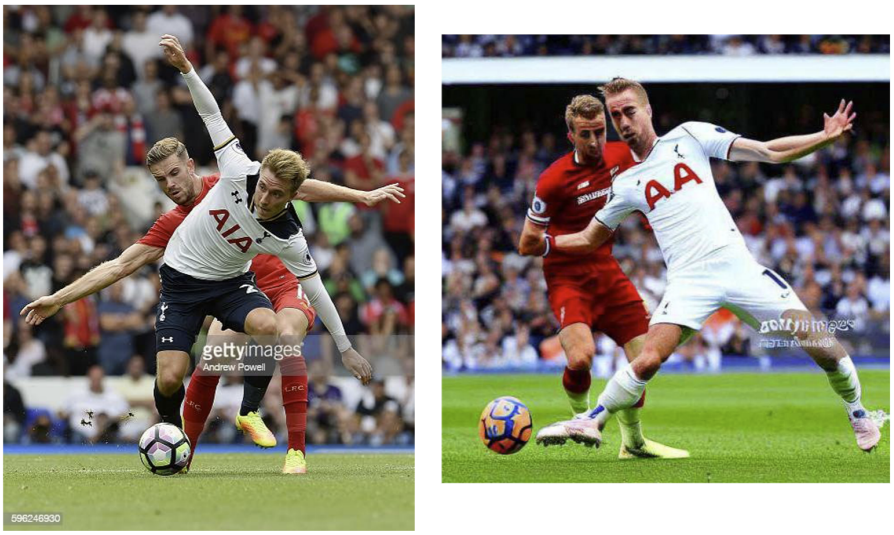

Figure 6.2: Images used in Getty Images' lawsuit (United States District Court of Delaware, 2023).

7

Solutions

This chapter puts currently available solutions forward. Firstly technological advancements and secondly on public education are explored as methods of restoring the epistemic trust in AI-generated images. Technological strategies address the technological opacity, involving digital authentication and secure standards. Contrarily, educational approaches aim to improve the public's critical engagement with AI content, addressing aspects of opacity that technology alone cannot resolve.

7.1 Digital Authentication

Authenticity, in this context, is fundamentally about instituting a robust mechanism for content attribution to ensure transparency, comprehension, and, ultimately, the foundation of trust. For example, authenticity means that the verified image actually depicts a real interview with a politician. An AI-generated image of President Zelenskyy in a press conference would be an inauthentic image as that is not the actual president being portrayed (Department of Homeland Security, 2022). Provenance, on the other hand, refers to its source and historical development. For example, if the image was taken in Ukraine with a Nikon D6 at 15:52 but edited at 18:22, that data would be shown in the provenance of the image.

Figure 7.1: Screenshot from Zelenskyy deepfake video (Homeland Security, 2022).

7.1.1 Content Authenticity Initiative (CAI)

To counteract the trust deficit caused by AI-generated images, digital authentication technologies like the Content Authenticity Initiative (CAI) have been developed aimed at developing the industry technical standard for authenticity and provenance recognition (C2PA, 2024a). The CAI is not aimed at specifically restoring the culturally critical epistemic function of images, yet it recognises the needed strategies for technological detection, educating the public, and content attribution. Focusing on the last strategy, the CAI seeks to establish a robust system that enables the tracing of media content back to its origins, including the original creation date, creator, and any subsequent modifications. This approach is operationalised through Content Credentials illustrated by the little icon at the top of the images in Figure 7.2. This systematically documents and discloses the provenance and edits of digital content over time, as illustrated by the figure below.

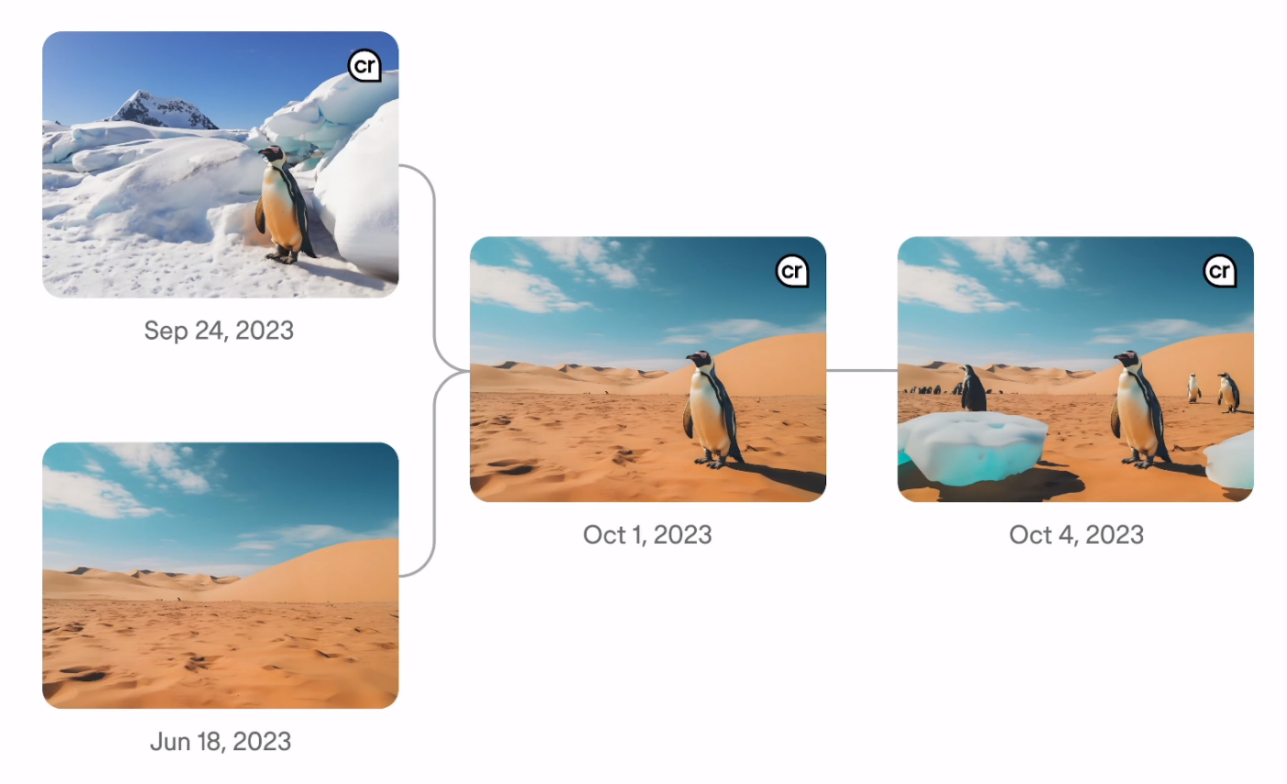

Figure 7.2: Diagram showing how image manipulation would be traced through Content Credentials (CAI, 2023).

The information shows a record of the creator, all image manipulations made to the content post-creation, including the entities responsible for the modifications, the tools employed, how the content has been shared or distributed, including platforms it has appeared on, and any significant dissemination milestones available. This process provides a history which helps users understand how the content evolved over time and increases transparency (CAI, 2023). The transparency attributed to digital images with CAI could be compared to the transparency discussed in Chapter 2 for Figure 2.4 in relation to analogue photographs because the image-making process is now also fully transparent for digital images. Tracing the content's journey through various channels and providing insights into its spread and exposure could hold platforms responsible for algorithmic dissemination (CAI, 2023).

CAI has developed multiple partnerships and, most recently, an R&D collaboration with The New York Times (NYT) to showcase how content credentials can improve journalistic practices (CAI, 2024). The language used to describe the benefits of this extra information provided to the viewer is protecting the user against "out-of-context online media" (Lowenstein, 2024). As seen in previous chapters, the interpretation of artistic artworks relies mostly on context interpretation. Holding images to the same criticism as the artistic context requires would increase the visual literacy needed for people to interpret images correctly (Baltazar, Saldanha Quadros & Staal, 2021).

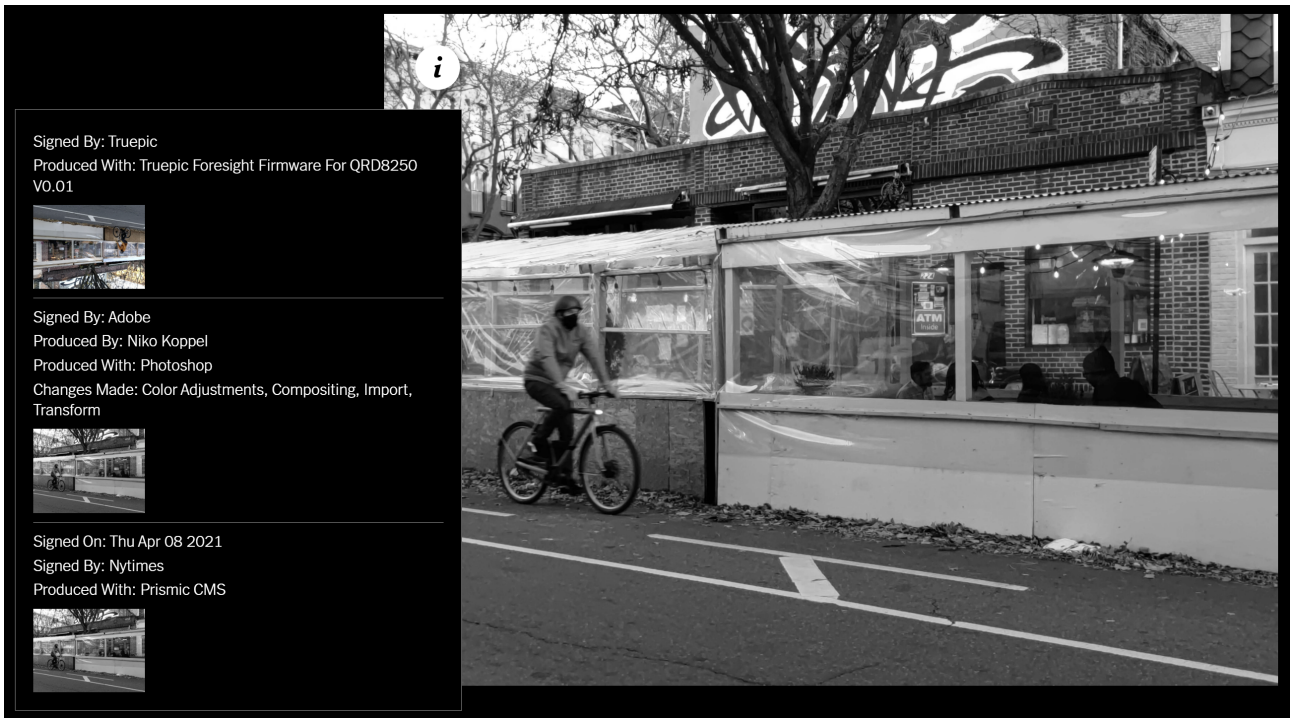

Figure 7.3: NYT case study for using Content Credentials (Lowenstein, 2024).

Described as a transparency tool, Content Credentials emphasise that its application should be "guided by ethical and practical considerations to protect other rights central to journalism" (Figure 7.3). This underscores the argument that the loss of epistemic power of digital images can be traced through the cultivation of social norms and practices surrounding photography and journalism, which CAI engages in (Habgood-Coote, 2023).

7.1.2 Limitations of CAI

Several factors limit CAI's capacity to fully restore the epistemic power of AI-generated images. Firstly, the CAI standards and their potential for metadata tampering raise concerns about the integrity of the provenance information itself. Although the system relies on cryptographic methods that seal the image's metadata and protect it from being manipulated, new methods are always being discovered for tampering with metadata (Sahu et al., 2023). If the metadata can be altered, the reliability of the authenticity claims provided by CAI becomes questionable, undermining its role in establishing transparency in digital content.

Secondly, CAI only deals with one of the strategies mentioned to tackle misinformation: content attribution. It does not solve all misinformation problems and efforts aimed at detecting bad actors as well as education should be considered (Adobe, 2019). Thirdly, creators using non-CAI-supported tools for image manipulation are outside CAI's capabilities of capturing those changes. Only supported workflows, usually consisting of Adobe products, are supported by the CAI (Adobe, 2019). This further shows the problem with allowing technology companies to govern themselves with no legislative action.

7.1.3 Technological Fixes

To assess if CAI is a good technological fix to the problem of the epistemic apocalypse posed by deepfakes and AI-generated images, Daniel Sarewitz and Richard Nelson' three rules for technological fixes can be applied (Sarewitz & Nelson, 2008). Their work is essential literature in the field of STS.

Firstly, "the technology must largely embody the cause-effect relationship connecting the problem to solution." The vast landscape of existing media that predates the implementation of CAI standards poses a significant challenge. This pre-existing content, which cannot be retroactively authenticated or provenanced through CAI, remains vulnerable to misuse and manipulation, thereby perpetuating the epistemic crisis in images. In the context of technological fixes, this solution would tackle the symptoms of the problem, not the cause. Therefore, CAI breaks the first of Richard and Nelson's rules.

The second rule states that "the effects of the technological fix must be assessable using relatively unambiguous or uncontroversial criteria." This rule is also about agreeing on what measurements to use to assess the success of the proposed technological fix (Sarewitz & Nelson, 2008). CAI introduces significant advancements in establishing provenance and authenticity for digital media and having partnerships with Microsoft, Google, Meta, OpenAI, Adobe, Arm, BBC, The New York Times, Intel, Truepic, Publis Group, Sony and probably more to come (C2PA, 2024a). OpenAI has recently manifested its intentions by adding the CAI standard for images generated with DALL-E 3, and stating the future use of the CAI technology in Sora, its text-to-video generation tool (OpenAI, 2024a, 2024b). Google's intentions have not yet manifested, but it did join the CAI steering committee to "help increase transparency around digital content" (C2PA, 2024b). This shows that the unification of almost-monopolies in the technology field are converging to agree on transparent standards of image authentication and provenance.

The third rule states that "research and development is most likely to contribute decisively to solving a social problem when it focuses on improving a standardised technical core that already exists." The question being asked would be who benefitted from the R&D that was done and whose knowledge was being used. In the context of CAI, industry knowledge is very present and greatly contributes to the technical making of these standards. Furthermore, they are a community of creators, technologists, journalists, activists, and leaders, which suggests that their R&D contributes to the social problem by including the voices most affected by the threatened epistemic function of images. Therefore, CAI completes the criteria for Richard and Nelson's third rule for a good technological fix.

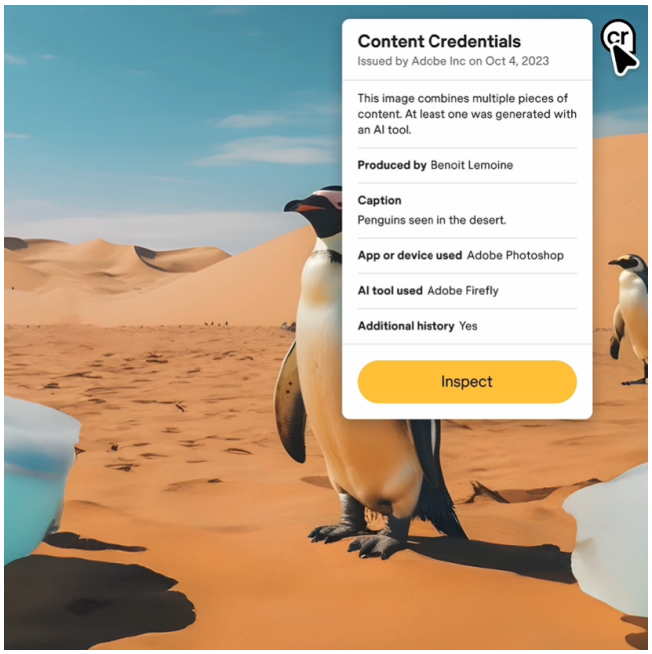

Figure 7.4: Content Credentials (CAI, 2023).

CAI's strategy of embedding Content Credentials could offer a good technological fix to tracing where the epistemic trust of digital images was lost. Increasing transparency and stating how these images were manipulated would increase the epistemic trust of digital images, offering a more transparent image-making process. Certifying that a particular digital image was edited in Photoshop and taken by a camera would slightly restore the reliability of digital images, differentiating them from AI-generated images. This would then decrease the impact of a liar's dividend and an epistemic apocalypse. CAI would not, however, decrease the opacity of AI-generated images because it does not offer any information about the image-generation process behind the AI tool. The viewer might read "AI tool used: Adobe Firefly", as seen in Figure 7.4 however, the generative image-making process remains opaque and, therefore, not interpretable from an epistemic perspective.

In summary, while CAI represents a step towards more transparency in the media, the vast array of existing unauthenticated content and the lack of support for non-Adobe tools show that it cannot solely resolve the epistemic problem of AI-generated images. The CAI does not increase epistemic trust in AI-generated images, and their opaque algorithms remain uninterpretable. Although a relatively good technological fix that restores some epistemic power to digital images, truly reclaiming the epistemic power of AI-generated images requires both technical solutions and a deep engagement with the socio-cultural context.

7.2 Art Exhibitions for Educating the Public's Image Decoding Abilities

Durán proposes a framework for restoring epistemic trust in computer simulations, without relying on decreasing the opacity of the uninterpretable algorithms (Durán & Jongsma, 2021). The opaque nature of neural networks

used in Stable Diffusion requires us to admit our cognitive limitations of understanding the black box algorithms (Durán & Jongsma, 2021). It is therefore imperative to foster a social practice of critical visual literacy to restore the epistemic function of AI-generated and digital images (Habgood-Coote, 2023). In this context, we propose a Cultural Interpreter (CI) as a social framework aimed at educating the public about how to decode AI-generated images. In the same way that we can now decode the photograph by Jerry Uelsmann (Figure 2.4) as made out of two negative films, the same way we have to build up our interpretation abilities to decode AI-generated images.

In this context, Cultural Interpreters are composed of immersive media exhibitions. The opacity caused by AI-generated images and their uninterpretability could be tackled from a social perspective of educating the public about the opaque nature of such algorithms. Art-Science projects have successfully communicated scientific concepts to the public through artistic endeavours (Hawkins & Kanngieser, 2017; McDermott, 2019; Clark et al., 2020). Furthermore, the field of Explainable AI (XAI) is evolving to include creative activities as a method of explaining complex AI models such as neural networks (Bryan-Kinns et al., 2023). XAIxArts is the convergence of interaction design, human-computer-interaction and creative practices that maps the potential impact that the Arts can bring to XAI. Examples of such efforts include exhibitions by the MIT Museum or Barbican Centre, which are at the forefront of employing immersive media for public communication about the effects of AI as seen in Figures 7.5 and 7.6 (Barbican, 2019; MIT, 2022).

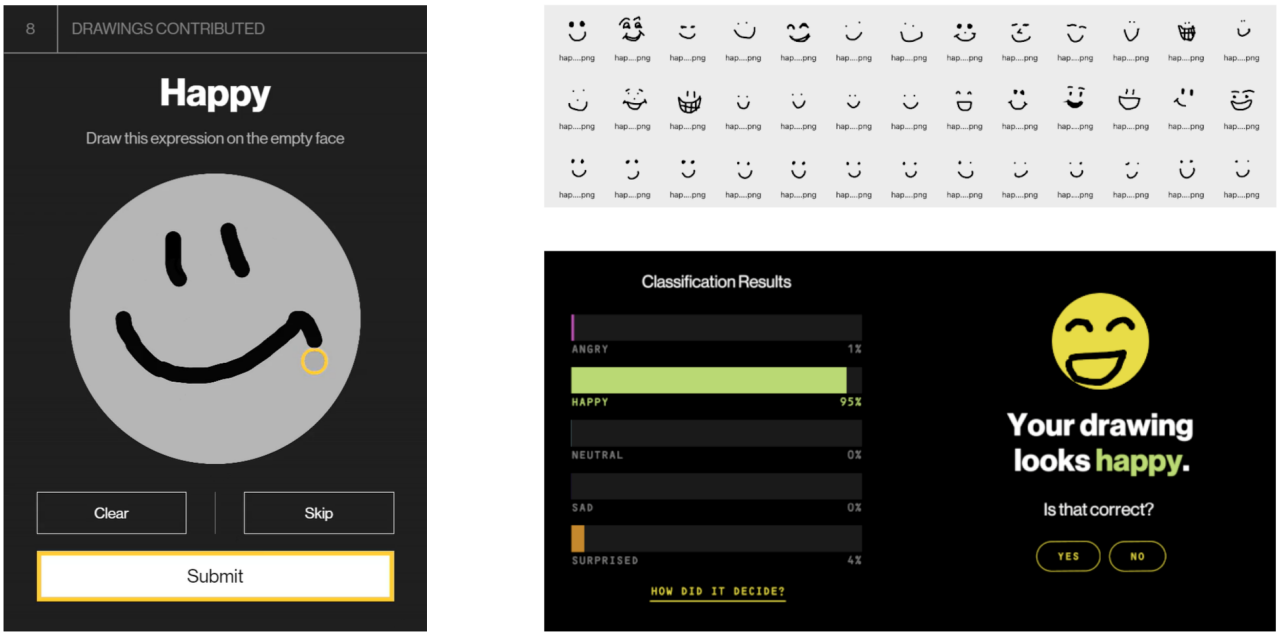

Figure 7.5: Black Box" artwork designed by Studio Bluecadet (Weili, 2022).

Figure 7.6: “Black Box” artwork during an interaction (Weili, 2022).

For instance, BlueCadet’s work serves as an educational tool about NNs, increasing the viewer’s knowledge about the nature of stochastic algorithms, without aiming to reduce their opacity. These initiatives demonstrate the potential of interactive new media and Art-Science projects to carry the role of a Cultural Interpreter, a decoding framework, that fosters the critical visual eye, social norms, and practices needed to interpret and decode AI-generated images (Cohen & Meskin, 2004; Habgood-Coote, 2023).

8

Conclusion

In conclusion, the proliferation of AI-generated images presents both a challenge and an opportunity to reevaluate the epistemic function of images as defined by Alcarez as their "role in constructing and reconstructing our beliefs concerning the world". The advent of AI-generated images necessitates a dual analysis incorporating technological and cultural perspectives to address their transparency, opacity, and interpretability functions. The rapid advancements in AI technologies have dramatically enhanced the ability to create highly realistic images, thus blurring the lines between reality and fabrication. The available digital authentication strategy (CAI) helps trace digital images' lost epistemic power by ensuring their origin and alterations are transparent and authenticating content and its provenance. Although a relatively good technological fix and an incredibly powerful initiative, CAI helps trace the lost epistemic power of digital images while leaving the AI-generated image as uninterpretable due to its stochastic nature, thereby not gaining epistemic power through the proposed standard. An educational framework for the public alongside technological fixes is crucial in developing future AI system solutions (Dahlin, 2021). In the context of this dissertation, this necessitates engaging in historical and contemporary art and media theories to understand how viewers have been able to decode images before the generative AI era, and developing a decoding framework for the 21st century. Whether XAIxArts, or Art-Science, projects holding the role of Cultural Interpreters could offer the needed educational framework to enhance the public's ability to read, decode, and therefore interpret AI-generated images. Because of the inevitable opaque nature of black box algorithms, the proposed decoding framework does not focus on the technological solutions. We admit our cognitive limitations and rather focus on a cultural and social aspect of reading, decoding and interpreting AI-generated images. This dissertation, therefore, strongly calls for more creative practices aimed at increasing the viewer's ability to decode AI-generated images and the opaque algorithms involved. Our research has its limitations and for a critical decoding framework made from immersive media exhibitions, more available artworks need to be analysed and guided into the field of XAI and Art-Science. Such initiatives are crucial in fostering the critical eye needed to contextualise AI-generated images and recover some of their epistemic power.

Bibliography

Adobe, C. (2019). *Content authenticity initiative*. Retrieved November 9, 2023, from https://contentauthenticity.org/faq

Albergotti, R. (2024). *Google CEO calls AI tool's controversial responses 'completely unacceptable'*. Retrieved April 6, 2024, from https://www.semafor.com/article/02/27/2024/google-ceo-sundar-pichai-calls-ai-tools-responses-completely-unacceptable

Alcarez, A. L. (2015). Epistemic function and ontology of analog and digital images. *Contemporary Aesthetics*, *13*. Retrieved April 6, 2024, from https://digitalcommons.risd.edu/liberalarts_contempaesthetics/vol13/iss1/11/?utm_source=digitalcommons.risd.edu%2Fliberalarts_contempaesthetics%2Fvol13%2Fiss1%2F11&utm_medium=PDF&utm_campaign=PDFCoverPages

Anand, A., & Bianco, B. (2021). The 2021 innovations dialogue conference report: Deepfakes, trust and international security. Retrieved April 12, 2024, from https://unidir.org/files/2021-12/UNIDIR_2021_Innovations_Dialogue.pdf

Baltazar, M. J., Saldanha Quadros, T., & Staal, J. (Eds.). (2021). *Image in the post-millenium: Mediation, process and critical tension*. Onomatopee.

Barbican. (2019). *AI: More than human*. Retrieved April 18, 2024, from https://www.barbican.org.uk/whats-on/2019/event/ai-more-than-human

Bazin, A. (1960). The ontology of the photographic image. *Film Quarterly*, *13*(4), 4–9. https://doi.org/10.2307/1210183

Beaumont, R. (2022). *LAION-5b: A new era of open large scale multi-modal datasets*. Retrieved April 6, 2024, from https://laion.ai/blog/laion-5b/

Benaich, N. (2023). *State of AI report 2023*. Retrieved April 6, 2024, from https://www.stateof.ai/

Benaich, N., & Hogarth, I. (2022). *State of AI report 2022*. Retrieved April 6, 2024, from https://www.stateof.ai/2022

Benjamin, W. (1972). A short history of photography. *Screen*, *13*(1), 5–26. https://doi.org/10.1093/screen/13.1.5

Born, G., & Barry, A. (2010). ART-SCIENCE: From public understanding to public experiment. *Journal of Cultural Economy*, *3*(1), 103–119. https://doi.org/10.1080/17530351003617610

Boucher, B. (2019). *Some people think cattelan's banana is genius. this law professor thinks it's illegal*. Retrieved March 15, 2024, from https://news.artnet.com/art-world/cattelan-banana-basel-illegal-1732932

Brittain, B. (2023). *Getty images lawsuit says stability AI misused photos to train AI*. Retrieved April 10, 2024, from https://www.reuters.com/legal/getty-images-lawsuit-says-stability-ai-misused-photos-train-ai-2023-02-06/

Bryan-Kinns, N., Ford, C., Chamberlain, A., Benford, S. D., Kennedy, H., Li, Z., Qiong, W., Xia, G. G., & Rezwana, J. (2023). Explainable ai for the arts: Xaixarts. *Proceedings of the 15th Conference on Creativity and Cognition*, 1–7. https://doi.org/10.1145/3591196.3593517

Burleigh, T. J., Schoenherr, J. R., & Lacroix, G. L. (2013). Does the uncanny valley exist? an empirical test of the relationship between eeriness and the human likeness of digitally created faces. *Computers in Human Behavior*, *29*(3), 759–771. https://doi.org/10.1016/j.chb.2012.11.021

C2PA. (2024a). *C2pa*. Retrieved April 11, 2024, from https://c2pa.org/

C2PA. (2024b). *Google to join c2pa to help increase transparency around digital contentC2pa in DALL·e 3*. Retrieved April 11, 2024, from https://c2pa.org/post/google_pr/

CAI. (2023). *Content credentials*. Retrieved April 11, 2024, from https://contentcredentials.org/

CAI. (2024). *New york times r&d team launches a pragmatic use case for CAI technology in journalism*. Retrieved April 14, 2024, from https://contentauthenticity.org/blog/new-york-times-rd-team-launches-a-pragmatic-use-case-for-cai-technology-in-journalism

Chayka, K. (2023). *The uncanny failures of a.i.-generated hands*. Retrieved April 12, 2024, from https://www.newyorker.com/culture/rabbit-holes/the-uncanny-failures-of-ai-generated-hands

Christie's. (2011). *Andreas gursky (b. 1955) rhein II*. Retrieved March 15, 2024, from https://www.christies.com/en/lot/lot-5496716

Christie's. (2022). *Edward steichen (1879-1973) the flatiron*. Retrieved March 15, 2024, from https://www.christies.com/en/lot/lot-6397095

Citron, D. K., & Chesney, R. (2019). Deep fakes: A looming challenge for privacy, democracy and national security. *107 California Law Review 1753*. Retrieved April 10, 2024, from https://scholarship.law.bu.edu/faculty_scholarship/640

Clark, S. E., Magrane, E., Baumgartner, T., Bennett, S. E. K., Bogan, M., Edwards, T., Dimmitt, M. A., Green, H., Hedgcock, C., Johnson, B. M., Johnson, M. R., Velo, K., & Wilder, B. T. (2020). 6&6: A transdisciplinary approach to art–science collaboration. *BioScience*, *70*(9), 821–829. https://doi.org/10.1093/biosci/biaa076

Cohen, J., & Meskin, A. (2004). On the epistemic value of photographs. *The Journal of Aesthetics and Art Criticism*, *62*(2), 197–210. Retrieved March 15, 2024, from https://www.jstor.org/stable/1559203

Commission, E. (2022). *2022 strengthened code of practice on disinformation*. Retrieved April 7, 2024, from https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation

Conway Morris, R. (2016). How photography cast new light on art. *The New York Times*. Retrieved November 9, 2023, from https://www.nytimes.com/2016/06/16/arts/international/how-photography-cast-new-light-on-art.html

Crimp, D. (1979). Pictures. *October*, *8*, 75. https://doi.org/10.2307/778227

Croitoru, F.-A., Hondru, V., Ionescu, R. T., & Shah, M. (2023). Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(9), 10850–10869. https://doi.org/10.1109/TPAMI.2023.3261988

Dahlin, E. (2021). Mind the gap! on the future of AI research. *Humanities and Social Sciences Communications*, *8*(1), 71. https://doi.org/10.1057/s41599-021-00750-9

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, medethics–2020–106820. https://doi.org/10.1136/medethics-2020-106820

Edwards, B. (2024). *Playboy image from 1972 gets ban from IEEE computer journals*. Retrieved April 7, 2024, from https://arstechnica.com/information-technology/2024/03/playboy-image-from-1972-gets-ban-from-ieee-computer-journals/

Enking, M. (2022). *Is popular a.i. photo app lensa stealing from artists?* Retrieved April 14, 2024, from https://www.smithsonianmag.com/smart-news/is-popular-photo-app-lensas-ai-stealing-from-artists-180981281/

Fischhoff, B., & Scheufele, D. A. (2019). The science of science communication III. *Proceedings of the National Academy of Sciences*, *116*(16), 7632–7633. https://doi.org/10.1073/pnas.1902256116

Flusser, V. (1983). *Towards a philosophy of photography*. Reaktion Books. Retrieved February 27, 2024, from http://imagineallthepeople.info/Flusser_TowardsAPhilosophyofPhotography.pdf

Flusser, V. (2011). *Into the universe of technical images*. University of Minnesota Press.

for Media Studies, P. I. (2022). *A global study on information literacy understanding generational behaviors and concerns around false and misleading information online*. Retrieved April 6, 2024, from https://www.poynter.org/wp-content/uploads/2022/08/A-Global-Study-on-Information-Literacy-1.pdf

Frye, B. L. (2021). SEC no-action letter request. *Creighton Law Review*, *54*(4). Retrieved March 15, 2024, from Frye,%20Brian%20L.%20%22SEC%20No-Action%20Letter%20Request.%22%20Creighton%20Law%20Review,%20vol.%2054,%20no.%204,%202021,%20pp.%20537-558.%20HeinOnline,%20https://heinonline.org/HOL/P?h=hein.journals/creigh54&i=583.

Genus, A., & Stirling, A. (2018). Collingridge and the dilemma of control: Towards responsible and accountable innovation. *Research Policy*, *47*(1), 61–69. https://doi.org/10.1016/j.respol.2017.09.012

Gillard, C. (2024). *The deeper problem with google's racially diverse nazis*. Retrieved April 6, 2024, from https://www.theatlantic.com/technology/archive/2024/02/google-gemini-diverse-nazis/677575/

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 80–89. https://doi.org/10.1109/DSAA.2018.00018

Goldman, A. I. (1979). What is justified belief? In G. S. Pappas (Ed.), *Justification and knowledge* (pp. 1–23). Springer Netherlands. https://doi.org/10.1007/978-94-009-9493-5_1

Grba, D. (2022). Deep else: A critical framework for AI art. *Digital*, *2*(1), 1–32. https://doi.org/10.3390/digital2010001

Habgood-Coote, J. (2023). Deepfakes and the epistemic apocalypse. *Synthese*, *201*(3), 103. https://doi.org/10.1007/s11229-023-04097-3

Hawkins, H., & Kanngieser, A. (2017). Artful climate change communication: Overcoming abstractions, insensibilities, and distances. *WIREs Climate Change*, *8*(5), e472. https://doi.org/10.1002/wcc.472

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. https://doi.org/10.48550/ARXIV.2006.11239

Honnef, K. (1990). *Andy warhol, 1928-1987: Commerce into art*. Taschen.

Howarth, S. (2000). *Marcel duchamp, fountain, 1917, replica 1964* [Tate]. Retrieved March 15, 2024, from https://www.tate.org.uk/art/artworks/duchamp-fountain-t07573

HuggingFace. (2024). *The AI community building the future.* Retrieved April 14, 2024, from https://huggingface.co/

Hurlbut, J. B. (2020). Imperatives of governance: Human genome editing and the problem of progress. *Perspectives in Biology and Medicine*, *63*(1), 177–194. https://doi.org/10.1353/pbm.2020.0013

Ichikawa, J. J., & Steup, M. (2018). The analysis of knowledge. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Summer 2018). Metaphysics Research Lab, Stanford University. Retrieved April 12, 2024, from https://plato.stanford.edu/archives/sum2018/entries/knowledge-analysis/

Inie, N. (2024). What motivates people to trust 'AI' systems? https://doi.org/10.48550/ARXIV.2403.05957

Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *iScience*, *24*(11), 103364. https://doi.org/10.1016/j.isci.2021.103364

Law, E. U. (2022). *Regulation (EU) 2022/2065 of the european parliament and of the council of 19 october 2022 on a single market for digital services and amending directive 2000/31/EC (digital services act) (text with EEA relevance)*. Retrieved April 7, 2024, from http://data.europa.eu/eli/reg/2022/2065/oj

Lopes, D. M. (2016). *Four arts of photography: An essay in philosophy* (1st ed.). Wiley. https://doi.org/10.1002/9781119053194

Lowenstein, S. (2024). *Using secure sourcing to combat misinformation*. Retrieved April 14, 2024, from https://rd.nytimes.com/projects/using-secure-sourcing-to-combat-misinformation/

Magnus, P. D. (2023). Generative AI and photographic transparency. *AI & SOCIETY*. https://doi.org/10.1007/s00146-023-01817-8

Mandel, B. R. (2009). Art as an investment and conspicuous consumption good. *American Economic Review*, *99*(4), 1653–1663. https://doi.org/10.1257/aer.99.4.1653

Mathys, M., Willi, M., Graber, M., & Meier, R. (2024). Synthetic image generation in cyber influence operations: An emergent threat? https://doi.org/10.48550/ARXIV.2403.12207

McDermott, A. (2019). Artists and scientists come together to explore the meaning of natural sound. *Proceedings of the National Academy of Sciences*, *116*(26), 12580–12583. https://doi.org/10.1073/pnas.1908588116

Merchant, B. (2023). *Blood in the machine: The origins of the rebellion against big tech* (First edition). Little, Brown; Company.

Meske, C., & Bunde, E. (2020). Transparency and trust in human-AI-interaction: The role of model-agnostic explanations in computer vision-based decision support. In H. Degen & L. Reinerman-Jones (Eds.),

Artificial intelligence in HCI (pp. 54–69, Vol. 12217). Springer International Publishing. https://doi.org/10.1007/978-3-030-50334-5_4

Miller, B., & Record, I. (2013). JUSTIFIED BELIEF IN a DIGITAL AGE: ON THE EPISTEMIC IMPLICATIONS OF SECRET INTERNET TECHNOLOGIES. *Episteme*, *10*(2), 117–134. https://doi.org/10.1017/epi.2013.11

Minow, M., & Minow, N. (2023). *Social media companies should pursue serious self-supervision — soon: Response to professors douek and kadri* [Harvard law review]. Retrieved April 7, 2024, from https://harvardlawreview.org/forum/vol-136/social-media-companies-should-pursue-serious-self-supervision-soon-response-to-professors-douek-and-kadri/

MIT. (2022). *Turning MIT inside-out*. Retrieved April 18, 2024, from https://www.technologyreview.com/2022/12/19/1064060/seen-on-campus-jf-2023-tbd/

Mizrahi, V. (2021). Seeing through photographs: Photography as a transparent visual medium. *The Journal of Aesthetics and Art Criticism*, *79*(1), 52–63. https://doi.org/10.1093/jaac/kpaa009

Mnookin. (1998). The image of truth: Photographic evidence and the power of analogy. *Yale Journal of Law & the Humanities*. Retrieved April 9, 2024, from https://ssrn.com/abstract=135311

Mori, M., MacDorman, K., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, *19*(2), 98–100. https://doi.org/10.1109/MRA.2012.2192811

Morozov, E. (2013). *To save everything, click here: The folly of technological solutionism*. PublicAffairs.

Mulvin, D. (2021). *Proxies: The cultural work of standing in*. The MIT Press. https://doi.org/10.7551/mitpress/11765.001.0001

Nanotech, N. (2018). A note on the lena image. *13*(12), 1087–1087. https://doi.org/10.1038/s41565-018-0337-2

Nicoletti, L., & Bass, D. (2023). *Humans are biased. generative AI is even worse.* Retrieved April 6, 2024, from https://www.bloomberg.com/graphics/2023-generative-ai-bias/

Niépce, J. N. (1827). *Untitled 'point de vue'*. Retrieved April 9, 2024, from https://www.hrc.utexas.edu/niepce-heliograph/

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York university press.

Getty Images vs Stability AI. Retrieved April 10, 2024, from https://tmsnrt.rs/3Y5yjI7

United States v. Mackey. Retrieved April 7, 2024, from https://www.justice.gov/usao-edny/pr/social-media-influencer-douglass-mackey-sentenced-after-conviction-election

of Homeland Security, D. (2021). *Increasing threat of deepfake identities*. Retrieved April 6, 2024, from https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf

of Homeland Security, D. (2022). *Increasing threats of deepfake identities – phase two*. Retrieved April 6, 2024, from https://www.dhs.gov/sites/default/files/2022-10/AEP%20DeepFake%20PHASE2%20FINAL%20corrected20221006.pdf

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy* (First edition). Crown.

Ong, J. Y. (2023). *AI's 'accidental' ethnically ambiguous homer simpson*. Retrieved April 14, 2024, from https://thechainsaw.com/nft/ai-accidental-ethnically-ambiguous-homer-simpson/

Online Safety Act 2023 (c. 50) (2023). Retrieved April 7, 2024, from https://www.legislation.gov.uk/ukpga/2023/50/enacted

OpenAI. (2023). DALL·e 3 system card. Retrieved November 9, 2023, from https://cdn.openai.com/papers/DALL_E_3_System_Card.pdf

OpenAI. (2024a). *C2pa in DALL·e 3*. Retrieved April 11, 2024, from https://help.openai.com/en/articles/8912793-c2pa-in-dall-e-3

OpenAI. (2024b). *Sora* [Creating video from text]. Retrieved March 29, 2024, from https://openai.com/sora#:~:text=00%20/%200%3A00-,Prompt%3A%20Reflections%20in%20the%20window%20of%20a%20train%20traveling%20through%20the%20Tokyo%20suburbs.,-Prompt%3A%20Reflections%20in

Peebles, W., & Xie, S. (2022). Scalable diffusion models with transformers. https://doi.org/10.48550/ARXIV.2212.09748

Perry, G. (2016). *Playing to the gallery: Helping contemporary art in its struggle to be understood*. Penguin Books.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. https://doi.org/10.48550/ARXIV.2103.00020

Raghavan, P. (2024). *Gemini image generation got it wrong. we'll do better.* Retrieved April 6, 2024, from https://blog.google/products/gemini/gemini-image-generation-issue/

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents. https://doi.org/10.48550/ARXIV.2204.06125

Review, H. L. (2024a). United states v. mackey. Retrieved April 7, 2024, from https://harvardlawreview.org/print/vol-137/united-states-v-mackey/

Review, H. L. (2024b). Voluntary commitments from leading artificial intelligence companies on july 21, 2023. Retrieved April 7, 2024, from https://harvardlawreview.org/print/vol-137/voluntary-commitments-from-leading-artificial-intelligence-companies-on-july-21-2023/

Rini, R. (2020). Deepfakes and the epistemic backstop. *Philosophers' Imprint*, *20*(24), 1–16. Retrieved 2024, from http://hdl.handle.net/2027/spo.3521354.0020.024

Ritchin, F. (2009). *After photography*. W. W. Norton.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2021). High-resolution image synthesis with latent diffusion models. https://doi.org/10.48550/ARXIV.2112.10752

Sahu, A. K., Umachandran, K., Biradar, V. D., Comfort, O., Sri Vigna Hema, V., Odimegwu, F., & Saifullah M. A. (2023). A study on content tampering in multimedia watermarking. *SN Computer Science*, *4*(3), 222. https://doi.org/10.1007/s42979-022-01657-1

Sarewitz, D., & Nelson, R. (2008). Three rules for technological fixes. *Nature*, *456*(7224), 871–872. https://doi.org/10.1038/456871a

Scheufele, D. A., & Krause, N. M. (2019). Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, *116*(16), 7662–7669. https://doi.org/10.1073/pnas.1805871115

Schoonover, N. (2023). *Pro-trump troll sentenced to seven months over election day shenanigans*. Retrieved April 14, 2024, from https://www.courthousenews.com/pro-trump-troll-sentenced-to-seven-months-over-election-day-shenanigans/

Shloss, C. (1980). The privilidge of perception. *The Virginia Quarterly Review*, *56*(4), 596–611. Retrieved 2024, from http://www.jstor.org/stable/26436109

Slaughter, K. (2024). A peripheral vision: Framing the cultural bias in the center of photography. *Critical Inquiry*, *50*(2), 317–334. https://doi.org/10.1086/727644

Smith, K., Todd, M. J., & Waldman, J. (2009). *Doing your undergraduate social science dissertation*. Routledge.

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. https://doi.org/10.48550/ARXIV.1503.03585

StabilityAI. (2022). *Stable diffusion* [GitHub]. Retrieved April 11, 2024, from https://github.com/CompVis/stable-diffusion

StabilityAI. (2024). *Stable diffusion 3*. Retrieved April 11, 2024, from https://stability.ai/news/stable-diffusion-3

Suleyman, M., & Bhaskar, M. (2023). *The coming wave: AI, power and the twenty-first century's greatest dilemma*. The Bodley Head.

Sun, H. (2023). Regulating algorithmic disinformation. *The Columbia Journal of Law & the Arts*, *46*(4). https://doi.org/10.52214/jla.v46i3.11237

The White House. (2023). *FACT SHEET: Biden-harris administration secures voluntary commitments from leading artificial intelligence companies to manage the risks posed by AI*. Retrieved 2024, from https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/

Van De Poel, I. (2016). An ethical framework for evaluating experimental technology. *Science and Engineering Ethics*, *22*(3), 667–686. https://doi.org/10.1007/s11948-015-9724-3

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. https://doi.org/10.48550/ARXIV.1706.03762

Wade, N. J., & Finger, S. (2001). The eye as an optical instrument: From camera obscura to helmholtz's perspective. *Perception*, *30*(10), 1157–1177. https://doi.org/10.1068/p3210

Walton, K. L. (1984). Transparent pictures: On the nature of photographic realism. *Critical Inquiry*, *11*(2), 246–277. https://doi.org/10.1086/448287

Wang, K., Xu, Z., Zhou, Y., Zang, Z., Darrell, T., Liu, Z., & You, Y. (2024). Neural network diffusion. https://doi.org/10.48550/ARXIV.2402.13144

Weili, S. (2022). *AI: Mind the gap*. Retrieved April 18, 2024, from https://shi-weili.com/

Wildberger, A. (1994). Alleviating the opacity of neural networks. *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, *4*, 2373–2376. https://doi.org/10.1109/ICNN.1994.374590

Wollheim, R. (1998). On pictorial representation. *The Journal of Aesthetics and Art Criticism*, *56*(3), 217. https://doi.org/10.2307/432361

Yang, J. (2024). Ultra-high-resolution image synthesis with pyramid diffusion model. https://doi.org/10.48550/ARXIV.2403.12915

Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. https://doi.org/10.48550/ARXIV.2302.05543

Zhang, Q.-s., & Zhu, S.-c. (2018). Visual interpretability for deep learning: A survey. *Frontiers of Information Technology & Electronic Engineering*, *19*(1), 27–39. https://doi.org/10.1631/FITEE.1700808

Zhou, M., Abhishek, V., Derdenger, T., Kim, J., & Srinivasan, K. (2024). Bias in generative AI. https://doi.org/10.48550/ARXIV.2403.02726