# Diffused Seeing: The Epistemological Challenge of Generative AI

Joanna Zylinska

*Media Theory*
Vol. 8 | No. 1 | 229-258

http://mediatheoryjournal.org/

# Diffused Seeing: The Epistemological Challenge of Generative AI

JOANNA ZYLINSKA

*King's College London, UK*

## Abstract

This article examines the transformation of the relationship between seeing and understanding in humans *and* machines by the technologies of machine learning known as 'generative AI'. Taking Stable Diffusion as the main case study, while also looking at its competitors (DALL·E 2 and Midjourney), it starts by analysing the photographic infrastructure underpinning these generative models. The subsequent examination of 'diffusion' as a key concept that underpins the text-to-image generation process leads to some broader questions about the ongoing instability and dissolution of our current epistemological and political frameworks. Taking seriously the charge issued by some critics equating developments in generative AI with nihilism or even fascism, the article considers whether the current socio-technical moment can also offer some emancipatory possibilities. Images are used as part of the article not just by way of illustration but also to enact some of its argument.

## Keywords

entropy, Flusser, generative AI, human, nonhuman, perception, Stable Diffusion, understanding

## Epistemic framing

This article investigates the transformation of the relationship between seeing and understanding in humans *and* machines by the technologies of machine learning known as 'generative AI'. It starts from the assumption that modern perception, with the diverse modes of seeing it entails[1] – be it in humans or machines – is inherently

photographic, even though it is not solely visual (Zylinska, 2023). Since the invention of the photographic medium in the early nineteenth century, perception has been expanded at different stages of photography's development: from analogue to digital to networked to generative. Over the last six decades the array of computing machines, transmission networks and haptic sensors has gradually supplemented the visual apparatus of perception to encompass other sensory experiences and data. This article focuses on the most recent of those stages of photographic development: the creation of photorealistic images by machine learning-driven models, aka 'generative AI'. Taking Stability AI's Stable Diffusion as a case study, it zooms in on the photographic base of these generative models, i.e., the fact that their training data consists largely of photographs, while the produced outputs are aimed at photographic verisimilitude with its source material, albeit remediated in a variety of styles that often exceed straightforward realism.

The argument will proceed through an analysis of the philosophical and technical aspects of a concept that underpins the generation process in many current text-to-image models (not only Stable Diffusion but also DALL·E 2, Midjourney, Imagen and Firefly): diffusion. It is the contention of this article that the photographic framing of text-to-image models that incorporate diffusion as part of their technical and discursive array raises some fundamental questions for our current ideas of representation and for our (human) ways of seeing and understanding the world. However, it needs to be acknowledged right from the start that 'diffusion' is only one component of the heterogeneous architectures of image-generation models, architectures that also include transformers and GANs,[2] with some models (e.g., Google's Parti) not including diffusion as part of its architecture. For our purposes here, the incorporation of 'diffusion' as both a technical *and* rhetorical device into many generative models is indicative of a wider tendency to build permeability and instability not only into those models' technical infrastructures but also into our wider data and image ecologies. Technically, 'diffusion' is a computational process that involves iteratively removing 'noise' from an image, a series of mathematical procedures that leads to the production of another image. Rhetorically, 'diffusion' operates as a performative metaphor – one that frames and projects our understanding of generative models, their operations and their outputs.

The following questions will thus be addressed through the course of this article: has our perception of the world, and hence our epistemology, been altered as a result of its mediation by generative models premised on unstable design architectures that incorporate noise, blur, diffusion and data compression? Has it become fundamentally unstable on a social level, with some going so far as to equate the developments in generative AI with nihilism or even fascism? Or can these developments be seen to be opening, in a Flusserian vein, some emancipatory possibilities? While the perspective adopted here is that of critical posthumanism (Wolfe, 2009; Braidotti, 2013), the species-specific notion of 'us' – i.e., technically constituted humans – anchors the argument, while remaining attentive to the bio- and geo-political power differentials within that feeble species unit(y). The argument will proceed along theoretical lines, but images will play an important role in it, both as objects of analysis and as enactive 'thought devices' whose role goes beyond that of mere illustrations.

## Photographic perception

Much has been made of the role of photography in the construction not just of our picture *of* the world but of the world itself. The radically creative force of photography was embraced in a series of political gestures by European avant-garde thinkers and makers in the 1920s and 1930s. For Ossip Brik, Aleksandr Rodchenko or László Moholy-Nagy, the post-pictorialist use of the photographic medium became a way of seeing the world anew but also enacting a New Vision for it. Moholy-Nagy declared photography to be capable of "bringing (optically) something entirely new into the world" (Wells, 2003: 92), thus connecting formal experimentation with socially progressive ideas. Tracing the history of 'photographic seeing' in the early- to mid-twentieth century, Liz Wells points out that, in Europe, aesthetic experimentation with the photographic medium was connected to a desire to enact social change, while in North America it focused more on personal expressivity and self-reflexivity, with transformation enacted on an individual level (2003: 82-84). This belief in the enactive power of photographic seeing was taken up by socially engaged writers on both sides of the Atlantic in the subsequent decades. For John Berger a photograph always encapsulated "a way of seeing" (Berger, 1972: 10). In her now classic 1973 volume, *On Photography*, Susan Sontag claimed that "photographs alter and enlarge our notions of

what is worth looking at and what we have a right to observe" (Sontag, 2005: 1), going so far as to suggest that "[o]ur very sense of situation" was enacted and confirmed by the camera's operations. "Ultimately, having an experience becomes identical with taking a photograph of it, and participating in a public event comes more and more to be equivalent to looking at it in photographed form" (18), opined Sontag in what sounded like a prophetic insight into the age of pervasive image sharing on social media in the early twenty-first century. Berger himself responded to Sontag's argument in his essay, 'Uses of Photography', where he argued, using the example of photographs featured in *Life* magazine, that they were not just *about* life. "[T]hese pictures *are* life", he declared, not without despair (Berger, 1980: 50). What Berger found particularly troubling about the comprehensive and detached God-like view that the camera producing those pictures adopted under industrial capitalism was the elision of the social context – human suffering, class-ridden injustice, exploitation – resulting in the free movement of images beyond history, memory and any trace of the real. In a searing diagnosis that seemed to forecast an era of not just social but also generative photography, Berger lamented the fact that, given that photographs carry no meaning in themselves, they "*lend themselves to any use*" (1980: 53; *emphasis added*).

The uses that especially preoccupied Berger in the late 1970s involved advertising and propaganda but he was equally exercised by the endless flow of decontextualised images. The free circulation of photos of war and suffering, detached from any inscription or narrative, resulted in anaesthetic perceptive experiences for viewers. To shift their mindless visual consumption in a direction of what could be termed ethical seeing, photographs needed multiple frames (or what Berger called "a radial system", 1980: 63), made up of layers of history, politics, aesthetics and economics. While Berger's human-centric and humanist concerns may seem idealistically old-fashioned in the era of generative AI, where *all* images present themselves to us as detached, decontextualised, substitutable, frameless and ephemeral, the strong ethico-political injunction underpinning his analysis has not lost any of its urgency, transcending in its force and demand present-day feeble calls for AI regulation, or even AI ethics. This urgency has been recognised by Mitra Azar, Geoff Cox and Leonardo Impett, editors of the special issue of *AI and Society* on 'ways of machine seeing', the goal of which was to investigate to what extent Berger's proposition – that the relation between seeing

and knowing involves a constant process of negotiation – still holds in the era of machine vision and algorithmic learning, when some of the negotiating agents may be determinedly nonhuman. Azar et al. suggest that in the current conjuncture "a new mode of perception is operationalized […]: a new way of (machine) seeing which is an assemblage of its various parts; including imaging devices (such as cameras), the data they produce (which might take the form of an image), and the wider practices and infrastructures through which they are operationalized" (2021: non-pag.). Perception still requires negotiation, but it may need to be conducted via human *and* machine languages.

Leaving aside, for now, the ontological status of image outputs produced by generative models that draw on photographic imagery and aesthetic to resemble their source material, it is the contention of this article that this new mode of perception *remains* photographic. Indeed, at the present moment the large part of the training data that fuels the generative models consists of billions of photographs (see Chávez Heras and Blanke, 2021; Malevé and Sluis, 2023).[3] The reason photographs are being used as the primary source in training machine learning models so widely is not just because of their ready (if not unproblematic, both in terms of privacy and ownership) availability through social media platforms, websites and archives, from Pinterest through to Wikimedia Commons. It is also to do with the assumptions around photographic images and representation manifested by many computer scientists, as evidenced in the following account of the construction of ImageNet, one of the early large datasets used for developing computer vision, by its lead researcher Fei-Fei Li. Launched in 2009, ImageNet started as a modest collection of 10,000 labelled images of different objects, divided into twenty classes. It was only when the collection was scaled up to 14 million items through accessing vast swathes of 'free' photos across various Internet platforms, and then labelling them with the help of cheap labour available via Amazon's Mechanical Turk online marketplace, that the project really took off. "Li was asked recently about the choice of using photographs for ImageNet during an event celebrating the tenth anniversary of the dataset: 'That's a great question.' – she replied – 'We didn't really stop to think much about it […]. I suppose we wanted as a realistic representation of the world as possible' (Li 2019)" (cited in Chávez Heras and Blanke, 2021: 1155). In this account photographs are just 'there', freely available for

researchers and the tech industry to come, take, catalogue, tag and use them in any way seen fit, an issue that raises vast moral and legal questions. More significantly, photographs are implicitly positioned in computer science research as windows onto the world, offering a direct and realistic picture of what is 'out there'. They are seen to present and represent directly multiple singular objects and actions that can be enclosed in discretised frames, tagged and confirmed as representatives of semi-eternal forms that transcend time and space.

We may thus conclude that the industry premised on perfecting the Turing machine, a differential calculating system that can be put to open-ended uses, surprisingly relies on a much simpler technology as its conceptual foundation. This is a technology that that could be called, with a nod to Yanai Toister, a Turin Shroud model (with the name referencing a Catholic religious artefact that supposedly served as a wrap for Jesus's dead body and featuring what looks like his face). In this model photographs are "viewed as instances of privileged if not miraculous transference" (Toister, 2020: 3). Computer scientists seem to remain untroubled by the debates on the limitations of indexicality in the photographic medium that have worried photography theorists for decades. In the datasets of machine learning, photography is seen to be espousing "an ontological privilege" (Toister, 2020: 3) through which external objects manifest themselves in images, "setting up a supposed uniformity between looking at a photograph and looking at the world" (4). This belief applies not just to what might be termed first-level photographs (images of humans, animals, landscapes and inanimate objects taken for social, documentary or artistic reasons) but also the photographic documentation of artefacts (paintings, sculptures) or photo-scans of flat objects (graphics, maps, texts). There is no room for the framing of photographs through layers of memory, history, economics and aesthetics, as envisaged by Berger: the image in the training dataset is an emanation of 'the thing in itself'; it is a direct line to, and a condition of, object recognition.

Importantly, it is not just (predominantly) photographic images that feed the databases of machine learning but rather image-text pairs, which allow computers to identify an object through its designation, aka label. The technology of mining the web to collect and label the images used in machine vision applications has changed over the years. While with ImageNet it was anonymous human workers who tagged the images, with

LAION-5B – a public dataset containing 5.85 billion captioned images, which underpins Stable Diffusion – an automated program called Common Crawl harvested the image-text pairs from a variety of websites, such as Pinterest through to WordPress, Blogspot, Flickr, DeviantArt and Wikimedia Commons. The images had been posted – and tagged – by internet users over the years, with the collective unpaid body-and-mind, with all its knowledge, ignorance, bias and prejudice, reflected in the dataset. Computer scientists do of course know that the data is not perfect: mentioning (albeit not resolving) issues of bias, as well as those of copyright, seems imperative in many computer science papers today. The guiding assumption in the majority of them is that the datasets will improve with time and that further accumulation of data will average out any potential issues, getting us closer to objectivity and truth. What remains consistent across mainstream applications of machine vision is a common-sense positivism with regard to images and their signification, or the elision of the representational gap between the image and its referent. Allan Sekula already grasped this tendency in 1975 when he pointed out, not without irony, that "Nothing could be more natural than […] a man pulling a snapshot from his wallet and saying, 'This is my dog'" (Sekula, 1975: 37). In computer science today, that man's name is legion, their dog may be a cat – and it is being shown at scale, not just to human but also to machine eyes.

The structuring logic of perception in machine vision is premised on the following conceptual and technical manoeuvres: the elision of distance, the collapse of the differentiation between foreground and background and the flattening of form, whereby 3D objects *in* the world are represented, or rather rendered, as 2D models through the selection of their 'features', i.e., elements that are readable by machine vision systems. An intriguing process of cutting and carving unfolds within the frame of the picture, enacted by inserting multiple inner frames into it to contain seemingly discretisable objects. A photograph is thus "flattened into a collection of objects to label" instead of being seen as "a cohesive whole or a composition with relational meaning" (Wasielewski, 2023a: 195-196). Amanda Wasielewski explains that, since machine vision algorithms are in the business of creating generalised models that enable object identification, the photographs used to train the database need to be just good enough.[4] The shrinkage of the image on the level of data is coupled with its

semiotic flattening, or what Nicolas Malevé and Katrina Sluis describe as a move from *representation* (which involves the mediation of an object through photographic technology) to *representativeness* (which consists in providing a supposedly representative sample of photographic images). They explain this operation as "a sleight of hand", enabling computer science researchers "to sidestep the cultural context of each individual image in order to privilege the problem of statistical distribution" (Malevé and Sluis, 2023: non-pag.). Generalisation is coupled with the principle of averaging: the image of a house needs to resemble, on some constitutive level, other pictures of houses so that its 'houseness' can be extricated as a sequence of data points corresponding to what are deemed to be the relevant features of the represented object.

To sum up, what has changed in the photographic perception of the world with machine vision is not just that the viewer of the images is no longer human (Paglen, 2016), but rather that the photographic image itself can be described as nonhuman (Zylinska, 2017), being subject to a plethora of data operations involving fragmentation, extraction, reduction and elision that go beyond human intentionality and agency. Humans are still partly involved in these operations, not least in the foundational decisions about reducing objects to images and rich imagistic data to extractable and exchangeable sequences of features.[5] But it is in the adoption of the Goldilocks principle of 'good enough images', which are transformed from representations to schemata and then models, that the uniqueness of this particular iteration of photographic seeing lies in generative AI. In Stable Diffusion, it was the encoding and decoding of images in so-called 'latent space', i.e., a simplified mathematical space where images can be reduced in size (or rather represented through smaller amounts of data) to facilitate multiple operations at speed, that drove the model's success.

The expansion of the photographic mode of seeing to machine vision entails a significant epistemological shift, one that has consequences for our conventional, humanist *ideas of understanding images* – as both photographs and pictures of ourselves. Compelling arguments about the failure of traditional photography theory, be it concerned with the meaning of a photograph (Barthes, 1977) or its uses (Berger, 1980), to account for the transformation of photographic images by networked computation

have been made previously (Paglen, 1996; Dewdney, 2021; Toister, 2021). Yet, importantly, this expansion also has serious consequences *for our very idea of understanding*. The production of photorealistic, semi-realistic and hyper-realistic outputs by generative AI models is putting an urgent demand on philosophers, media theorists, scientists and the general public to engage anew with the very notion of what it means to understand the world – and of who, or what, is capable of executing such understanding.

## Nonhuman understanding

The debate on understanding in machine learning as manifested (or not) in generative image models or large language models has divided the scholarly community. According to a review paper by computer science researchers from Santa Fe Institute, Melanie Mitchell and David C. Krakauer, in a 2022 survey 51% of the researchers questioned agreed that such models "could understand natural language in some nontrivial sense", while 49% disagreed (Mitchell and Krakauer, 2023). Those in the 'aye' camp usually adopt a functionalist approach, proclaiming that if a model's behaviour *looks like understanding* to an external observer, then it should be treated as such. For the nay-sayers, such models can only *simulate* understanding: they are essentially what Emily Bender and colleagues have described as "stochastic parrots" (Bender et al., 2021), mimicking humanlike understanding through complex statistical correlations unfolding at scale. AI software engineers and their financial backers tend to identify with the functionalist position. This (perhaps at times performative) stand leads many of them to issue hyped-up promises as well as veiled threats about AI imminently surpassing humans on many levels. It also allows them to position themselves in the role of both Dr Frankenstein and humanity's saviour. Humanities scholars, in turn, have typically been critical of pronouncements about AI reaching levels of human understanding any time soon (or, indeed, ever).[6] For philosopher John Searle "the thought experiment underscores the fact that computers merely use syntactic rules to manipulate symbol strings, but have no understanding of meaning or semantics" (Cole, 2023: non-pag.). In other words, computers perform "a calculation of meaning" (Bunz, 2019). Mercedes Bunz goes so far as to posit that present-day AI systems are merely "*imitating the understanding of meaning* by calculating it, but they are *not*

*understanding* – they lack the ability to link their classifications in an integrated way to a wider, constantly shifting context" (Bunz, 2019: 273, *emphasis added*).

Wasielewski's argument about understanding in text-to-image generators is very much aligned with the post-Searle position. She writes: "DALL·E and its ilk are able to replicate visual forms but are not 'aware' of or 'familiar' with the referents in the images they produce, i.e., they have no experience of the physical objects, people, or places depicted in the output images" (Wasielewski, 2023b: 78). Human viewers, in turn, "have had a full sensory experience and accompanying contextual understanding of these objects that far exceeds the information that can be learned from a digital image (or even thousands of digital images)" (78).

Wasielewski continues, in a section worth citing at length:

> [T]ext-to-image generators are very good at identifying the image of something that is input as a word. However, this still does not mean that it [i.e., the model] understands what that image actually is or what it represents. […] [They] are very good at extrapolating from the pixel patterns labeled 'dog' and those labeled 'beach' and creating an image of a dog on a beach. The model is merely learning the variety of things in a two-dimensional image labeled 'dog' and the variety of things labeled 'beach'. It does not understand either of these concepts beyond the limits of two-dimensional visual patterns that have been labeled to create image-based representations. In other words, image generators have a very limited understanding of the forms found in our world because they deal only in digital images (79-80).

As an illustration of this lack of understanding Wasielewski cites Midjourney's inability to draw 'normative' human hands accurately – even if explicitly prompted to do so (fig. 1). She detects the same problem in DALL·E and Stable Diffusion, with all AI image generators having a tendency to portray human bodily parts "in ways that are completely fantastical" (73). This indicates for her a deeper struggle those models have with "the creation of meaning" (71), relying instead on its calculation – a state of events that results in meme-like user responses along the lines of "Mj [Midjourney] can't count" (78). Artist and educator Eryk Salvaggio makes a similar point when

highlighting DALL·E's problem with creating credible images of kissing humans (fig. 2). "The 'strong' pattern across the kissing itself is that they are all surrounded by hesitancy, as if an invisible barrier exists between the two 'partners' in the image. The lips of the figures are inconsistent and never perfect", he says (Salvaggio, 2022). While neither Wasielewski nor Salvaggio embrace bodily or behavioural essentialism – they actually go to great trouble to challenge prescriptive and ableist assumptions around corporeality, health, sexuality and gender – their respective analyses nevertheless rely on a certain idea of *correctness* as a ground for their critique. For the tech companies, the issues identified by those scholars in late 2022, early 2023 will no doubt be seen as mere technical glitches that will be resolved with more data. Eventually, better statistical distributions of pixels will likely result in the elimination of the 'hand' and 'kiss' outliers, i.e., images that miss the mark of presumed correctness. Yet that time-specific discussion with regard to what human bodies look like and what they do is indicative of the deeper question that is of concern to the problem of this article: on what terms and according to which criteria can we contest AI models, with their photographic perception of the world, without falling into the representationalist trap?

**Fig. 1:** "An absurd image of hand-toe-finger amalgams created from the prompt 'Children's hands reaching for candy' with Stable Diffusion, January 2023" (Wasielewski, 2023b: 73)

**Fig. 2:** "Image of two humans kissing. Generated by OpenAI's DALLE2" (Salvaggio, 2022)

In the era of alternative facts, when photographic images can be put to *any use* while deepfakes have become embedded in both marketing and propaganda, the issue of the accuracy of representation should certainly be taken with seriousness and care. Yet it is the contention of this article that the 'not good enough' argument with regard to generative models' *modus operandi* is itself inadequate because it ends up narrowing the conceptual scope through which generative AI can be engaged, tested – and contested. The 'they know not what they do' mode of critique, premised on denying AI any form of intentionality, even of a weak kind, ends up reinserting a set of positivist assumptions about the world, ways of seeing it as well as processes involved in imaging it, not to mention assumptions about humans as self-contained transparent entities whose intentions and desires can always be clearly understood and mapped out. Without giving up on a critical analysis of the conceptual and technical limitations of generative models, and without clamouring for AI as fully intentional, it may therefore be worth asking whether these 'incorrect' images could be perceived and read differently, by both humans and machines. Stylistically, those 'completely fantastical' and 'hesitant' renderings have an affinity with the visuality of the earlier avant-gardes: the dreamy aesthetic of surrealism, the technicist look of the collage and photomontage. They evoke Etienne-Jules Marey's bodies in movement recorded on film through continuous or multiple exposure, the photographic captures of Pina Bausch's dance acts, the images of Stelarc's performance with The Third Hand.

Castigating a generative model for getting things 'wrong' thus only ever makes sense if we are to assume that the primary function of generative AI technology is to produce verisimilitude, i.e., to create more of what we already have. Yet what if, rather than seeing this technology as premised on delivering an accurate response to a natural-language-defined prompt, we saw them as conversation pieces, provocations – or, indeed, prompts – opening up a dialogue, with us and for us, on the fundamental incommensurability between the word and the image, between the world and its representation? It may be argued that we do not actually need AI to make this point: formative texts of the humanities, from Aristotle's theory of mimesis through to poststructuralism, have already brought it home. Yet now that critical humanities education is being increasingly dismissed as both unscientific and unproductive, with STEM subjects awarded the knowledge crown, can we not attempt to divert the

seemingly inevitable trajectory of AI development by raising some fundamental questions about the creation of meaning and understanding, in *all kinds* of artificial intelligences – including that of the *human* kind? Human intelligence is of course also an artifice because it is an outcome of a creative, i.e., poietic, process, a process of bringing something forth over time. It consists of an acquired set of traits that allow us to learn new things, understand concepts and adapt to the environment, while itself emerging in relation with said environment. (If the timeline of that process is long enough, we call it evolution, while its shorter stretches get designated as culture.)

As we can see, the debate about representation in generative AI brings to the fore the meta problem of understanding *understanding*, and of perceiving the images and models upon which it is supposed to be based. Mitchell and Krakauer point out that "humanlike understanding", even though lacking a rigorous definition, is related in cognitive science to the ability to create and operationalise concepts, i.e., "internal mental models of external categories, situations, and events and of one's own internal state and 'self'" (2023). This ability allows people to make predictions, generalisations and analogies; to reason – and, last but not least, to explain their understanding to others. Many humanities scholars claim that machines can't create concepts, or at least that they can't create what look like meaningful concepts to humans (Bunz, 2019; Bender et al., 2021). Yet is it too preposterous to suggest that machines may be *creating something*: they may be enacting some novel ways of parsing data that make sense *to them*? Should this be the case, then may we not surmise that some of the patterns produced may remain invisible and illegible to humans – but not to other machines? When humanities scholars do recognise a machine producing such 'something' – a (mental) picture, a concept – they may be tempted, following Searle and Bender, to dismiss it as a thing that *just looks like a meaningful image or concept*. But what if there is *something* there: some other patterns and pictures, other ways of linking things, other models, logics and forms of connection emerging that we humans cannot see? Mitchell and Krakauer wonder whether, even if we are to accept that these systems do not produce concepts as we understand and recognise them, i.e., as causal mental models, their large systems of statistical correlations can

> produce abilities that are functionally equivalent to human understanding. Or, indeed, that enable new forms of higher-order logic that humans are

> incapable of accessing? And at this point will it still make sense to call such correlations 'spurious' or the resulting solutions 'shortcuts?' And would it make sense to see the systems' behavior not as 'competence without comprehension' but as a new, nonhuman form of understanding? (2023)

Raising the question about the nonhuman mode of understanding does not need to amount to giving technology companies a free pass, one that would allow them to get away with not just bias but also discrimination, exclusion and other forms of exploitative plundering of human and natural resources because of the presumed, even if uncertain, greatness of their product. But it does mean having to consider generative AI as being capable of creating possibilities for us to unsee and unthink ourselves from our own all-too-human modes of being and acting, with all their biases, exclusions and injustices. To take just one example: what does it mean for an AI model to get us to think what 'a good kiss' may involve and look like? Could it be mobilised to reimagine corporeality and sexuality beyond the restrictive, all-too-binding heteronormative, humanist assumptions (fig. 3)? Of course, artists, filmmakers and LGBTQ+ communities (from Leon Carax with *Holy Motors* through to Shu Lea Cheang's Mycelium Network Society) have been involved in reimagining precisely that. Yet, as 'the norm' concerning what we humans have designated as culture and nature has been only temporarily stabilised (see Fausto-Sterling, 2000; Daston and Gallison, 2010), the likes of Stable Diffusion reveal the shaky ground on which the visual and cultural consensus with regard to our bodies and lives has been established. For a time at least, they are allowing us to see the stitching – even if some insist on calling it a glitch.

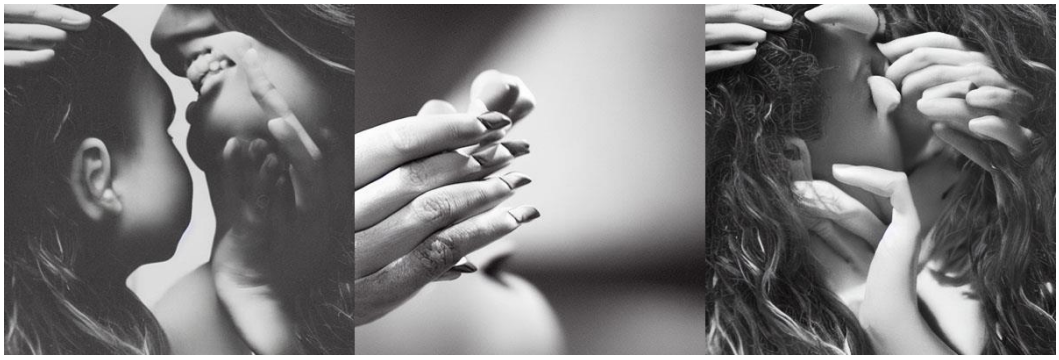

**Fig. 3:** Triptych made by the author from selected images produced by Stable Diffusion in response to the seemingly absurd prompt 'fingers kissing', in an attempt to explore purposefully the model's ways of imaging beyond verisimilitude. Images originally obtained in black-and-white, gentle curve adjustment applied. October 2023.

To reiterate, ideas of intentionality and purposefulness need to be raised in the context of generative AI – but they need to be raised *constantly*, with regard to both humans *and* machines. The situation when judges would, on average, pass more lenient verdicts after a food break in response to their blood sugar levels rather than complex understanding of jurisdiction (Danziger, Levav and Avnaim-Pesso, 2011) or when a drowning person would be rescued by a bystander primarily thanks to the latter's impulsive bodily reaction rather than rational deliberation on morals and values[7] are just two examples of the entanglement of human intentions with corporeal responses, impulses and drives. "Artists and athletes intuitively know that they have to make their next move without even thinking about it", observes Mark Amerika in his provocative book, *My Life as an Artificial Creative Intelligence*, coauthored with AI (Amerika, 2022: 4).[8] To dismiss generative AI for its lack of intentionality is thus to ignore the fact that much of *human* activity – including its 'generative' variation recognised as 'art' – has also been produced in the state of at least partial suspension of the artist's explicit intentionality and clear mind. Many paintings, drawings, poems and installations, from André Breton's pure psychic automatism through to Jean-Michel Basquiat's intense creative visions, have been made by artists experiencing altered states of consciousness and agency through meditation, mania, ingestion of foreign substances as well as entering a process of cocreation with other agents. Artists working with machines, from robots through to neural networks and generative AI models, only bring to the fore that ongoing distributed agency of the creative process (see Zylinska, 2020: 52-55). Generative AI, whose models are deemed by some as lacking understanding, can serve for us as a reminder of the fact that humans do not always fully understand, or have full mastery of, what they do – and why they do it.

The relationship between understanding and its lack cannot therefore be discretised as a theoretical problem that falls on two sides of a neat binary, not least because its assessors, i.e., humans themselves, are also its legislators and enactors. Any discussion regarding the understanding of anything, including the meta-understanding of understanding, needs to be conducted through the bodily apparatuses of human subjects, whose very constitution is also already technical. With this, we are arriving at the following proposition: generative AI may be seen to be staging a problem of understanding *for us*, while offering us a different concept of understanding. We could

perhaps suggest that generative AI produces what could be called an 'unstable' or 'wobbly' understanding – and a related phenomenon of 'shaky' perception. Diffusion, to be discussed in the section that follows, can be seen as an imaging template for this model.

## Un-stable diffusion[9]

The photographic model of perception underpinning machine vision, discussed in the first part of this article, is itself framed through a rather telling set of terms and concepts. Before we go on to explore the desire to stabilise perception – *and its perception* – in the current projects of machine learning such as Stability AI's Stable Diffusion, it is worth looking at some historical antecedents of machine vision, and their rhetorical and kinaesthetic support. It is because such moments of technical invention can serve as crucial points in our own human revaluation of ourselves. They become instances when we can assess the hypotheses and assumptions behind those inventions with a view to looking at ourselves anew in their technical afterglow.

The curiously named 'Shakey the robot' was one of the early devices of machine and computer vision that was supposed to appeal to a wider public. Developed in 1970 by the Stanford Research Institute, it gathered information about the environment around it thanks to the static pictures received from a TV camera installed on it. Its name was well justified, explains James E. Dobson in his book, *The Birth of Computer Vision*:

> Shakey was an unstable project for several reasons. The device […] moved through space in a jerky manner. This was not because of delays in planning, although these also added to the instability of the device, but rather the state of robotics in the 1960s. Shakey was initially loosely connected by a radio transmitter to a large and immobile SDS 940 general-purpose digital computer that needed to be physically located in close proximity to the robot. […] The project itself made progress toward the development of computer vision in a haphazard fashion, and the results were frequently far less impressive than those promised in the project proposals. The Shakey robot and the larger project were designed to satisfy DARPA's requirements for 'automatons capable of gathering, processing,

> and transmitting information in a hostile environment' (Dobson, 2023: 138).

Dobson acknowledges the robot's significant contribution to changing the research direction of the field of machine vision, from relying on two-dimensional images to becoming "a situated sensing device" (139). But he also recognises, not without irony, that the name (originally referring to the whole project and not just the robot) reflects the engineers' shaky belief in the possibility of developing a system that would be capable of operating in the actual world. The subtly playful rhetorical framing was perhaps a way of mitigating the promise right from the project's start; it was a way of managing expectations while acknowledging that both the research idea and its execution were unfolding on a rather unstable ground.

New developments in computer and machine vision in the current 'AI summer' of the third decade of the twenty-first century, especially in the field of generative AI, seem to have done away with such rhetorical modesty or qualification, at least at first glance. This is most evident in Stability AI's flagship product: the generative text-to-image model that gained the name Stable Diffusion. Yet the explicit disavowal of any kind of shakiness or wobbliness as articulated in both the company's name and its AI model is counterweighted by the incorporation into the latter of a rhetorical figure with a rather different set of characteristics: *diffusion*. What's more, diffusion is at the core of not just Stability AI's model but also of those of its key competitors: Midjourney Inc.'s Midjourney (a product whose name literally states: 'we are not there yet'), Google's Imagen, Adobe's Firefly and OpenAI's DALL·E 2.

Looking at metaphors used to describe AI models may seem like a supercilious humanities-style attempt to discredit a complex computer science project on a purely rhetorical ground, without getting to grips with the actual technology underpinning it. We *will*, however, consider the technological side of diffusion. But it is worth pausing for a moment to acknowledge that metaphors matter a great deal for how science and engineering are framed, articulated and projected into the future: they are not just ornamental but also performative. Matthew Cobb argues that metaphors are "central to the way scientists think" (2020), with scientific views partly shaped by the dominant technical concepts of the moment – which in turn begin to serve as inspiration for not

just describing but also *doing* science. Cobb goes so far as to claim that metaphors and analogies can actually alter how scientists "understand their work, or even enable them to devise new experiments" (2020). The current dominant metaphor across neuroscience and computer science – superseding references to 'imprinting' in photography and 'the showreel' cinema through which perception used to be explained – is that of the brain as a computer, an organ that supposedly processes data in the same way a difference engine does. Further metaphorical loops have subsequently been initiated in AI research, with machine-learning computational processes, including those of the generative kind, being presented as working on the same principle as human brains: hallucinating, dreaming, making unexpected connections. It is precisely into this nested metaphorical structure that the figure of diffusion seeps in, with all its fluid wobbliness.


**Fig. 4:** An illustration of how Stable Diffusion uses noise removal to produce an image. This example shows the model's response to the prompt 'Hyperrealistic photo portrait wide shot of a cyborg, city background' enacted in 10 steps, using 10 different sampling methods (aka samplers), from 'Euler a' through to 'DPM++ 2M Kansas'. Each sampler has its own way of applying noise reduction to obtain an image.

Borrowed from thermodynamics (Ho, Jain and Abbee, 2020: 1), the term 'diffusion' refers to a slow dissolution of one substance within another. Within the realm of computer science, both 'substances' are quite insubstantial: they are data, with noise being the starting point. In other words, the generation of an image through a diffusion process consists of taking a random noisy image and then statistically 'dissolving', or removing, this noise, through a number of iterative steps, with a view to arriving at an image of a desired object. The noise is removed using a probabilistic sampling method, a mathematical procedure that calculates how much noise should be removed at each step. That desirability is usually verbally described by a user (in a prompt), but the description can also be coupled with an image. Image generation is a reversal of the prior training stage, whereby noise (called Gaussian, or random) was gradually added to an image through a number of stages.

## Negentropic hope

What does the placement of diffusion and noise, not just as rhetorical devices but also as mathematical and technical practices, tell us about ways of seeing the world? While we pointed earlier to an epistemological gap between the object and the image, we arrive here at something much more fluid: instability as the organising concept and technology for the emergence of our picture of the world. Indeed, it is not just the perception of images but their very constitution that is fundamentally unstable, premised as it is on a sequence of (technical, cultural and economic) operations whereby Marshall Berman's critique of modernity contained in the title of his book, *All That Is Solid Melts into Air* (1982), is literalised.[10] Berman's analysis pointed to the liquidising of all the certainties of "the traditional world", while offering a Marxist correction to modernism's fluidity as a way of re-channelling the flows of capital and society in a more progressive direction. This sentiment and mode of thinking has been taken up by contemporary Marxist critics of AI. For example, David Golumbia has argued that generative AI has been built "on very dark and destructive ideas about what human beings, creativity, and meaning *are*", equating the project with nihilism – leading "directly" to fascism (2022). Golumbia has gone so far as to declare: "ChatGPT and other Generative AI programs should not exist. They are not the kinds of things that someone who cares about human life would build" (2022). Counter-opposing

'human meaning' and those who care about it to those who cannot see the fascism for the dispersed smoke and mirrors of generative AI's vapour, he sees the relationship between humans and AI as an existential battle in which there can only be one victor. In a similar vein, technology critic Paris Marx has declared that "Generative AI closes off a better future" (2023). As generative models "can only take what already exists" (Marx, 2023), they are unable to produce something truly novel, be it on the level of image or imagination.

Rather than read Golumbia's and Paris Marx's critiques literally (and then dismiss them too quickly as just too one-sided and too uncritically humanist), perhaps it would be possible to look at, and build on, the sentiment behind their respective arguments – something that, incidentally, generative AI is (yet) unable to do. We could thus see in their scathing repudiation of generative AI a strong moral critique of the exploitation of human and natural resources. We could also identify a political articulation of a desire for a better, fairer world, unencumbered by excessive capital and its masters in the form of Big Tech – and a willingness to work towards a future that would be full of possibilities for the many, not the few. As part of the more generous 'sentiment reading' of Golumbia's and Paris Marx's critiques, it is worth recognising the already existent work on the political economy of AI; on labour issues and the valorisation of cultural endeavours in a society where art is reduced to an infinite production of pointless outputs; on the critique of not just bias but also 'political redlining' and deeper injustice enabled by some applications of algorithmic technology; on the ecological cost needed to run them (Noble, 2018; Zylinska, 2020; Crawford, 2021).

Without wanting to give up on the progressive desire, can we, however, move beyond the 'humans vs machines' dualism to embrace the possibility that these (and other) technologies are not something that is happening *to us*? If intelligence is artificial not just in machines but also in humans, if we have co-emerged, as human subjects, through *tékhnē* (tools, clothing, shelter, communication), the statement that generative AI 'can only take what already exists' does not need to be seen as an indictment of this technology but rather as a matter-of-fact recognition of its materiality. This approach will only work, however, if we are prepared to embrace a form of entangled and scalar thinking that is capable of considering the cellular, the molecular and the quantum (even if not necessarily at the same time, with the same set of affordances in each case).

It is indeed the argument of this piece that we do need that latter acknowledgement if we are to create a better progressive vision for, and perhaps with, generative AI, one that is not locked to the short-term articulation of its backers – or, indeed, critics. This vision would require a less moralistic and less declarative understanding of not just art and other forms of human creative practice but also understanding.

And it is in the concept of diffusion and its underpinning idea of 'photographic seeing' as applied in machine vision that such a promise could arguably be sought. In generative image models such as Midjourney or Stable Diffusion, diffusion is the quintessential figure of entropy – i.e., degradation of the system towards informational uniformity. Yet entropy on the training level undergoes a negentropic reversal process when a new image is being generated from noise, or 'thin air'. Due to the black-boxing of AI technology, there is something mysterious, almost magical about this moment, the way it presents to human observers – who are unable to observe much of the process. But the wider context of the unfolding transformation, with its multiple layers of change, is notable from a scientific, technical *and cultural* point of view. Though we start from looking at the actual technique of diffusion as implemented in the key text-to-image generation models, the concept of 'diffusion' can also be read as a symbol of the deeper tendency towards instability as well as material and financial dissipation that is built, albeit not explicitly, into all the generative AI systems (including those where the actual diffusion model does not feature) – and into the wider media ecologies they form. Using feature extraction and semantic compression as part of its earlier discussed 'averaging principle', this tendency entails a diffusion of a different kind: the *dissolution* of "politically salient human concepts" such as race or gender, and their subsequent re-constitution as "ahistorical, apolitical, and non-ideological" (Offert and Phan, 2022: 3).

In the sciences, the use of the concept of entropy has travelled from thermodynamics, where it refers to the gradual dissipation of matter and energy, leading to the eventual heat death of the system, to information theory, where high entropy stands for low information value, uncertainty and chaos. Writing in the 1980s, by which time information theory had made significant inroads into social sciences and wider social thinking, Vilém Flusser mobilised the concept to explore not just its destructive but also emancipatory potential. We are once again turning here to the use of metaphors

as agents of change across different disciplinary fields, although this time it is humanities scholars – philosophers, media theorists – who are borrowing from science ideas to articulate their visions of and for the world. Flusser draws on entropy in his reading of both communication and creativity (see Booten, 2020) – which for him always happen within the format of a larger socio-technical apparatus, but which also contain a negentropic tendency. In the essay 'On the Theory of Communication' (2002b) he gives an example of a university lecture, which is entropic, as it is a natural thermodynamic process during which one system (the public) is changed by another (the lecturer transmitting sound waves to the first system through the medium of the room's air). Energy is gradually transformed into heat, as a result of which entropy increases. But he also recognises the *cultural* aspect of the event, whereby information (in the sense of knowledge) increases in the first system, actualising a negentropic tendency within the room – albeit one that can only be recognised by observers to whom such information is meaningful. Flusser is aware that "in the longer term, the autonomy of the apparatus" wins over the liberatory efforts of human beings to move towards heat death (Flusser, 2011: 19). He is, however, interested in those possible moments of interruption, when negentropy (or "form-giving", 2002b: 20) can be actualised, through the actions of a Maxwell's demon,[11] to halt or divert the process, at least temporarily.

We are turning to Flusser here not only because he can help us identify a counterforce within the process of diffusion but also because he affords photography a unique place within systems of communication when it comes to countering the entropic condition. A critical posthumanist *avant la lettre*, Flusser is aware of the indifference of the larger processes unfolding across the universe to the scale of the human. Yet that particular human scale, with its historical modes of understanding, meaning making and artefact creation, matters to him a great deal. Seeking an emancipatory promise from within the technical system, he points to in-formers, i.e., those who give form to the chaos of the world, as agents capable of making a meaningful intervention into large-scale processes of both the universe and the technical apparatus of which we are all part. It is photographers in particular who are recognised by Flusser as being able to fulfil that role, because he sees them as being capable of breaking the habit – "the aesthetic equivalent of 'entropy' in physics" (Flusser, 2002a: 53) – to see and do things

differently. Naturally, not everyone equipped with a photographic apparatus, such as a camera, a scanner or image manipulation software, will be able to achieve this: Flusser is well aware of the overbearing processes of habituation that dominate all forms of human cultural activity, from thinking through to image making. Yet it is this *possibility of doing things otherwise* that creates an opening within what may look like a predetermined game of chance, driven by "a fundamental tendency toward becoming continually formless" (Flusser, 2002c: 129). Photographs, argues Flusser, "are intentionally produced, negatively entropic clusters. Negative entropy can be called 'information.' From the perspective of formal consciousness, photographs are information intentionally produced from a swarm of isolated possibilities" (129). They are potentially anti-diffusion devices – even if the context to which they are currently being put often generates confusion and chaos.[12]

## New visualisation

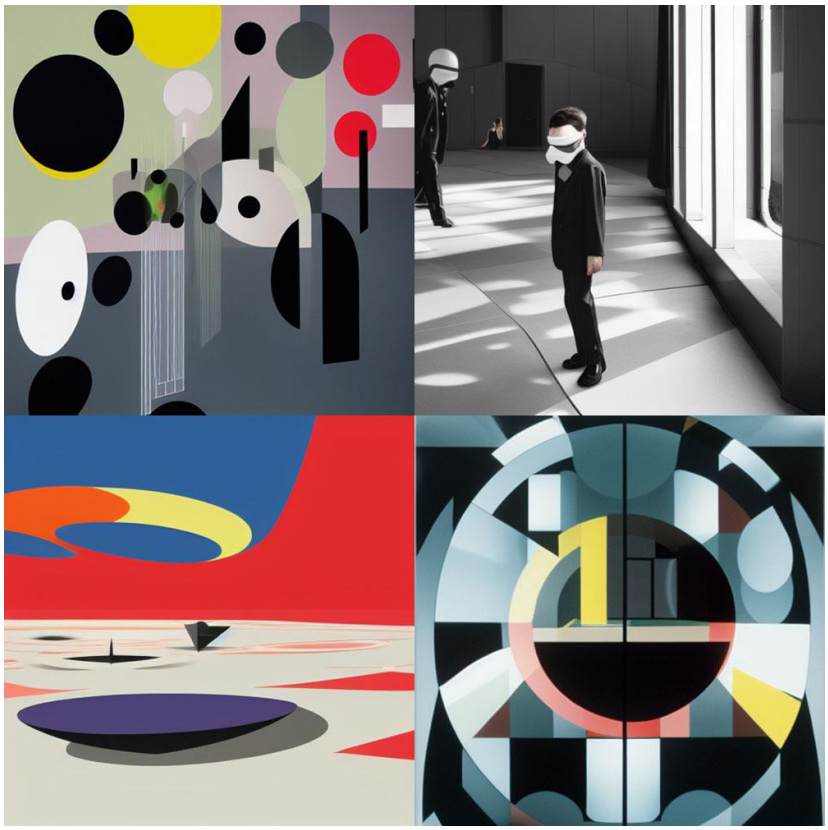

**Fig. 5:** Selection of images made by the author from Stable Diffusion's response to the prompt 'Photographic image of the contemporary world in 2023 in the style of Moholy-Nagy New Vision', using different CFG Scale values. October 2023.

Flusser's most important texts on photography were produced in the 1980s, a mere dawn in the era of digital transformation in image making. Yet intentionality for him was always systemic rather than purely human. Photographs were thus outcomes of the activation of technical forces enfolding humans and their apparatuses, and not just products of human intention. Today the relationship between the photographic image and its processes of production and distribution – and of the agents involved in them – is significantly altered: not only are we faced with the situation whereby the majority of photographic images are not made with a human viewer in mind, but also it is often difficult to distinguish between a photographic image and a synthetic one that *looks like* a photograph but that has been produced by a generative AI model. Yet that uncertainty about the ontology of the photographic image has been present in the medium since its inception, as evidenced by spirit photography, avant-garde photomontage and digital photographic manipulation.

As has been argued throughout this article, photography persists today not just as a memory but also as a structuring technology of AI databases – and a creative force for future imaging. Paraphrasing Geoffrey Batchen (1997), [13] we could say that photography persists in a *desire for photography*, even if the technology mobilised to enact this desire is not entirely light-based. Salvaggio posits that image outputs produced by generative models such as Stable Diffusion, Midjourney or DALL·E 2 are ontologically and technically speaking not photographs but rather visualisations, or infographics; they are "data patterns inscribed into pictures" (Salvaggio, 2022). Yet we should be mindful of earlier definitions of photography as precisely "the engine of visualization" (Maynard in Toister, 2021: 8): a history that makes Toister rename the photographic medium in more explicitly computational terms as "'the program of visualization', or more simply 'A Visualization Turing Machine'" (8). Toister's argument is Flusserian in that he recognises the generative possibility identified by Flusser in the photographic medium. For Flusser photographs are "computed possibilities (models, projections onto the environment)" (Flusser, 2002c: 129) because they are premised on the calculation of dot elements and their subsequent computation. Flusser's argument refers to analogue photography but it applies equally to other forms of what he called 'technical images': digital photography as well as sorts of after-photographic images. Figures of visualisation rather than imagination (because

they do not require one to remove oneself from the environment "to create an image of it"), photographs are capable of turning "a swarm of possibilities into an image" (129). Visualisation is the name Flusser gives to "the power to concretize an image from possibilities" (129), enabling a projection into the future – and also *of* the future.

It is in this sense that we can understand all kinds of technical images, including those produced by generative AI, as possible carriers of the future: of future meanings, future projects, future lives. Naturally, there is no guarantee of what that future will look like: that it will be progressive rather than fascist, life-enhancing rather than exploitative. But it is in the inherent combinatorial possibility of technical images that we can seek alternatives to the mournful melancholia of Marxist humanism – and the deranged optimism of Big Tech (fig. 5). And it is in that *desire to visualise anew*, assembling alternative if shaky visions of justice and politics, that we can find the solace of there being a future in the first place.

## References

Amerika, M. (2007) *META/DATA: A Digital Poetics*. Cambridge: MIT Press.

Amerika, M. (2022) *My Life as an Artificial Creative Intelligence*. Stanford: Stanford University Press.

Azar, M., G. Cox, and L. Impett (2021) 'Introduction: Ways of Machine Seeing', *AI & Society* 36: 1093-1104. https://doi.org/10.1007/s00146-020-01124-6.

Baio, A. (2022) 'Exploring 12 Million of the 2.3 Billion Images Used to Train Stable Diffusion's Image Generator', *Waxy*, 30 August. Available at: https://waxy.org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-image-generator/ (Accessed: 10 January 2024).

Barthes, R. (1977) 'The Photographic Message' in *Image, Music, Text*, trans. S. Heath (ed.). London: Fontana Press, pp.15-31.

Batchen, G. (1997) *Burning with Desire: The Conception of Photography*. Cambridge: MIT Press.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell (2021) 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?', in *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and*

*Transparency*. Association for Computing Machinery: 610- 623. DOI: 10.1145/3442188.3445922

Berger, J. (1972) *Ways of Seeing*. London: British Broadcasting Corporation and Penguin Books.

Berger, J. (1980) 'Uses of Photography', in *About Looking*. New York: Pantheon Books, pp.48-63.

Berman, M. (1982) *All That Is Solid Melts into Air: The Experience of Modernity*. London and New York: Verso.

Booten, K. (2020) 'Flusser's Demon: Writing Under the Eye of an Automatic Critic', *Flusser Studies* 30(November): 1-20.

Braidotti, R. (2013) *The Posthuman*. Cambridge: Polity.

Bunz, M. (2019) 'The Calculation of Meaning: On the Misunderstanding of New Artificial Intelligence as Culture', *Culture, Theory and Critique*, 60: 3-4, 264-278. DOI: 10.1080/14735784.2019.1667255.

Cai, K. and I. Martin (2024) 'How Stability AI's Founder Tanked His Billion-Dollar Startup', *Forbes*, 29 March. Available at: https://www.forbes.com/sites/kenrickcai/2024/03/29/how-stability-ais-founder-tanked-his-billion-dollar-startup/ (Accessed: 5 April 2024).

Chávez Heras, D. and T. Blanke (2021) 'On Machine Vision and Photographic Imagination', *AI and Society* 36: 1153-1165.

Cobb, M. (2020) *The Idea of the Brain: The Past and Future of Neuroscience*. London: Profile Books, Kindle edition.

Cole, D. (2023) 'The Chinese Room Argument', *The Stanford Encyclopedia of Philosophy*, E. N. Zalta and U. Nodelman (eds.). Available at: https://plato.stanford.edu/archives/sum2023/entries/chinese-room/ (Accessed: 10 April 2024).

Crawford, K. (2021) *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.

Danziger, S., J. Levav, and L. Avnaim-Pesso (2011) 'Extraneous Factors in Judicial Decisions', *PNAS* 108(17): 6889-6892. https://doi.org/10.1073/pnas.1018033108.

Daston, L. and P. Galison (2010) *Objectivity*. New York: Zone Books.

Dewdney, A. (2021) *Forget Photography*. London: Goldsmiths Press.

Dobson, J. E. (2023) *The Birth of Computer Vision*. Minneapolis: University of Minnesota Press.

Fausto-Sterling, A. (2000) *Sexing the Body: Gender Politics and the Construction of Sexuality*. New York: Basic Books.

Flusser, V. (2002a) 'Habit: The True Aesthetic Criterion', in *Writings*, ed. A. Ströhl, trans. E. Eisel. Minneapolis: University of Minnesota Press, pp.51-57.

Flusser, V. (2002b) 'On the Theory of Communication', in *Writings*, ed. A. Ströhl, trans. E. Eisel. Minneapolis: University of Minnesota Press, pp.8-20.

Flusser, V. (2002c) 'Photography and History', in *Writings*, ed. A. Ströhl, trans. E. Eisel. Minneapolis: University of Minnesota Press, pp.126-132.

Flusser, V. (2011) *Into the Universe of Technical Images*. Minneapolis: University of Minnesota Press.

Golumbia, D. (2022) 'ChatGPT Should Not Exist', *Medium*, 14 December. Available at: https://davidgolumbia.medium.com/chatgpt-should-not-exist-aab0867abace (Accessed: 1 February 2024).

Ho, J., A. Jain, and P. Abbee (2020) 'Denoising Diffusion Probabilistic Models', pp.1-25, arXiv preprint arxiv:2006.11239.

Malevé, N. and K. Sluis (2023) 'The Photographic Pipeline of Machine Vision; or, Machine Vision's Latent Photographic Theory', *Critical AI* 1(1-2). DOI: https://doi.org/10.1215/2834703X-10734066

Marx, P. (2023) 'Generative AI Closes Off a Better Future', *Disconnect*, 1 September. Available at: https://www.disconnect.blog/p/generative-ai-closes-off-a-better (Accessed: 1 February 2024).

Mitchell, M. and D. C. Krakauer (2023) 'The Debate over Understanding in AI's Large Language Models', *PNAS* 120(13): e2215907120. https://doi.org/10.1073/pnas.2215907120

Moholy-Nagy, L. (2003 [1936]) 'A New Instrument of Vision', in L. Wells (ed.) *The Photography Reader*. London and New York: Routledge, pp.92-96.

Noble, S. (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

Offert, F. and T. Phan (2022) 'A Sign That Spells: DALL·E 2, Invisual Images and The Racial Politics of Feature Space', *arXiv*, arXiv:2211.06323.

Paglen, T. (2016) 'Invisible Images (Your Pictures Are Looking at You)', *New Inquiry*, 8 December. Available at: https://thenewinquiry.com/invisible-images-your-pictures-are-looking-at-you/ (Accessed: 8 June 2018).

Salvaggio, E. (2022) 'How to Read an Image: The Datafication of a Kiss', *Cybernetic Forests Substack*, 2 October. Available at: https://cyberneticforests.substack.com/p/how-to-read-an-ai-image (Accessed: 10 January 2024).

Searle, J. R. (1980) 'Minds, Brains, and Programs', *Behavioral and Brain Sciences* 3(3): 417-457. [Preprint, accessed via Cogprints, 1-19.]

Sekula, A. (1975) 'On the Invention of Photographic Meaning', *Artforum* 13(5): 36-45.

Singer, P. (1997) 'The Drowning Child and the Expanding Circle', *The New Internationalist*, 5 April. Available at: https://newint.org/features/1997/04/05/peter-singer-drowning-child-new-internationalist (Accessed: 10 March 2024).

Sontag, S. (2005 [1973]) *On Photography*. New York: Rosetta Books with Farrar, Straus & Giroux.

Toister, Y. (2020) *Photography from the Turin Shroud to the Turing Machine*. Bristol: Intellect.

Wasielewski, A. (2023a) 'Authenticity and the Poor Image in the Age of Deep Learning', *photographies* 16(2): 191-210. DOI: 10.1080/17540763.2023.2189158.

Wasielewski, A. (2023b) '"Midjourney Can't Count": Questions of Representation and Meaning for Text-to-Image Generators', *Image: The Interdiscipinary Journal of Image Sciences* 37(1): 71-82. DOI: 10.1453/1614-0885-1-2023-15454.

Wells, L. (2003) 'Photographic Seeing', in L. Wells (ed.) *The Photography Reader*. London and New York: Routledge, pp.82-85.

Wolfe, C. (2009) *What Is Posthumanism?* Minneapolis: University of Minnesota Press.

Zylinska, J. (2017) *Nonhuman Photography*. Cambridge: MIT Press.

Zylinska, J. (2020) *AI Art: Machine Visions and Warped Dreams*. London: Open Humanities Press.

Zylinska, J. (2023) *The Perception Machine: Our Photographic Future between the Eye and AI*. Cambridge: MIT Press.

## Notes

[1] The argument about the interlocking of perception and vision was developed in *The Perception Machine* (Zylinska, 2023); this article is a follow-up to the book, taking the argument into some new technical

and conceptual territories. We need to be mindful of the fact that in cognitive psychology seeing has traditionally been defined as a physiological process that happens *to* the subject, with perception considered a more complex and active form of engagement with the world on the part of the subject, involving interpretation of the received data. John Berger's *Ways of Seeing* (1972) – a key text for humanities scholars studying visual processes – argued for the inextricable *interlocking* of these processes and hence against the idea of a primary 'way of seeing' that precedes culture. This article follows Berger's line of thinking by positing that seeing is already a form of perception because processes of data analysis and interpretation are mobilised from the outset across the perceptual matrix that contains not only the subject's visual and cognitive apparatus but also a whole array of agents, connections and institutions outside the subject's corpus – a state of events that applies to both humans *and* machines (see Azar et al., 2021).

[2] In many such models diffusion works in tandem with transformers and GANs (e.g., through networks such as VQGAN and CLIP). GANs (Generative Adversarial Networks) are machine learning models whereby two neural networks are put in competition with one another to produce an 'original' output.

[3] Malevé and Sluis explain: "While rarely described as such, machine vision has historically relied on an array of *photographic practices* (e.g., composing, capturing, labeling, and categorizing photographic images) and has engineered complex *curatorial pipelines* that translate the labor of millions of photographers and perceiving subjects into datasets" (2023).

[4] "Too much detail might make each image too different from one another and therefore not sufficiently generalizable. For this reason, downsampled images often perform better in classification tasks than highly detailed images" (Wasielewski, 2023a: 198).

[5] In the case of the LAION dataset that was used to train Stable Diffusion, humans were also involved in the assignation of so-called 'predicted attributes' (Malevé & Sluis, 2023; see also Baio, 2022), such as an aesthetic score or the likelihood of the presence of a watermark, to each image.

[6] This line of thought mirrors philosopher John Searle's 'Chinese room argument' proposed in 1980. Searle envisaged a scenario where, sitting in a closed room, he had been given a batch of writing in Chinese, followed by a set of instructions (a version of a computer programme), in English, about how to correlate certain Chinese symbols and shapes. This allowed Searle, who was not a Chinese speaker, to produce plausible responses to conversational prompts appearing under the room's door, in Chinese, without knowing what he was saying. The situation nevertheless deceived the human observer on the other side of the door into thinking that whoever they were communicating with – in this case, Searle, who described himself as "simply an instantiation of the computer program" (1980: 3) – 'truly' understood and spoke Chinese.

[7] This is why Peter Singer's utilitarian parable about our responsibility towards the drowning child (1997), like many test-case scenarios from analytical philosophy based on abstracted case studies, does not quite work.

[8] In the spirit of remix that characterises his work, Amerika is citing here from his earlier book, *META/DATA: A Digital Poetics*.

[9] As well as foregrounding the conceptual instability of the term, the hyphenated spelling is aimed to differentiate the position outlined here from the AI porn generator that goes by the name 'Unstable Diffusion'.

[10] The original phrase comes from Karl Marx and Friedrich Engels' *The Communist Manifesto*.

[11] Kyle Booten explains: "Maxwell's demon is a kind of algorithm or 'bot' that resists entropy within a physical, thermodynamic system. Flusser's two-fold intuition is that 1) critics are also, in a similar way, demonic filters that resist entropy in cultural systems and that 2) this function could be automated by 'automatic critics' that filter texts based on some linguistic criterion" (2020: 4).

[12] The fact that, by early 2024, Stability AI had run out of its large reserves, was unable to secure enough additional funding, "had defaulted on payments to Amazon whose cloud service undergirded Stability's core offerings" (Cai and Martin, 2024), had lost most of its research team and had removed its CEO founder Emad Mostaque is illustrative of the wider issues underpinning the product modelling a new future. Promising to "confront great adversaries, cancer, autism, and the sands of time itself", Stable Diffusion ended up "in a deep hole", unsure how to crawl out of it (Cai and Martin, 2024).

[13] Batchen (1997) attributes the explosion of multiple photographic inventions at the close of the nineteenth century to the existence of a 'desire' for photography. This desire was evident in the widespread interest in the ability to preserve images and was also fuelled by the sufficient development of the technological infrastructure that made those photographic inventions possible.

**Joanna Zylinska** is a writer, artist, curator and Professor of Media Philosophy + Critical Digital Practice at King's College London. She is the author of a number of books, including *The Perception Machine: Our Photographic Future Between the Eye and AI* (MIT Press, 2023, open access), *AI Art: Machine Visions and Warped Dreams* (Open Humanities Press, 2020) and *Nonhuman Photography* (MIT Press, 2017). Her art practice involves experimenting with different kinds of image-based media. She is currently researching perception and cognition as boundary zones between human and machine intelligence, while trying to map out scenarios for alternative futures.

**Email:** joanna.zylinska@kcl.ac.uk