WOR(L)D-IMAGE *TRANS*-FORMATION: LOOKING THROUGH DALL-E 2 AND MIDJOURNEY

Natalia Stanusch

The relationship between words and images, between rational and sensible, has been a persistent query of representational theory in philosophy, semiotics, and art history. In his *Short History of Photography*[1], Walter Benjamin pondered the relationship between image and caption, wondering whether the caption would become the most meaningful element of an image. But what if the caption becomes the image? The relationship between caption and image has been opened to new exploration by recent Artificial Intelligence (AI) text-to-image generation models, such as DALL-E 2 and Midjourney. Models such as DALL-E 2 and Midjourney generate images when the user submits a written prompt, as little as a single word or a phrase. Might it be that text-to-image generation models will enable us to rethink the centuries-old relation between words and images that situates technology and art in the very center of this seemingly binary phenomenological clash?

Yuk Hui's concept of cosmotechnics is useful here, and so is too Joanna Zylinska's post-humanist paradigm. Cosmotechnics is meant to address the question of technology of today by positioning technological activates within relational frameworks (cosmoses and moralities) which are conditioned by localities of different dynamics[2]. Joanna Zylinska's post-humanist paradigm asks us to discard the distinction between human made and non-human made objects. She notes that human art and non-human made objects – an «assembly with a plethora of nonhuman agents»[3] – constantly influence each other, and are therefore profoundly and consequentially connected.

Hui and Zylinska both return us to the ancient concept of technē – a word that includes technology and art without acknowledging a significant separation between them. In the understanding of the ancient Greek technē, art and technology were not separated, but rather inseparable. Greek technē did not distinguish between artistic and other forms of making, between engineering and so-called creative production. Instead, technē referred to the act of bringing something forward. Hui suggests that while the understanding of technology in twenty-first century is radically different from that of the Greeks, contemporary discourse of art has to account for «the question of technology»[4]. Zylinska notes that because Greeks saw no distinction between art and technology within technē, we are invited to expand our

1 W. Benjamin. *A Short History of Photography*, Monogram, London 1972 (1931), p. 25.
2 Y. Hui, *Cosmotechnics as Cosmopolitics*, "e-flux", 86, 2017.
3 J. Zylinska, *Ai Art: Machine Visions and Warped Dreams*, Open Humanities Press, London 2020, p. 54.
4 Y. Hui, *Art and Cosmotechnics*, "e-Flux", New York, 2021, p. 77.

perception of the technical quality of «technical assemblages» such as algorithms and databases, and the act of creation[5]- AI technologies, including the recent text-to-image generation models, allows us to talk about artistic and technological making within a single discourse.

What differs models such as DALL-E 2 from previous AI-based image generating models (e.g. GANs - generative advisory network – "artworks" such as *Portrait of Edmond de Belamy* sold at Christie's in 2018), is the scale, accessibility, and internal logic of the model. Text-to-image generation models discussed in this paper were "trained" using machine leaning techniques applied in AI language generation models, and share with them a similar basis for image generation. For example, DALL-E 2 was trained using a language prediction model GPT-2, a model used to predict the next word in a sentence based on a database of human texts[6].

The text-to-image generators, such as DALL-E 2, are large transformer models based on «next pixel prediction»[7]. The prediction logic relies upon reversing a so-called "diffusion" process. Diffusion is a process in which an image is gradually turned into a pixelated noise, until it reaches a stage of illegibility. Image-generation models start with "an image of noise," a pixelated chaos, which is gradually transformed, pixel by pixel, into a desired image. In order to generate a picture, these models base their "predictions" upon being fed great quantities images. Thus, text-to-image generation models have on their disposal a chaotic emulation of the Internet, a twenty-first century offspring of Warburg's *Bilderatlas Mnemosyne*. As appealingly phrased by Eryk Salvaggio, «DALL-E 2 doesn't make pictures that have "never existed," it makes every picture that has ever existed»[8].

Text-to-image generation models gained sudden visibility in the summer of 2022, following their major opening for the public use. Some became freely available to the public (Stable Diffusion), some offered a limited time of a free Beta version (Midjourney) or introduced a subscription model (DALL-E 2 and Midjourney). The growing public interest in these developments can be illustrated in the foundation of businesses such as promptbase.com, a company which specializes in selling written 'prompts' which retrieve best results for a given model. The prompt writing is not a clear-cut action, as using a word «very» three times in a row can give better results than a single «very»[9] (as will be discussed later on). Upon the opening up of the access to these models, "#AIArt" images begun to flood the feeds of Twitter, Instagram, and TikTok.

5 J. Zylinska, *Ai Art*, cit., p. 32.

6 Openai, *Image GPT*,"OpenAI",17 June, 2020.

7 *Ibid.*

8 E. Salvaggio, *Ghosts of diffusion. How the new generation of image-producing AIs tries to reverse entropy,* "Cybernetic Forests", 21 August 2022; https://cyberneticforests.substack.com/p/ghosts-of-diffusion.

9 K. Wiggers, *A startup is charging $1.99 for strings of text to feed to dall-e 2*, "TechCrunch", 29 July 2022, https://tcrn.ch/3Q5f8cR.

The speed and abundance with which "#AIArt" is shared mirrors the sense of being overwhelmed by contemporary image production. The language used in addressing this image production has been compared to that of natural disasters, creating an «impression of an unmanageable and unstoppable cascade of images»[10]. The growing number of AI-generated images, conditioned by models such as Midjourney and DALL-E 2, can be argued to contribute to the «mass image» as theorized by Cubitt[11]. Cubitt claims that the mass image «excludes the world it depicts […]. [It] inscribes the world as data»[12]. Similarly, Hito Steyerl defines the spam-like online circulation and production of «poor images» as «flattening-out of visual content—the concept-in-becoming of the images—positions them within a general informational turn, within economies of knowledge that tear images and their captions out of context into the swirl of permanent capitalist deterritorialization»[13].

Text-to-image generation models are based on questionable practices which had already led to criticism of "AI art," with the critiques often rightfully circling around the unresolved grey area of copyright. These models are deeply disturbing in their incorporation of intrinsic biases inscribed in large datasets on which they were trained[14]. The output of AI image generation models is conditioned by its dataset[15]. These datasets consist of images scrapped from the internet, coming from databases such as ImageNet and LAION-400M. Aside from their potential misuses, as already exemplified with deepfake technology, text-to-image generation models are an extension of a long-time quest of using computing powers to "see," a quest which is «inherently social and political»[16]. This quest of machine vision assumes that AI and big data can construct a new comprehensive understanding (or a complete image of) the world[17]. Yet this «illusion of objectivity» dissolves the moment machine vision reveals its «unconscious: that of the social structures engrained on the training dataset[18]. Hence, from the visual and epistemological points of view, one of the most significant ambiguities is what (or who) these text-to-image generation models are making (in)visible.

10 T. Dvořák, J. Parikka, *Introduction*, in Id. (eds.), *Photography Off the Scale. Technologies and Theories of the Mass Image*, Edinburgh University Press, Edinburgh 2021, p. 4.

11 S. Cubitt, *Mass Image, Anthropocene Image, Image Commons*, in T. Dvořák, J. Parikka (eds.), *Photography Off the Scale*, cit., pp. 25-40.

12 Ivi, p. 26.

13 H. Steyerl, *The Wretched of the Screen*, Sternberg Press, Berlin 2012 (B), p. 41.

14 For comprehensive studies of biases and harmful content inscribed in large-scale datasets of images see B. Abeba, V. U. Prabhu, E. Kahembwe, *Multimodal datasets: misogyny, pornography, and malignant stereotypes,* "arXiv preprint", arXiv:2110.01963, 2021.

15 C. Bueno, M. J. Schultz Abarca, *Memo Akten's Learning to See: from machine vision to the machinic unconscious*, "AI & Society", 36, 2021, p. 1183.

16 K. Crawford, T. Paglen, *Excavating AI: the politics of images in machine learning training sets*, "AI & Society", 36, 2021, p. 1106.

17 C. Anderson, *The end of theory: the data deluge makes the scientific method obsolete*, "Wired," 2008, https://www.wired.com/2008/06/pb-theory/.

18 C. Bueno, M. J. Schultz Abarca, *Memo Akten's Learning to See*, cit., p. 1184.

From the visual standpoint, the algorithmic logic constructing "AI art" has been criticized for following an already established visual dictionary with slight variations under the guise of "creativity." Zylinska characterizes most of AI-generated art, particularly the kind derived from datasets, as «crowd-sourced beauty»[19]. She claims that the results are dull and substanceless images which are «dazzling viewers with the mathematical sublime of big datasets, rapid image flows and an intermittent flicker of light»[20]. Thus, Zylinska argues that "AI art" often has «a pacifying effect, anaesthetising us into the perception of banal sameness that creates an illusion of diversification»[21]. Such anaesthetizing effect in text-to-image generation models could lie in reproducing the past through historical data. It has been noted that «the use of predictive models based on historical data is inherently conservative (…). [It] tends to reproduce and reinforce assessments and decisions made in the past»[22], leading to a creation of «a future based on the past»[23]. In AI models, the processes of reproducing (or reconfiguring) the past are grounded in interpretation of (generated) data. It is worth recalling that, as Lisa Gitelman states, «raw data is an oxymoron»[24]. Data does not exist "out there"; it has to be manufactured.

Midjourney generating an image "in style of Beksiński" or DALL-E creating "a Faith Ringgold painting" deploys a translation of a style into a pattern, stripping it from nuances of cultural and anthropological contexts in which given (sets of) artifacts were created by human actors. These embodied contexts of lived experiences are unlikely to be comprehensively translated (or rather created) in a form of data; «read as data, "jazz" and "Monk" are mere patterns, not authored, musical/political acts. As such, we have to consider how algorithmic classifications profoundly disrupt what we refer to when we talk about race and legacies of blackness, as well as new forms of whitewashing and cultural appropriation»[25]. «Algorithmic classifications» that Cheney-Lippold amply discusses, respond to a broader project of rationalization through datafication. Referring to another AI-art project, the Google DeepDream, Hito Steyerl designates AI-generated images as products of «the networked operations of computational image creation, certain presets of machinic vision, its hardwired ideologies and preferences»[26]. Ideologies embedded in the networks that condition AI vision share the assumption that there exist «universal correspondences» between ideas and pictures, and that these correspondences reflect «uncomplicated, self-evident, and measurable ties between images, referents, and labels»[27]. This project of «universal labeling» or «transcoding» is

19 J. Zylinska, *Ai Art*, cit., p. 49.

20 Ivi p.72

21 Ivi, p. 83

22 O. H. Gandy Jr, *Exploring Identity and Identification in Cyberspace*, "*Notre Dame*", 14, J.L. Ethics & Pub. Pol'y 2000, p. 1101.

23 W. Hui Kyong Chun, *Programmed Visions: Software and Memory*, MIT Press, Cambridge 2011, p. 9.

24 L. Gitelman, "*Raw Data" Is an Oxymoron*, MIT Press, Cambridge 2013.

25 J. Cheney-Lippold. *We Are Data: Algorithms and the Making of our Digital Selves*, New York University Press, New York 2019, 72.

26 H. Steyerl, *A Sea of Data: Apophenia and Pattern (Mis-)Recognition*, "e-Flux", 72, 2016.

27 K. Crawford, T. Paglen, *Excavating AI: the politics of images in machine learning training sets*, "AI & Society", 36, 2021, p. 1113.

placed in an ambiguous relation to human sensibility, and becomes prominently illustrated in text-to-image generation models. Michelle Henning makes a fitting point while referring to social media photography and emojis, stating that «to make pictures language-like also seems to be to make them calculable, quantifiable - it is a rationalization of both pictures and human feeling»[28].

Text-to-image generation models can be said to embody the affordance to situate human imagination within a rationalization process of language>computation>image. A user generating an image expresses (translates) an image they wish to see into a spelled-out text prompt (first rationalization). Next, an AI actor – the model – uses data (second rationalization) to translate, or perhaps transcode, an image (third rationalization). While trying to avoid the trap of techno-determinism, it is worth questioning what *may* happen to human sensibility if, as Yuk Hui incisively notes, «every painting is already finished before it is painted, because the canvas is already calculated as a finite set of possibilities»[29]. Is it (yet another) death of art?

Unseeing technē

One way of approaching the differences between technē and AI models is to focus on the notion of human and nonhuman vision, visibility, and invisibly. Text-to-image generation models share a peculiar relation with the notion of "visuality" and "seeing." The way in which AI generates images does not correspond to traditional art historical conceptualizations of image making, but to a new algorithmic or computational regime of "seeing." Jussi Parikka calls this machine vision «seeing as measurement»[30]. Similarly to other AI 'vision' and pattern recognition technologies, text-to-image generation models follow a logic of nonhuman vision, while the processes and outputs that these models construct remain largely invisible to the human eye.

Text-to-image generation models bring about an automated dissemination and simplification of image making, the two qualities that invite a comparison with photography. In analog photography, being invisible was most often about being outside of the lens; however, in image generation models such as DALL-E 2, what is hidden can just as well be buried inside the image and not outside of it: inside the dataset, or within the algorithmic prediction of pixel sequences transcoded from the text prompt. In *Monty Python Flying Circus'* sketch "How Not to Be Seen", being invisible depends on the physical capacity to hide from the camera. In 2022, invisibility is a much more complicated and tech-dependent quality (what is present in dataset, what is learnt by an algorithm) rather than a bodily one[31]. While in analog pho-

28 M. Henning, *Feeling Photos: Photography, Picture Language and Mood Capture*, in T. Dvořák, J. Parikka (eds.), *Photography Off the Scale*, cit., p. 80.

29 Y. Hui, *Art and Cosmotechnics*, cit., p. 215.

30 J. Parikka, *On Seeing Where There's Nothing to See*, cit., p. 186.

31 See Hito Steyerl's film *How Not to be Seen. A Fucking Didactic Educational .MOV File* (2013).

tography people hide from the camera, in AI-generated pixilation, the infinite density allows one to hide (or be hidden) within the image.

Today, one could argue, it is inside the image where objects and people can be made visible or invisible. And just like in *Monty Python*, explosions are included. These images often make visible what the human eye cannot see; they are produced, reformatted, and viewed using specialist image recognition systems. In Steyerl's article on the relation of digital image to politics and power, she gives an example of the prominent role of non-human technological actors in human embodied experiences. Steyerl suggests, «just look at the NSA [National Security Agency] training manual for unscrambling hacked drone intercepts. (…) you need to bewitch the files with a magic wand (Image Magick is a free image converter)»[32].

Steyerl's comparison of technology to a magic wand is a particularly seducing one. Magic tricks must have a magician and a spectator, therefore defining roles of the maker and the viewer. The medium of AI image generation – if we can categorize text-to-image models such as Midjourney and DALL-E 2 as a medium – reflects a profound shift in contemporary visuality which can be thought of using the magician analogy. Most of those who consume these images have no knowledge of what actually happens beyond the pixels and lines of code, and, more importantly, why does it happen. The grey zone is occupied by actors who participate in image 'postproduction:' prompt creators, followers, influencers, scholars, artists, trolls. Text-to-image generation models are also actors within this new structure of influence. As Zylinska points out, AI image-generating technologies «are both objects to be looked at and vision-shaping technologies, for humans *and* machines»[33], and, as such, they bring as much new circulation of visibility, as they bring invisibility.

One more look at Midjourney and DALL-E 2

Midjourney and DALL-E 2 generate sets of four images per single prompt. While generating prompts, we can see that these models provide unique results despite using the same prompt. For example, using as a prompt the single word "serendipity" in two separate queries provides two different sets of results (fig.1 and fig.2 display DALL-E 2 results, fig. 3 and fig. 4 show the results from Midjourney). While these results are 'unique' for my query, we could be tempted to point out stylistic differences inherent in Midjourney and DALL-E 2, that one is biased towards silk-screen and painterly aesthetics while the other prompts a more representational and stock-photographic style. Yet that is not inherent in DALL-E 2 or Midjourney. Whatever comes out of these models is tangentially related to my intention as user or the aesthetic preference of the model. The very outcomes of these models, and the fact that these examples were distinctively stylized, is serendipitous.

32 H. Steyerl, *A Sea of Data*, cit.
33 J. Zylinska, *Ai Art*, cit., p. 106.

The way in which these text-to-image generation models depict abstract concepts, such as singular words or verses of poetry, is precisely an algorithmic serendipity. Midjourney or DALL-E 2 take from the datasets that they were trained on and propose images built through their "vision" of human "sensibility". These technologies are mashed into a complicated array of interrelations, especially once triggered by the users. Aside from relying on algorithmic serendipity, prompt-writing requires a re-adjustment of image describing and querying. Rather than constructing a visual analysis of a painting one would look at it in a museum, or searing for an image via Google images, text-to-image generation models require to think of the prompt as an equation. This new language of equations, specific to each model, is derived (or driven) from the datasets of ALT texts, pixels, keywords, and data points. The prompts and later "variations" of generated images serve as a way of steering the machine, with following users' aesthetics preferences or a pre-structured vision of desired composition.

On the epistemological level, the process of generating images implies the possibility of navigating through the logic of the model. Trying to generate a satisfactory image from a prompt is similar to stabbing the «black box» with a stick (the «black box» is a term coined by Frank Pasquale to refer to the hidden logics of the algorithms and data processing). Rather than seeing right through the black box, one is able to cast light in different directions, see the sides and angles inside of it. While we might feel tempted to assign a level of autonomy to these models, for example by claiming that Midjourney, when not directed, generates painterly compositions of organic web-like complexities which create illusions of sophistication, that would be analogous to casting the light in only one direction, missing all the other directions.

Back to Benjamin

To follow in the footsteps of Walter Benjamin, rather than asking whether these models, and the (in)visibilities which they produce, are art, one should ask how can these models transform art. Text-to-image generation models can be seen as both actors (tools) within a larger network of visuality and as a technology/art with a potentiality to suppress or amplify human sensibility of visuality in the twenty-first century. In his book *Art and Cosmotechnics*, Hui argues that a refusal to engage with technological developments such as "AI art" is a refusal to engage with a complex yet «intimate relation between art and technology»[34]. The relation between art and technology has largely dissolved from our consciousness as the understanding of technē became dismantled into, what seems today, an unreconcilable binary. As Hui further expands on this binary,

> Today, when we say "art and technology", we mean global art and we mean digital technology, and the "and" implies art using technology. But what does "use" mean here? Does it mean the instrumentalization or appropriation of technology by art, as we see in works using augmented

34 Y. Hui, *Art and Cosmotechnics*, cit., p. 219.

> reality and virtual reality appropriated from industrial products? Or does it, more precisely, in quasi-Heideggerian language, reopen the question of Being as the Here we must reiterate that questioning the Unknown or Being for Heidegger is an attempt to reopen the question of technology and *locality*: technology in the sense that art can also open the possibility of technology by providing the imagination of a technodiversity; locality in […] reopening the question of the unknown through technology affirms the irreducible difference of the multiplicity of modes of thinking (aesthetic, technical, moral, philosophical…) in different cultures and territories[35].

One way of approaching the question of «art and technology» is to use text-to-image generation models as critical tools against art and against technology. Hui points out that while art requires technological frameworks to operate, it has a potentiality to go beyond the technology which frames it. Art's capacity to go "outside" of dominant techno-scientific ways of thinking makes it capable of "revealing" what is invisible, or «making what is invisible sensible»[36]. Art has to speak to the viewer by bringing forth the sensibility which is based upon non-rational, what Hui speaks of as crucial for the relevant aesthetic experience today[37]. In order to resist the rationalization of techno-scientific thinking, art has to embrace the non-rational and even make «science become a stranger to itself»[38], argues Hui.

With regards to AI, Hui focuses on algorithmic «recursivity» which he explains as the algorithmic possibility of non-linearity[39]. He also suggests that today the need for expanding art is similar to the actions of avant-garde artists of early twenty century who «enlarged the medium of art»[40] by expanding with the uses of canvas and cinema rather than destroying or rejecting them. To reveal new forms of sensibilities, art has to situate itself within the modern discourse of power and technology. Hui brings attention to the fact that art has to go back to its continuous quality of questioning and challenging of technology and of itself[41].

To question technology-through-art and art-through-technology we can turn to what Joanna Zylinska calls a «posthumanist art history» She argues that the critical possibilities of art can be expanded once art is seen as a process of collaboration within networks of actors and triggers[42]. Art has to situate itself in a critical dialogue with technologies by which it operates, but it also has to the socio-political structures that condition these technologies and which these technologies are part of, and in itself, it has to be critical of limitation of art[43].

35 Ivi, p. 50.
36 Ivi, p. 28.
37 Ivi, p. 125.
38 Ivi, p. 116
39 Y. Hui, *Augmentation of the Senses (or the Machine Becomes an Idea that Makes Art)*, "The National Gallery of Victoria", 2022, https://www.ngv.vic.gov.au/augmentation-of-the-senses-or-the-machine-becomes-an-idea-that-makes-art/.
40 *Ibid.*
41 Y. Hui, *Art and Cosmotechnics*, cit., p. 276.
42 J. Zylinska, *Ai Art*, cit., p. 54.
43 Ivi, p.14.

Following this understanding of art, Zylinska further states that algorithmic art can be used as an expansion of visuality by generating «new visions and vistas for the world to come»[44].

This collaboration or co-creation using the recent developments such as text-to-image generation models can introduce what Zylinska calls a nonhuman vision, or «the possibility of seeing otherwise – for both the machine and the human»[45]. She further adds that "AI art" could be imagined as a bridge between a work of art and the problem of human vision, going beyond human capacity to see and, perhaps even, imagine[46]. Thus, there is a possibility for a new critical sensibility of the "outside" or "non-human" which can be explored through technologies such as text-to-image generation models.

Conclusion

Contemporary culture is a visual culture[47]. The arts, computer technology, and pop culture have all pointed in the direction of further practices of visual documentation and reproduction, such that the rise of AI generated images was predictable, if not absolutely inevitable in the grand march toward a fully documented, rationalized, and visually datafied world. One notes that Dall-e and Midjourney apparently have different methods of processing language, such that the user of these programs soon learns different strategies of language input in order to manipulate the programs to a desired end. We are still only at the beginning of exploring the many ways in which one AI systems differs from others, and, in short, we can't essentialize what AI does, because we are still learning what AI is capable of.

What is at stake here is whether sentience, human aesthetic experience, can be modified and distinguished from purely digitalized pattern recognition. Most people would like to think that human aesthetic experience, sentience, is always beyond what computers can do, but posthuman narrative will tell us that this is merely a sentimental, humanist illusion.

The models such as DALL-E 2 point towards the artificiality of objective and unfiltered, human and non-human, vision[48], "seeing" allow for a further questioning of «what this seeing means in the broader sense of processing patterns, organizing culture, or designing large-scale distributed nonhuman agent systems»[49]. Hence, while acknowledging that the text-to-image generation models such as Midjourney and DALL-E 2 can provide a safe detachment from the real by offering a digital spectacle, these models can participate in an expansion of our visual field of possibilities. The ambiguities of these technologies should not be overlooked but rather revealed in the critical spotlight.

44 Ivi, p.45.

45 Ivi, p. 150.

46 Ivi, p. 142.

47 N. Mirzoeff, *How to See the World*, Penguin Books, London 2015; A. L. Boylan, *Visual Culture*, MIT Press, Cambridge 2020.

48 C. Bueno, M. J. Schultz Abarca, *Memo Akten's Learning to See*, cit., p. 1185.

49 J. Parikka, *On Seeing Where There's Nothing to See*, cit., p. 189.

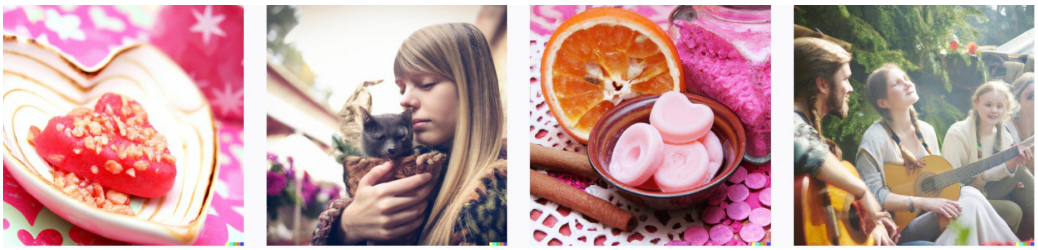

Fig. 1

Fig. 2

Fig. 3

Fig. 4