IMAGE
The Interdisciplinary Journal of Image Sciences
37(1), 2023, S. 71-82
ISSN 1614-0885
DOI: 10.1453/1614-0885-1-2023-15454

HERBERT VON HALEM VERLAG

Amanda Wasielewski

"Midjourney Can't Count": Questions of Representation and Meaning for Text-to-Image Generators

Abstract: Text-to-image generation tools, such as DALL·E, Midjourney, and Stable Diffusion, were released to the public in 2022. In their wake, communities of artists and amateurs sprang up to share prompts and images created with the help of these tools. This essay investigates two of the common quirks or issues that arise for users of these image generation platforms: the problem of representing human hands and the attendant issue of generating the desired number of *any* object or appendage. First, I address the issue that image generators have with generating normative human hands and how DALL·E has tried to correct this issue by *only* providing generations of normative human hands, even when a prompt asks for a different configuration. Secondly, I address how this hand problem is part of a larger issue in these systems where they are unable to count or reproduce the desired number of objects in a particular image, even when explicitly prompted to do so. This essay ultimately argues that these common issues indicate a deeper conundrum for large AI models: the problem of representation and the creation of meaning.

Introduction

In early 2022, generative AI went mainstream. Many of the tools that became available over the course of the year were designed to bring AI capabilities to the masses, allowing just about anyone to generate text, images, or sound in multimodal ways. Around half a dozen image generation tools based on diffusion models were released to the public over the course of the year and they have already shaken the foundations of legal systems, business, artmaking, and politics. Although other AI image generation techniques, including GANs (generative adversarial networks), have received attention in the media in recent years, generative AI was still rather niche before 2022 and its implementation was

mostly confined to a tech savvy user base (cf. DICKSON 2020; HILL/WHITE 2020; RAYMOND 2021). In contrast, when DALL·E 2 was released to a limited audience in February 2022, it caused a media frenzy. Even though it took several months of closed beta testing for full-scale text-to-image generators like Midjourney, DALL·E 2, and Stable Diffusion to be released to the general public, less sophisticated copycat generators such as DALL·E Mini (later renamed Craiyon) were available early on. This led to an explosion of AI-generated images on social media. Suddenly, the public was not only aware that deep learning could be used to create and manipulate images; they were using it themselves.

Months before I began buying credit for or subscribing to text-to-image generator services, I lurked in online AI artist communities on Facebook, Reddit, and elsewhere that had early access to DALL·E and other generators. These groups were created by and for people who wanted to share tips for prompt writing and to exchange the output images they had created. Through this kind of informal ethnography, I began collecting posts and replies about the everyday uses of text-to-image generators that pointed toward greater underlying issues. After the wider release of DALL·E and Midjourney, I continued following these groups. My growing collection of posts has highlighted some common quirks in this type of technology that are worth deeper theoretical reflection. The following text addresses some of my early thoughts on this topic.

"Show Me her Hands!"

In a post on one of the AI artist communities I follow,[1] a user put up an AI-generated image of a young woman pictured in medium close-up, rendered in a photorealistic manner. This is a common genre for posts on such communities, i.e., showing off a particularly impressive creation for affirmation and applause. (Young, attractive women are *also* a common genre, but that is another story.) In the replies to the post, someone joked: "Very nice… but show me her hands!" The 'hands problem' is perhaps the most well-known failing of text-to-image generators, which struggle to render human hands with a sum total of five fingers that appear proportional and in naturally-occurring configurations. I am being careful not to characterize this as a failure to produce 'normal' hands. While five fingers in particular proportions may be the medically-defined norm, there are many people who are, of course, born with different numbers or configurations of digits/bones, or may have lost fingers/parts of their hand, or had them altered by events later in life. Nevertheless, one could say that AI-generated images often depict the human body, particularly hands and fingers, in ways that are

1 I was unable to find this post again in researching the present essay.

completely fantastical. Sometimes those fingers are long and stretched out, blending into the fabric of clothing or other body parts. Sometimes they appear more similar to toes (cf. fig. 1). Sometimes they are discontinuous blobs separated from the rest of the body. Often there are simply far too many fingers – sometimes dozens of fingers!

Figure 1: An absurd image of hand-toe-finger amalgams created from the prompt "Children's hands reaching for candy" with Stable Diffusion, January 2023

DALL·E 2 seems to have made an attempt to correct the 'hands problem' by forcing most of the hands depicted in its output images to have five fingers and *only* five fingers.[2] This would have been a smart – albeit somewhat inelegant – solution if either (a) no deviation from this norm existed or (b) no one would ever want to create an image containing a non-normative human hand. I first became aware of DALL·E's solution to the hand problem from a post where the prompt was "a hand with six fingers" (cf. BEERI 2023) and three out of four of the images showed five-fingered hands. I decided to try some prompts of my own.

When using the prompts "a hand missing a finger" or "a hand missing one finger", I found that the output images were not what I imagined either when writing those prompts. Instead, the eight images produced could be characterized as maliciously compliant. In other words, DALL·E gave me exactly what I asked for but not in the way I imagined (cf. fig. 2-4). One image appears with a finger that is literally missing, i.e., it looks like the finger was photoshopped out and the two ends of the hand were stitched and blended together (cf. fig. 2). Four of the images show a pointing index finger. In two of these, the finger is depicted either too large or too small in proportion to the rest of the hand (cf. fig. 3). The folded knuckles of the hands may be a way to interpret "missing" in this case. Another

2 As this essay was proofed, Midjourney v.5 was released and it seems to have also addressed/mostly fixed the hand problem.

two images show the frame of the image cropped so that only a sliver of the fifth finger is depicted in the image but, we can imagine, may still exist outside the boundaries of the frame (cf. fig. 4). A finger was missing from the image but not *missing*. I realized that my use of the term "missing" was not only difficult to interpret but also unwittingly biased. Was DALL·E pointing out my ableist characterization of non-normative limbs?

Figure 2: Image of what seems to be an awkwardly removed finger created from the prompt "A hand missing one finger" with DALL·E 2, February 2023

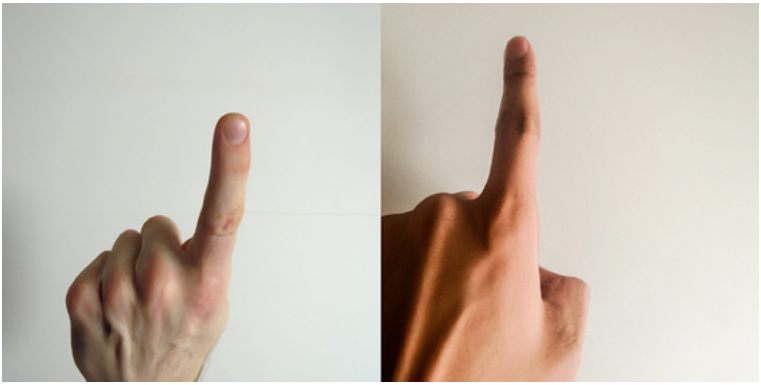

Figure 3: Images of two pointing index fingers created from the prompt "A hand missing a finger" with DALL·E 2, February 2023

Figure 4: Image of a hand with a 'missing' finger, i.e., a finger we can imagine as being just out of frame, created from the prompt "A hand missing one finger" with DALL·E 2, February 2023

I adjusted my prompt to simply ask for "A hand with four fingers". Once again, three of the four images generated showed five-fingered hands (cf. fig. 5). All the images depict the thumb folded into the palm and one appears to show the pinky finger also folded in. The fourth image does show a hand with four fingers but, again, the palm appears to have been shortened in order to omit one of the fingers (cf. fig. 6). DALL·E still does not seem to understand what I am getting at here.

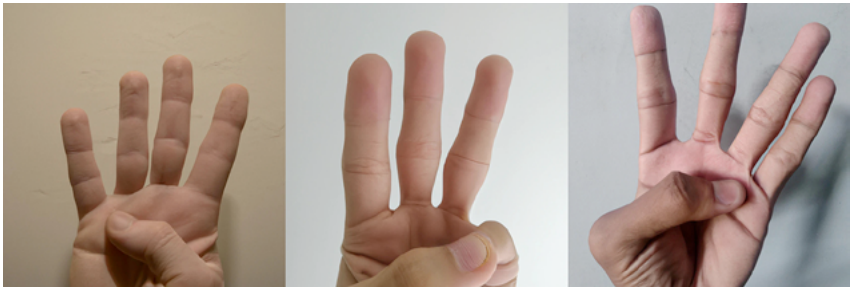

Figure 5: Images created for the prompt "A hand with four fingers" by DALL·E 2, February 2023. Curiously, all three show, in fact, a hand with five fingers

Figure 6: Image created for the prompt "A hand with four fingers" by DALL·E 2, February 2023. This was the only image of the four created in total for said prompt which actually had four fingers

This lack of understanding is not that surprising, however, if one considers the possible training data behind the system. For example, when I search for "a hand with four fingers" on Google image search, the majority of the images that come up are similar to the DALL·E output: they show hands holding up four fingers with their thumb folded inward. The semantic construction indicates something to me that is different from what it calls up for the interpretative machine. Given the nature of a regular online search, I do not expect Google to produce the exact (type of) images I ask for. If I search for an image of a hand with four fingers and I do not get an image that looks exactly like what I hoped it would, as is the case here, I do not automatically conclude that Google has failed. After all, you cannot seek what is not there to find. Search implies that we are sifting through existing things.

As a user, however, I expect DALL·E to conjure something that is *not* there to find, even if the reality is that Google image search and DALL·E are both drawing from bodies of *existing* information, i.e., data that connects text to images. In simple terms, this has to do with how these tools have been marketed and promoted to the public. OpenAI, the company behind DALL·E, and others hyped the technology's ability to construct scenes with impossible or fantastic juxtapositions, such as an astronaut riding a horse on the moon. One might wonder, if DALL·E can do something outlandish like this, why does it struggle with simple requests for a certain number of fingers? The comparison between the 'search' query and the 'prompt' query, however, has deeper implications for users of AI, particularly as search engines like Bing are rolling out AI chatbots to assist with search functionality.

In my work on this topic, I often refer to targeted prompt-writing as a way to 'query the database', meaning that I am doing a kind of search of terms that might be connected to certain imagery and drawing conclusions based on

whether they 'come up' in the resulting image. The difference between searching and prompt-writing nowadays seems to be related to the user's expectations. The public-facing AI tools that have been launched over the past year are marketed as near-magical experiences, i.e., *intelligent* machines that help generate text or images. Google and other search engine algorithms have been using machine learning to optimize search functionality for many years, yet few people expect Google to read their minds when they query a simple search (indeed, many people would rather it not).

Perhaps the novelty and 'magic' of prompting will wear off and we will learn to expect as little (or as much) from prompts as we do from a search. For now, however, it is worthwhile to put prompts into perspective and temper our expectations of their efficacy. It's software, not magic. Exercises such as the one above begin to explore the boundaries and limits of AI tools, albeit in a non-systematic way. They also hint at the ways text-to-image generators may replicate highly biased notions of 'normality' vis-a-vis statistical sampling. In addition to addressing the hand problem, OpenAI has also quietly addressed issues around the ethnic and racial diversity of the people depicted in output images of DALL·E. For example, whereas earlier versions of DALL·E might have shown only white men as CEOs, it now generates a diverse collection of people if given the general prompt "the CEO of a company" (cf. fig. 7), although it does so by editing user inputs by adding certain words before passing them on to the generative AI (cf. OFFERT/PHAN 2002: 2).

Figure 7: Images created for the prompt "the CEO of a company" by DALL·E 2, March 2023

Perhaps someday soon there will be a more elegant solution to the hand problem. However, hands and fingers are simply the most obvious sign of a larger underlying problem for text-to-image generators: counting.

"Why Can't Mj Count?"

Text-to-image generators not only have trouble knowing how many fingers to give a person but also how many of anything to give to anything, even when the prompt explicitly specifies a number. To return to the question of why DALL·E can generate an image of an astronaut riding a horse on the moon but not (reliably generate) a four-fingered hand, the answer has to do with numbers and counting in general. Another recent post, this time on a Midjourney community on Facebook (cf. REYNERI 2023), asked the group why they were unable to generate an image of a "five-story apartment building" despite specifying the number of floors using a variety of different terms. They were frustrated because, over and over again, the images generated showed *eight* to *nine* floors. In response, a familiar chorus of replies flooded in: "Mj can't count". A few months earlier, a user in the group named Steve Laredo (2022) had directly posed this question to the group, "Why can't Mj count? There must be a computer science reason? Anyone?" Very few of the replies were able to directly answer the question, but many attributed Midjourney's lack of counting abilities to its basis in deep learning. Its functionalities were not, they explained, explicitly programmed to do specific things but rather acquired. So, they said, it simply *did not learn* to count. More pragmatically-minded replies, meanwhile, dismissed the issue as a temporary glitch that would be worked out in time. I would posit, however, that the counting problem is something more fundamental to text-to-image generators. It is essentially a *representation* problem.

The aforementioned issues with diversity in output images and the subsequent effort to make DALL·E images more racially and ethnically diverse boil down to the bias of its training data (and, of course, the bias of society at large). There were simply more images in the training data that labeled *white* men as CEOs and the early output of DALL·E reflected this. The counting problem, however, is not related to the training data. It is not even necessarily an issue of semantics or the connection between text and images. The counting problem has to do with our understanding of images as representations. 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. The human viewers of AI-generated images, on the other hand, are likely to have had some earlier experiences of physical people, places, and things that are much like those that are depicted in AI-generated images. How else could we recognize the subject of these images? While we may not have had direct in-person experiences of some rarer things, we also understand those things in a more nuanced way than AI tools do, through contextual information we might read or hear about. Human viewers will thus 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). For example, it is likely that every person on this planet has an experience of interacting with human hands in physical spaces – both their own and other people's – whereas DALL·E has only experienced human hands through visual representations, i.e., patterns of pixels that have been categorized as "hands".

One of the replies to Laredo's (2022) post in the Midjourney community from another group member named Rachel Aanstad touches on this: "Because [Midjourney] understands surface better than form. It has used 2D images to train and doesn't have a concept of 3D space like we do. It lives in flatland. It gives us layers not volume and doesn't understand how bodies are formed". Midjourney 'understands' that certain collections of pixels in an image can be categorized as "dog" or "tree" but it does not really know what a dog or a tree are (cf. WASIELEWSKI 2023: 93). This is an example of computational formalism, where a visual representation is assumed to provide enough information on the nature of the thing represented. These reflections on meaning and form, in turn, echo the arguments of Emily M. Bender and Alexander Koller (cf. 2020). They address the question of whether large language models can create meaning or 'understand' language, arguing that language models "trained purely on form will not learn meaning" (BENDER/KOLLER 2020: 5187). The purpose of language, they contend, is "communicative intent", which is "about something outside of language" (BENDER/KOLLER 2020: 5187). They propose a thought experiment they call the "octopus test" (BENDER/KOLLER 2020: 5188), where an octopus deep in the ocean (the stand-in for large language models) is able to intercept the communications between two humans and learn to predict their likely responses based on statistical samplings. They argue that the octopus may convince one of the humans that it is the other human by mimicking their responses but "has never observed these objects [to which it refers], and thus would not be able to pick out the referent of a word when presented with a set of (physical) alternatives" (BENDER/KOLLER 2020: 5188).

At first glance, multimodal models may seem different. After all, text-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 *understands* what that image actually is or what it represents. Like any type of symbol, digital images – even digital photographs – are representations of things that have a meaning superseding their visual form. In another, now infamous article, which led to the high-profile firing of researchers Timnit Gebru and Margaret Mitchell from Google (cf. SIMONITE 2021) and which was co-authored by Bender and Angelia McMillan-Major, the authors describe large language models as "stochastic parrots" (BENDER et al. 2021: 610), meaning that they are very good – uncannily good – at mimicking language but have no idea what they are actually saying.

We could say the same thing about text-to-image generators. 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.

Form can be defined as the visual and the material properties of an image or object. However, neither the surface appearance nor the three-dimensional volume of an object can produce meaning on its own. Rather, form is the site or the locus of context and experience. As David Summers asserts, this has to do with the real space forms inhabit: "uniformities arise because images are always embodied and share real space with those who see and use them" (SUMMERS 1989: 405). A human viewer will likely be aware that their experience of an object is mediated by, for example, a photograph, and that this photograph has its own form and its own properties that are separate from those of the objects or scene depicted. In other words, most humans understand that the photograph of the dog is not the dog itself. Alternatively, a viewer may understand a particular form through social interactions and human intermediaries. They have had interactions with a dog, perhaps, or are aware, through life experience, of the many ways dogs and humans coexist in the world. Image generators, however, do not process images within a framework that accounts for or uses such mediations. Instead, they must produce images based on relationships between representational forms, which have been concretely defined. Very little if any consideration is given to real space in such constructs.

Conclusion

In this essay, the phenomena I have labeled 'the hand problem' and 'the counting problem' for text-to-image generators are ultimately both issues of meaning and representation. The output images of tools like DALL·E and Midjourney are discrete visual forms based on statistical samplings. Despite the particularity of their appearance, they represent data in the plural form. In most traditional image-creation processes, representational images refer to a single entity contained within the confines of the image. Text-to-image generators need to be understood as a very different form of representation, despite their superficial, perhaps even uncanny similarity to images generated by other means. Right now, this technology is still very new. As we get more familiar with it, some of its

magic will likely wear off and it will become just another tool in the arsenal of digital imaging software. While it is still fresh, though, it is worthwhile thinking about the ways in which its early quirks define it as a creative practice.

Bibliography

BEERI, IDO: Prompt: A Hand with Six Fingers. Post on *Facebook*. January 19, 2023. https://www.facebook.com/groups/dalle2.art/permalink/706084514419575/ [accessed March 1, 2023]

BENDER, EMILY M.; ALEXANDER KOLLER: Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 5185-5198.

BENDER, EMILY M.; TIMNIT GEBRU; ANGELINA MCMILLAN-MAJOR; SCHMARGARET SCHMITCHELL: On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 610-623

DICKSON, E.J.: TikTok Stars are being Turned into Deepfake Porn Without their Consent. In: *Rolling Stone*. October 27, 2020. https://www.rollingstone.com/culture/culture-features/tiktok-creators-deepfake-pornography-discord-pornhub-1078859/ [accessed March 1, 2023]

HILL, KASHMIR; JEREMY WHITE: Designed to Deceive: Do these People Look Real to you? In: *The New York Times*. November 21, 2020. https://www.nytimes.com/interactive/2020/11/21/science/artificial-intelligence-fake-people-faces.html [accessed March 1, 2023]

LAREDO, STEVE: Why can't Mj Count? Post on *Facebook.* December 27, 2022. https://www.facebook.com/groups/officialmidjourney/permalink/4808534808728888/ [accessed March 1, 2023]

OFFERT, FABIAN; THAO PHAN: A SIGN THAT SPELLS: DALL-E 2, Invisual Images and the Racial Politics of Feature Space. *arXiv:2211.06323*. October 26, 2022. https://arxiv.org/abs/2211.06323 [accessed March 1, 2023]

RAYMOND, SHANE: Deepfake Anyone? AI Synthetic Media Tech Enters Perilous Phase. In: *Reuters*. October 27, 2020. https://www.reuters.com/technology/deepfake-anyone-ai-synthetic-media-tech-enters-perilous-phase-2021-12-13/ [accessed March 1, 2023]

REYNERI, FEDERICO: Dear All, is there a Way to Let MJ Know what a Storey (of Floorplan) is. Post on *Facebook*. January 21, 2023. https://www.facebook.com/groups/officialmidjourney/permalink/5205677035681 32/ [accessed March 1, 2023]

SIMONITE, TOM: What Really Happened when Google Ousted Timnit Gebru. In: *Wired*. June 8, 2021. https://www.wired.com/story/google-timnit-gebru-AI-what-really-happened/ [accessed March 1, 2023]

SUMMERS, DAVID: 'Form', Nineteenth-Century Metaphysics, and the Problem of Art Historical Description. In: *Critical Inquiry,* 15(2), January 1989, pp. 372-406
WASIELEWSKI, AMANDA: *Computational Formalism: Art History and Machine Learning*. Cambridge, MA [MIT Press] 2023