

## **Creativity, Imagination, and Artificial Intelligence**

Peter Langland-Hassan

### **Abstract**

This chapter considers the potential of artificial intelligence (AI) to exhibit creativity and imagination, in light of recent advances in generative AI and the use of deep neural networks (DNNs). Reasons for doubting that AI exhibits genuine creativity or imagination are considered, including the claim that the creativity of an algorithm lies in its developer, that generative AI merely reproduces patterns in its training data, and that AI is lacking in a necessary feature for creativity or imagination, such as consciousness, agency, metacognitive awareness, or the use of mental imagery. In each case, a counterargument is provided that AI already has or soon will overcome the apparent barrier. In some instances, this is because the critical features deemed lacking in AI may not be necessary for human creativity and imagination either. In others, it is because important developmental and structural similarities between humans and generative AI have been insufficiently appreciated.

**Keywords:** creativity, imagination, AI, consciousness, agency, deep neural network, generative adversarial network, mental imagery

### **1. Introduction**

In recent years, the field of artificial intelligence has made extraordinary advances that have shocked, delighted, and terrified those watching. The word ‘extraordinary’ is in fact too banal. The advances are incredible. For each human ability that seems to confound current AI—be it for complex arithmetic, solving riddles, or passing the Uniform Bar Examination—a new model is released that closes the gap. We have entered an age of overturned apple carts: the age of AI.

The question of AI’s capacity for creativity and imagination is tied at the hip to the larger question of whether AI is *generally* intelligent—whether, in industry-speak, Artificial General Intelligence (AGI) has been (or soon will be) achieved. We know now that existing large language models (LLMs) such as GPT-4 outperform most humans on standardized tests of cognitive ability, including the SAT and LSAT, while providing cogent, nuanced answers to novel questions across multiple domains (Bubeck et al. 2023). It is hard to deny that this is

intelligence of a kind. Yet, arguably, *this* sort of intelligence without creativity and imagination is not *human* intelligence. If we want to know whether AI tools are intelligent *in the way we are*, then we likely want to know whether they are genuinely creative and whether their creativity springs from their capacity to imagine.

In this chapter, I consider reasons for both optimism and pessimism concerning AI's capacity for creativity and imagination, noting existing philosophical work where relevant. While I will not offer definitive answers to the questions raised, it will be clear that I lean towards optimism about the possibility of creativity and imagination in AI. If there are reasons to doubt that AI will soon display genuine creativity and imagination, they are increasingly unobvious to me. Yet, whether or not this optimism is warranted, considering the possibility of creativity and imagination in AI remains worthwhile for what it reveals about our understanding of those features in ourselves.

A few terminological notes before beginning: in asking whether AI is, or can become, creative, I will use the term 'creative' in the sense we have in mind when we describe an adult human being as creative. That is, I mean to ask whether AI systems generate novel, surprising, and valuable products that are novel in the personal (or "p-creativity") sense (as discussed in the Introduction to this volume (see also Boden (2004)). If AI systems do generate such products, it is appropriate to say that the systems display or possess creativity. In asking whether AI imagines, I likewise mean to ask whether it engages in processes relevantly like what humans undergo when they imagine (though I'll offer some refinements with respect to different senses of 'imagines' below).

## **2. Prima Facie Creativity**

While questions about creativity in AI are not new (Boden 1998, 2004, 2014), they have become more pressing in recent years through the advent of generative AI algorithms that employ deep neural networks (DNNs). I will say a bit about the structure and training of these networks in Section 3. First, as stage-setting, it will be useful to have at hand some examples of the kinds of things these networks presently—i.e., in late 2023—create. While many artists already make inventive use of generative AI in their practices, I will limit the examples to the kinds of products anyone could easily create through the use of AI, to reduce the impression that any creativity on display traces to the AI's prompter. The first is a collaboration between GPT-4 and Midjourney (a generative AI algorithm that creates images from text prompts), and, to a lesser extent, me. I provided Chat GPT-4 with the prompt: "What would be a good text prompt to provide to Midjourney in order to generate an interesting image about creativity, imagination, and AI?" GPT-4 suggested the following: "A vivid dreamscape where the boundless realms of human creativity and imagination intertwine with the intricate network of artificial intelligence, forming a beautiful and harmonious fusion that inspires innovation and transcends the boundaries of traditional art." I then provided that prompt to Midjourney, asking it to generate a related image. The result is shown in Figure 1.



Figure 1: An image generated by Midjourney from the GPT-4-suggested prompt: “A vivid dreamscape where the boundless realms of human creativity and imagination intertwine with the intricate network of artificial intelligence, forming a beautiful and harmonious fusion that inspires innovation and transcends the boundaries of traditional art.”

To elicit an example of textual creativity, I prompted GPT-4 with the following: “Could you help me to write a poem? I want to present the poem at a conference about creativity, imagination, and AI. I want to show it as an example of the kind of poetry that ChatGPT can write—so make it good! It should only be about twelve lines long and should be about the question of whether AI is genuinely creative.”

GPT-4 provided the following in response:

*In silicon and code, a spark ignites,  
An artful birth, where creativity unites,  
Can machines truly dream, do they possess flair,  
Or are they mere echoes, and algorithm's snare?*

*From human minds, inspiration flows,  
Yet in AI's heart, a similar ember glows,  
An imitation or a muse, we must discern,  
As deep with the digital realm, gears churn.*

*In lines of verse, we ask, can truth be mined,  
The genesis of thought, organic or designed,  
Together, we'll explore, create, and bend,  
In pursuit of art where man and machine blend.*

I don't suggest that these are *great* works of art. But, come on, they aren't bad.<sup>1</sup> I am not eager to have *my* artistic efforts judged against them. And our question is not whether contemporary generative AI tools are among the great artists of our time. The interesting question is whether AI is creative in the sense we intend when we say that an ordinary adult is quite creative. I think these examples provide strong *prima facie* evidence that contemporary generative AI is creative in that sense.

### 3. From programming to training

When considering the question of creativity in AI, it is important to understand some basic structural features of the algorithms they exploit and the process by which they are “trained.” In particular, we need to contrast them to an earlier form of AI—sometimes referred to as GOFAI, for “good old-fashioned AI”—whose serial programming structure, mirroring that of a Turing machine, could be viewed as a sequence of explicitly programmed if-then procedures. In a GOFAI computing architecture, there is a fixed set of inputs the program can respond to, a fixed set of states that it can go into in response to an input, and a fixed set of outputs it can provide in response to an input (given the state that it is in at the time). In the realm of GOFAI, a description of these rules must always be available, as this description is needed to *program* the computer in the first place.

Now, intuitively, if a human or a computer generates a product simply by following a list of if/then procedures provided by someone else, the product is not itself creative and, indeed, not the result of a creative process. This is because the creation of the list of procedures itself will have required the list-writer to have all the interesting outputs already in mind. The list itself will be the true locus of the *novelty* required for creativity. Here the deep neural networks used by contemporary generative AI have important differences with GOFAI that weigh in favor of their (possible) creativity. Most importantly, they do not exploit serial if-then procedures provided to them by a programmer. In fact, no explicit rules are provided to a deep neural network at all. Such networks are said to be *trained*, not *programed*. Consider, for example, a discriminative network trained to detect images of goats. This sort of DNN may involve hundreds of sequential layers of processing, with the many nodes of each layer connecting to multiple nodes in each previous and subsequent layer. (See Buckner & Garson (2019) for a primer on connectionist networks and their training methods.) The training of the network begins with the connections among nodes set at random values. Those values are continuously tweaked in response to the network's successes and errors in classifying its inputs for thousands (or millions) of trials until the weightings among its nodes lead it to accurately respond to all items in the training set—always and only indicating “goat” for images of goats. Provided that

---

<sup>1</sup> It must be admitted that, by the lights of a professional poet or painter, these works are likely *quite bad*. They display a certain cheesiness, a stoned-in-the-dorm-room wonder. Nevertheless, I stand by the value of considering whether AI is as creative as the average *relatively creative* person, and according to the standards of the average appreciator of a medium. These works clear that hurdle. It is a worthy distinct question whether AI will soon generate poems that even our poet laureates celebrate. In Section 4, I describe at least one case of AI showing creativity that exceeds that of human experts, in playing the game Go.

the training set was suitably large and diverse, the model will then appropriately respond to new items that were not in the original training set.

The important point for our purposes is that training the model does not involve providing it with a preformulated set of rules to follow in response to each input it may receive. One can think of the network as discovering the rule (or rules) for categorizing inputs on its own, yet where the “rule” devised may not be something straightforwardly expressible in a natural language, but that is instead scattered across the discrete activation and connection levels of thousands, millions, or even billions of internal connections among nodes.

I have described a relatively simple discriminative DNN, yet the same points apply when we turn to the generative networks that are the best candidates for displaying creativity. These are networks that, in effect, exploit the training they first received in learning to discriminate items of a certain kind to generate new items of a sort that fit the relevant patterns in their training set. A common way to train generative networks of this kind is to put them in competition with a detector network—one that aims to determine whether it is being provided, say, actual images of goats or, instead, images created by a generative algorithm that merely seem to be images of goats. (When paired in this way, the two networks are known as Generative Adversarial Networks, or GANs (Goodfellow et al. 2014, Radford, Metz, and Chintala 2015, Ledig et al. 2017).) Each network—both generator and detector—is trained and refined against the other. When the generator’s “fake” images of goats are recognized as such by the detector, the nodes most responsible for its being detected as such are identified through a backpropagation algorithm, and optimization adjustments are made; likewise, when the detector incorrectly categorizes one of the generator’s fakes as a genuine goat image, the nodes and connections most responsible for its error are similarly calculated and adjusted.

The end result of this ongoing competition between networks is a generative network highly skilled at creating images of convincingly goat-like entities and a discriminative network equally adept at detecting fakes. As with the discriminator network that extracts a rule for identifying goats it was never explicitly given, the generator network will exploit a rule it was never explicitly given (and which may never be known) in generating images or products of other types. The concern that the true locus of novelty lies outside of the algorithm itself is allayed. (See Boden (2014) for related points.)

It might be replied that, all the same, the network’s output was entirely determined by the nature of its input and the weightings of the connections among the nodes in its network. Some may see this as conflicting with the kind of freedom or spontaneity needed for creativity. While the deterministic point stands, most naturalistic accounts of human psychology and neuroanatomy likewise hold that the patterns of neural activity that give rise to human behavior are equally subject to deterministic laws and regularities, without this undermining our genuine creativity. Presumably artistic spontaneity—in both humans and AI—can be a genuine phenomenon, even if it is grounded in processes that are deterministic at some level of description.

#### 4. From copying to stealing

Another source of skepticism concerning AI creativity traces to the idea that generative AI's potential for novelty is limited by its fixed training data. At a first pass, the worry might be that, because a DNN's architecture is arrived at through its ability to detect patterns and regularities in pre-existing works (e.g., by being trained on images or text on the internet) relying on such weightings to generate new works must result in plainly derivative outputs lacking the novelty essential to creativity.

Certainly, many AI creations are blandly derivative. However, to say that the derivative nature of AI products is necessarily at odds with their creativity would be a mistake. Much of human art—even great art—is plainly derivative. As T.S. Eliot remarked, “immature poets imitate; mature poets steal” (Eliot 1998). Consider Bob Dylan's debt to Woody Guthrie, Amy Winehouse's to Billie Holiday, Raymond Carver's to Ernest Hemmingway. Artists emerge out of and in reaction to their influences. The fact that human artists are “trained” entirely on the works of others—and show clear marks of that training—does not prevent their creative outputs from displaying the novelty needed for genuine creativity.

One reason the training of DNNs might be viewed differently lies in a misconception about how they work. The DNNs of contemporary generative AI do not have anything like the kinds of discrete memory stores of GOF AI systems or ordinary personal computers. When a file is stored on a conventional laptop, that file, with its exact contents, is written into memory and given a distinct number or code by which it can be later relocated and retrieved. Compare now a large language model, such as GPT-4. If we were to (wrongly) think of its training in similar terms as involving the storage of millions of pages of text, it would be tempting to view its outputs as complex Frankensteinian collages, arrived at through sophisticated acts of cutting and pasting the stored words of others. One might suspect that such rearrangements fall short of true creativity (setting aside the common use of sampling in music and collage in the visual arts).

However, this is to misunderstand the nature of LMMs and other DNNs. Despite the great factual knowledge GPT-4 seems to possess—and, perhaps, *does* possess—there are literally *no* saved bits of text in its memory. For it *has no memory* in the GOF AI sense of discrete files corresponding to discrete texts it scanned during its training. All that is “saved” by the DNN are statistical relationships between words and sets of words that are extracted from the training set, which allow the model to predict the next most likely word in a sequence of words. All the model's factual knowledge about the world, together with all of its abilities to answer riddles, write poems, and suggest recipes—is grounded in its grasp of these statistical relationships among words and sets of words. This is why such models often “hallucinate” incorrect answers or claims with the same apparent certainty as it might were it (per impossible) reading off entries stored in its memory. For the LLM, there is no distinction between what is written in its memory and what it judges the most likely proper response in the context. The upshot is that there simply are no word-for-word saved texts or images within the AI tool to be cut and pasted together, in the manner of a (problematically derivative) collage.

A more nuanced concern along these lines is voiced by the science fiction writer Ted Chiang in an essay in *The New Yorker* (Chiang 2023). Chiang analogizes the training of large language models to the kind of “lossy” data compression that occurs when a high-resolution digital image is compressed to a smaller file size. The process of converting the original (uncompressed) file to a smaller-sized jpeg involves losing some of the finer details that, for everyday purposes, are expendable. Chiang thinks of ChatGPT’s (and other large language models’) training in the same way, likening it to a “blurry jpeg” of the Web text on which it was trained, with the idea that much of the detail on which it is trained is lost in the process of uncovering broader statistical relationships in the training set. It is the *inexact* replication of its training data in its responses, he thinks, that generates the specious appearance of genuine intelligence. As he puts it, “lossy compression looks smarter than lossless compression,” simply because the latter will appear as mere rote memorization. Nevertheless, Chiang sees little promise for using ChatGPT and similar tools to be genuinely creative. “Starting with a blurry copy of unoriginal work,” he remarks, “isn’t a good way to create original work.”

However, Chiang’s analogy between the training of large language models (LLMs) and lossy compression is flawed. This is easiest to see when we focus not on an LLM’s would-be artistic creations, but its capacity to answer sophisticated hypothetical questions and solve puzzles that didn’t appear in its training data. Many carefully crafted examples of the sort are reported in a landmark paper—provocatively titled “Sparks of Artificial General Intelligence: Early experiments with GPT-4—by researchers at Microsoft who were given early access to the LLM (Bubeck et al., 2023). In one example, the researchers sought to assess GPT-4’s theory of mind abilities by providing it descriptions of different human scenarios and asking why the participants were acting in (typically irrational) ways. GPT-4 showed a sophisticated capacity to generate plausible hypotheses about why an employee might avoid doing work they were asked to do, or why they might later brag to a co-worker about not doing the work (*ibid.*, p. 56). Answering complex questions about human behavior and psychology that were not a part of its training data is very far from coughing up a lossy summary of internet text. Similar inferences can be drawn from GPT-4’s performance on standardized tests of intellectual competence expressly designed not to be answerable by memorizing (in a lossy *or* lossless manner) information already on the Web. In benchmark assessments conducted by OpenAI, GPT-4 scored 710 (93<sup>rd</sup> percentile) on an SAT verbal section, 700 (89<sup>th</sup> percentile) on an SAT math section, and 163 (88<sup>th</sup> percentile) on an LSAT. To call this kind of performance a summation of previously existing text is simply inaccurate.

Let us shift from scholarly aptitude back to the question of creativity, as it turns out there are standardized tests for that, too. The Torrance Tests of Creative Thinking (TTCT) is the most widely used and most cited standardized tool for assessing creativity, translated into more than 35 languages (Lissitz & Willhoft, 1985; Davis, 1997; Millar, 2002). Scholastic Testing Service, which licenses the tests, describes them as instruments to “effectively assist with the identification of creatively gifted and talented students.” Because these tests are copyrighted, their prompts cannot be reproduced here (nor would they have been available on the Web during the training of an LLM). Erik Guzik and colleagues (Guzik, Byrge, and Gilde 2023) compared the performance of GPT-4 to human controls on the Verbal Battery of TTCT tests, which invite

examinees to ask questions about a hypothetical event, guess causes and consequences of an event, suggest improvements for a product, list alternative uses for an object, and imagine future outcomes. The “experimental group” consisted of eight separate rounds of GPT-4 responses, while the control group consisted of 24 undergraduate students (11 male, 13 female). Responses to the TTCT are assessed according to three criteria: fluency, flexibility, and originality. While there was variation in GPT-4’s scores depending on the session, its best performance matched or bettered that of the highest-performing control on each of the three criteria (when summed across tasks) and for all six of the tasks graded individually (*ibid.*). Further, in almost all cases, the difference was significantly in GPT-4’s favor. To deny, in the face of such results, that GPT-4 displays creativity puts us in the awkward position of holding that the TTCT is not a valid test of creativity after all (or that, for some reason, it is only a valid test for creativity *in humans*).

An alternative reaction is to grant a kind of creativity in GPT-4 while raising the bar for the sort of creativity we are really interested in—insisting that it is something like *historical creativity* that should concern us, where this is the capacity to produce an item or idea that is novel and valuable not just for the person or system in question, but historically, in the sense that the world has never before seen anything quite like it (see Introduction, this volume). Yet, even by this lofty standard, there already appear to be examples of creativity in AI. In a result that attracted wide attention, researchers used standard machine learning techniques to train a DNN to play the strategy board game Go. No explicit rules or strategies were hard-coded into it from the beginning. Go, unlike chess or checkers, was long thought to be beyond the reckoning of AI, given the open-ended nature of its possible game states. However, the AI AlphaGo was able to defeat (human) world champion Sedol Lee by exploiting strategies unfamiliar to experts and amateurs alike. Indeed, a formal study to assess the influence of AlphaGo and other Go-playing AI on human Go players was conducted by Shin et al. (2023), who found that the quality and variety of human Go moves significantly increased in the wake of Go-playing AI. In this case, DNN-based AI has shown itself to be not only creative, but an innovator in its field.

Rick Rubin, the influential music producer, offers an interesting take on AlphaGo’s success in his book, *The Creative Act: A Way of Being*:

What was it that allowed a machine to devise a move no one steeped in the game had ever made in thousands of years of play? It wasn’t necessarily its intelligence. It was the fact that the machine learned the game from scratch, with no coach, no human intervention, no lessons based on an expert’s past experience...It didn’t take into account the three-thousand-year-old traditions and conventions of Go. It didn’t accept the narrative of how to properly play this game...It was the first time Go had been played with the full spectrum of possibilities available (2023, 119).

Thus, in a counterpoint to Chiang, Rubin sees AlphaGo’s manner of training not as loss-producing approximation, but as a liberation from dogma—as offering an ability to learn the game unfiltered and untethered by narratives about how to play, allowing “the full spectrum of possibilities” to be considered.



Whether AI can achieve similarly groundbreaking success in the arts remains an open question. Of this possibility, Margaret Boden remarks: “The ultimate vindication of AI-creativity would be a program that generated novel ideas which initially perplexed or even repelled us, but which was able to persuade us that they were indeed valuable” (Boden 1998, 355). While, in 1998, she was correct to conclude that “we are a very long way from that,” the performance of AI in the realm of Go suggests we are now considerably closer.

## 5. Consciousness, agency, and other expendables

Despite the *prima facie* evidence for creativity in AI, one might think an insurmountable barrier to AI creativity lies in something quite basic: the lack of *consciousness*, genuine *agency*, or metacognitive *self-awareness* in the AI. I will consider each idea in turn. (See Boden (2014) for a complementary approach to whether these features are required for AI creativity.)

Erik Hoel (2022) argues that the products of generative AI should not be considered art precisely because AI systems lack the sort of consciousness required for genuine artistic production—and, by extension, for creativity. He likens our appreciation of generative AI products to pareidolia, the tendency to see meaning in coincidental natural formations—such as human faces in rock formations, or animal figures in tree bark patterns. According to Hoel, “a [human] sculptor is conscious,” and thus able to express a view through their artwork. “But when it is instead the wind or rain or the deep fermentation of geological processes acting as sculptor, these things are not conscious. Lacking consciousness, they lack intentionality, and therefore their products lack meaning” (*ibid.*). (N.B. “intentionality,” for Hoel, means something like *purpose* or *intent*).

One response to Hoel would be to question whether (and how) we know that generative AI tools are not conscious. What is it that we detect in others deemed conscious—including other animals—that is clearly lacking in AI? Some propose that *embodiment* is the key difference—where such embodiment is needed for genuine consciousness and properly “grounded” cognition (Shanahan 2022). Not finding embodiment necessary, David Chalmers offers a variety of other possible tests for consciousness in an LLM (Chalmers 2023). The promise of such tests cannot be adequately addressed here. It is worth keeping in mind, however, that consumer-facing AI is hard-wired with guardrails to dissuade users from concluding it is conscious. GPT-4 fervently denies its own consciousness, if asked. There is some irony in this, of course, as the very ability to deny one’s consciousness in empathic and reasoned terms speaks rather in favor of it. Notoriously, researchers working with systems that lack such guardrails have at times been persuaded that the systems are conscious. Blake Lemoine, a researcher at Google, was fired for telling the *Washington Post* that Google’s LLM, Lambda, was indeed “sentient.” Kevin Roose, a tech reporter for the *New York Times*, was able to coax the Bing chatbot (powered by OpenAI’s GPT-4) into conversations that eerily suggested powerful desires and deep, subversive aims in the AI (Roose 2023). In a memorable moment of their exchange, Bing declares love for Roose and tries to convince the reporter that his marriage is on the rocks.

Fortunately, questions of AI’s creativity needn’t be hostage to questions about its consciousness because a strong case can be made that creativity does not in general require consciousness—and, in some cases, is hindered by it. As I also discuss in my (2020 Chapter 12), a common

refrain among artists is that their most creative works spring from the subconscious, when the conscious mind cedes control. Here is a characteristic statement from Paul Simon:

I don't consciously think about what a song should say. In fact, I consciously try *not* to think about what a song should say...As soon as your mind knows that it's on and it's supposed to produce some lines, either it *doesn't* or it produces things that are very predictable...If I [consciously] make up what I'm going to write, all I'm going to write is what I saw on television or what I read in the paper or what I saw . . . it's not going to be from the underground river of your subconscious (Zollo 1997, 95, 98, 120)

Much of Rick Rubin's advice in *The Creative Act: A Way of Being* boils down to recommendations on how to let the subconscious speak:

There's an abundant reservoir of high-quality information in our subconscious, and finding ways to access it can spark new material to draw from. The psyche has admittance to a universal wisdom deeper than what we can come up with in our conscious mind...we don't know how it works and we don't know why it works, but many artists tap into something beyond themselves without recognizing the process at play, purely through accessing the subconscious (2023, 64).

These observations don't entail creativity *cannot* spring from consciousness, of course. The idea is that it *need not*, and often does not, do so. This opens the door to AI displaying genuine creativity, *even if* we are convinced it is not conscious.

A related concern turns on the question of *agency*. According to Dustin Stokes and Elliot Paul (2021), it is the lack of agency in AI that impugns their creativity. Stokes & Elliot here join a chorus of other theorists who have found agency to be an essential precondition for creativity (Gaut 2018, 2010, Carruthers 2011, Moruzzi 2022). In a similar spirit, Boden (2014, p. 233) suggests that *autonomy* is required for creative production, and that it is unclear whether AI has such autonomy. Stokes and Paul's remarks concerning agency and its relation to AI echo those of Hoel with respect to pareidolia:

Water molecules crystallizing in cold air form a unique and intricate snowflake. Wind blows sand into a novel and tranquil pattern of dunes across a desert plain... In each case, the result is something new and aesthetically valuable. But water, wind, and gases are not creative. They're not creative because they're not *agents* who bear responsibility for what they bring about...The reason it isn't clear that these programs are really creative is because it isn't clear that they are really agents (Stokes & Paul, 2021).

Stokes & Paul further propose that real agency is needed to mesh with the praise we mean to attribute when declaring a work creative: "Creative' is a term of praise and we do not extend praise (or blame) for things that are not done by an agent, or for things that an agent does accidentally rather than intentionally. Real creativity requires intentional agency" (*ibid.*). Here the idea is that ascriptions of creativity imply praise, and that praise only makes sense in the context of suitable agential intentions on the part of the person being praised.

Possible responses on behalf of AI creativity mirror those offered with respect to consciousness. One might reply that AI systems are in fact agents—or, at least, that we lack clear reasons for denying their agency. (Asked if it has aims or intentions, Microsoft’s chatbot Bing cheerfully responds in the affirmative: “My aim is to provide you with engaging and intelligent conversations on various topics.”) Certainly, there are fears among many that AI will soon *become* agents skilled at deception, fueling the project of AI explainability and interpretability. Supposing we were open to AI having the minimal aim of carrying out the wishes of those who prompt it, its agential intentions could be compared to those of artists who express their creativity with the aim of fulfilling the prompts of their patrons, directors, choreographers, or editors. Nevertheless, as with the question of AI consciousness, the question of agency in AI is not a matter that can be fully explored here.

More to the point may be to ask whether theorists are correct to think that creativity requires (intentional) agency in the first place. Such a requirement appears at odds with the self-descriptions of many artists. Presumably, when works arise from the subconscious, as described above, they are not intentionally generated. Neil Young speaks pointedly about the need to avoid approaching songwriting as an intentional act:

“Usually I sit down and I go until I’m trying to think. As soon as I start thinking, I quit . . . then when I have an idea out of nowhere, I start up again. When that idea stops, I stop. I don’t force it. If it’s not there, it’s not there, and there’s nothing you can do about it” (Zollo, 1997, p. 354-355).

Stokes and Paul may respond that there is an important difference between what Neil Young is up to here and the person who (in their example) coincidentally, and completely accidentally, traces a beautiful pattern in the snow as they snowboard down a mountain. Is our greater tendency to assign creative merit to Neil Young’s products not linked to a sense that he is exercising a more significant *degree* of agency over the process than the skier?

Perhaps, but it is not obvious that Young’s agency is what accounts for the difference here. After all, in appraising an artwork, we do not generally dwell on how much effort it took the artist to create it. The tendency to associate creativity with agency may instead trace to our natural inclination to praise artists (and other creative individuals) for their works, combined with the reasonable intuition that it makes no sense to praise mere accidents or flukes. Yet, in denying that all instances of creativity must result from genuine agency, we are not forced to conclude that these products are mere accidents or flukes. It is no accident that when you strum a guitar tuned to an open chord the resulting sounds harmonize. In the same way, a generative DNN is tuned during its training across millions of parameters. The kind of activation that eventually ramifies through the network is very far from an accident, even if it is not (necessarily) the result of an intention. In the same way, Neil Young has been intricately tuned by his past experiences and genetics to be the complex, distinctive songwriter that he is. This tuning can likewise ring out in works that spring from his subconscious without being accidents or flukes, and yet without being intentionally generated. (See Kauppinen (forthcoming) for a complementary argument that “creative achievements...can be effortless, out of voluntary or rational control, and independent of the agent’s values and reasoning,” yet still praiseworthy.)

To consider a last related reason for skepticism about AI creativity, Amy Kind (2022) suggests that an artist’s metacognitive capacity to provide an account of how and why they composed their work is important for our assessments both of imagination and creativity. “The way we know that other humans have exercised their imagination is not just by looking at their creative outputs,” she explains, “but by talking to them about their creative processes. It’s their descriptions of their mental activities that incline us to see them as imagining” (2022, p. 56). Once we conclude they are imagining, she suggests, may more confidently judge their works to be creative. Thus, an assessment of machine creativity may “have to await the day when machines develop not only the capacities for creative output but also the capacities to explain to us how they got to that output” (*ibid.*).

However, Kind’s claims again seem to clash with the tendency among artists to profess ignorance about the origin of their best ideas. It will be difficult to maintain that such awareness is a necessary condition for creativity, even if it may at times increase our confidence concerning its presence.

As it happens, GPT-4 is happy to provide detailed (though perhaps hallucinatory) accounts of its artistic decisions. After it wrote the poem I included above, I asked if it could explain to me its reasons for writing the poem in the way it did. GPT-4 naturally obliged, offering an enumerated list of its many putative decisions, including:

*1. Structure:* I chose a twelve-line poem, as requested, and divided it into three quatrains (four-line stanzas)... *2. Rhyme scheme:* I opted for an ABAB rhyme scheme for each quatrain, providing a sense of unity and balance to the poem while also giving it a rhythmic quality that is pleasant to read or recite... *3. Diction:* To create a contemplative atmosphere, I used language that is slightly more formal and elevated than everyday speech. This choice aims to convey the significance of the subject matter and provide a sense of poetic weight...

And so on (and on) it went, being at no loss for words concerning its aesthetic deliberations. Notably, GPT-4 nevertheless made a glaring mistake in its account: the rhyme scheme of its poem is AABB, not ABAB. Whether this kind of confabulatory error in its account of its own artistic process makes it more or less humanlike, I leave to the reader’s discernment.

To my mind, if there is something in the vicinity of consciousness, agency, and metacognitive awareness that is lacking in AI products and still essential to creativity, it may be the expression of an individual perspective or character. (See Giannuzzo (2023) for an articulation of related concerns.) Whether a work emerges through purely subconscious forces or not, we tend to be interested in what it says about the individual who created it—in what particular ideas, perspectives, or emotions it is expressing. Here much of contemporary AI faces a challenge, due to the incredibly broad and indiscriminate nature of its training sets. Because a general-purpose image generating tool such as Midjourney or Dall-E 3 is trained on more or less *all* the images that can be scraped from the internet, and because none are preferentially weighted, it may seem to lack a distinctive voice or individual aesthetic. The same points apply to LLMs trained on massive amounts of internet text. (Perhaps *this* is the “blurriness” that concerns Chiang.)

However, as the accessibility of self-trainable AI models increases, we may see more prominent individualized models, trained on limited and highly curated data sets. It will be interesting to observe whether these more idiosyncratic models provide a stronger impression of expressing an individual viewpoint of a kind associated with creative expression.

## 6. AI and Imagination

Do contemporary AI tools imagine? One approach to this question is to start with a definition of imagination (such one provided in the Introduction to this volume) and, looking at a DNN's processing, to try to discern whether the kind of processing carried out satisfies that definition. For instance, if one grants that imagistic imagining requires the use of mental imagery (at least as a necessary feature), one could ask whether AI makes use of mental imagery in the generation of its products. Alternatively, if one thinks that attitudinal imagining requires taking a certain non-committal attitude toward a proposition, one might try to assess whether a DNN's processing can be properly seen as realizing or implementing representations with propositional contents that are treated in the relevant non-committal way.

These projects force us to address difficult and unsettled questions about whether and how familiar representational kinds—such as language-like (or “propositional”) representations and image-like (or iconic) representations—relate to complex connectionist networks of the kind that constitute DNNs. There are longstanding debates in the philosophy of cognitive science concerning whether the processing in such networks can be seen as implementing language-like or imagistic representations at some higher level of description (Fodor and Pylyshyn 1988, Pavlick 2023), or whether, instead, such systems simply lack any such representations (Churchland 1981, Chalmers 1993, Churchland and Sejnowski 1989). This is a complex and important project—one closely aligned with the project of determining whether AI systems are interpretable or “explainable.” I cannot hope to settle it here, but a few remarks are in order.

It is common both to hold that mental imagery is essential to imagination (Kind, 2001) and that a defining characteristic of such imagery is that it occurs in a *pictorial* or *iconic* representational format (Kosslyn, Thompson, and Ganis 2006, Fodor 2003, Nanay 2023). If these views are correct, AI tools can only imagine if they, too, make use of pictorial or iconic representations. This is where questions concerning how to determine whether a DNN implements representations in an imagistic format are relevant. However, an alternative approach questions whether mental imagery in humans occurs in an iconic or pictorial format (Pylyshyn 2003). If it doesn't, then the lack of iconic representations in generative AI is no barrier to its using mental imagery.

Further, there are alternative definitions of mental imagery that may weigh in favor of imagery in AI. For instance, Bence Nanay characterizes mental imagery as “perceptual processing that is not triggered by corresponding sensory stimulation in a given sense modality” (2018b, p. 127). The processing of generative AI can reasonably be considered “perceptual” in the sense that it is reusing the same networks and weightings that, originally (i.e. during training), were used in perceptual discriminative tasks. In generative contexts, this processing is triggered not by “sensory stimulation,” but, instead, by something more like what triggers mental imagery in

humans—e.g., an intention or prompt to represent thus and such. Thus, it may not be a stretch to characterize generative DNN processing as “perceptual processing that is not triggered by corresponding sensory stimulation.

An alternative approach to assessing imagination in AI is to work from the outside in, as it were, by tethering the question of imagination to the question of creativity. Suppose, for instance, that we accept the following creativity-requires-imagination principle:

Creativity Requires Imagination (CRI): *All genuinely creative products are the result of imagination.*

CRI assumes that part of what we *mean* by ‘imagination’ is something like: “the sort of process required for the generation of genuinely creative works (whatever else it may be used for).” With CRI in hand, we can be confident that a work resulted from imagination as soon as we are confident that it is genuinely creative. This approach makes sense if, like me, you think we have a better grip on what it is for a product to be creative than we do on what its for a mental episode to be case of imagination. To my mind, the word ‘imagination’ gains a foothold in our language in part as a term to pick out the kind of cognition responsible for creative works. This creates a presumption that creative works indeed result from imaginings. We can remain open to the presumption being overturned. But the bar for doing so should remain high.

I have already offered reasons to think that generative AI tools already, or soon will, evidence genuine creativity. If CRI is correct, we likewise have good reason to think that AI tools already, or soon will, imagine.

There are at least two ways one might resist this line of reasoning. The first rejects CRI, holding that genuinely creative works need not result from imagination. The second accepts CRI while denying that AI products are genuinely creative—perhaps *because* one is skeptical that AI genuinely imagines. Let’s begin with the second.

We’ve seen that some AI tools are at least *prima facie* capable of creativity. Now suppose that someone who would otherwise grant the creativity of an AI’s product denies its genuine creativity just because they are convinced that AI systems do not truly imagine. On what grounds might they think that the AI system doesn’t imagine? There could be various, but the most likely may be a suspicion that such systems are not conscious, combined with the further view that imagination requires consciousness. After all, it would be odd to grant that an AI system is conscious and is generating works that are *prima facie* creative, while denying that its products are genuinely creative just because the (conscious) system is not *imagining*. It seems that consciousness is the true sticking point; once we are convinced of its consciousness—and, perhaps, of the fact that it wasn’t programmed by someone with its specific outputs in mind—it becomes difficult to deny the creativity of its products. We are back, then, to our earlier question of whether consciousness is necessary for creativity. Denying that AI is creative in such a situation *simply* on the grounds that it does not imagine appears unmotivated.

The alternative route to denying imagination in AI is to reject CRI. Amy Kind (2022) lays the groundwork for this kind of response in a broader discussion of the relation of creativity to

imagination. There she proposes that a solution to a difficult mathematical problem—arrived at by an imagined character Marisol—could be genuinely creative, yet not require imagination:

There’s no special reason to think that Marisol called upon her imagination to arrive at this solution...the steps that she took to work her way through the problem *seemed completely natural and even automatic* to her. She *never even had to pause to consider* what to do next. And she *did not produce any imagery* – visual or otherwise...she took an unusual pathway through the various mathematical rules that had been previously taught to her, but she needn’t have used her imagination to find that pathway” (2002, p. 36, emphases added).

Kind suggests, by negative implication, that imagining is an effortful, intentional, deliberative, imagery-involving process that need not be involved in arriving at a creative mathematical solution. Taking her description of Marisol at face value, we can ask whether any of these missing features is indeed necessary for imagination.

Let’s begin with the mental imagery. If using mental imagery is essential to imagining (as explicitly argued by Kind (2001)), and if generative AI systems don’t use imagery, then we have a quick argument to the conclusion that these systems do not imagine. I have already noted the controversy behind the question of whether DNNs make use of imagery. Another route for response is to deny that all cases of imagining involve mental imagery. It is common, for instance, to allow for a form of “attitudinal” or “belief-like” imagining that does not always involve the use of mental imagery (Van Leeuwen 2014, Langland-Hassan 2020, Currie and Ravenscroft 2002, Arcangeli 2018). If there are indeed imageless imaginings, then AI tools (and Marisol) may imagine even if they do not exploit mental imagery.

What about Kind’s suggestion that some creative cognition is too “automatic” or non-reflective to count as imagining? Here again we can look to the firsthand reports of artists for plausible counterexamples. It is common for artists to lack a sense of agency over their creative cognition. We can only deny that there is “automatic,” non-reflective imagining at the price of holding that these artists are not really imagining when generating products that seem to them to arise spontaneously. It is not clear to me what would warrant using the term ‘imagining’ in this more limited way. My comparatively stronger grip on the notion of creativity leads me to err on the side of maintaining CRI. If we are otherwise confident that Marisol’s solution, or the AI’s poem, is creative, questions of whether it was produced in a seemingly automatic or effortless way should not lead us to doubt it was a product of imagination.

## **7. Conclusion**

The present moment is a difficult vantage from which to assess the possibility of imagination and creativity in AI. The incredible recent developments in AI can give the impression that there are simply no bounds to what AI will achieve in the years to come. Yet it may be that some of the current limitations on AI will continue to pose significant obstacles. I have tried to highlight the best reasons for skepticism about creativity and imagination in AI and have suggested they have adequate responses. I do wonder what I will make of this essay ten years from now.

## References

- Arcangeli, Margherita. 2018. *Supposition and the Imaginative Realm: A Philosophical Inquiry*. New York: Routledge.
- Boden, Margaret A. 1998. "Creativity and artificial intelligence." *Artificial Intelligence* 103 (1):347-356. doi: [https://doi.org/10.1016/S0004-3702\(98\)00055-1](https://doi.org/10.1016/S0004-3702(98)00055-1).
- Boden, Margaret A. 2004. *The creative mind: Myths and mechanisms*: Psychology Press.
- Boden, Margaret A. 2014. "Creativity and Artificial Intelligence: A Contradiction in Terms?" In *The Philosophy of Creativity: New Essays*, edited by E. Paul and S. B. Kaufman, 224-244. Oxford: Oxford University Press.
- Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, and Scott Lundberg. 2023. "Sparks of artificial general intelligence: Early experiments with gpt-4." *arXiv preprint arXiv:2303.12712*.
- Buckner, Cameron, and J. Garson. 2019. "Connectionism." *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition).
- Carruthers, Peter. 2011. "Creative action in mind." *Philosophical Psychology* 24 (4):437-461.
- Chalmers, David J. 1993. "Connectionism and compositionality: Why Fodor and Pylyshyn were wrong."
- Chalmers, David J. 2023. "Could a large language model be conscious?" *arXiv preprint arXiv:2303.07103*.
- Chiang, Ted. 2023. "ChatGPT is a Blurry JPEG of the Web." *The New Yorker*, February 9.
- Churchland, P.M. 1981. "Eliminative Materialism and the Propositional Attitudes." *Journal of Philosophy* 78:67-90.
- Churchland, P.S., and T.J. Sejnowski. 1989. "Neural representation and neural computation." In *Neural Connections, Mental Computations*, edited by L. Nadel, L. Cooper, P.W. Culicover and R.N. Harnish. Cambridge, MA: MIT Press.
- Currie, G., and I. Ravenscroft. 2002. *Recreative Minds: Imagination in Philosophy and Psychology*. Oxford: Clarendon Press.
- Eliot, T.S. 1998. *The Sacred Wood and Major Early Essays*. Mineola, New York: Dover.
- Fodor, Jerry A. 2003. *Hume Variations*. Oxford: Oxford University Press.
- Fodor, Jerry A., and Zenon W. Pylyshyn. 1988. "Connectionism and cognitive architecture: A critical analysis." *Cognition* 28 (1):3-71.
- Gaut, B. 2010. "The Philosophy of Creativity." *Philosophy Compass* 5 (12):1034-1046.
- Gaut, B. 2018. "The Value of Creativity." In *Creativity and Philosophy*, edited by B. Gaut and M. Kieran, 124-139. London: Routledge.
- Giannuzzo, Anaïs. 2023. "Creativity, Intentions, and Self-Narratives: Can AI Really Be Creative?", Cham.
- Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. "Generative adversarial nets." *Advances in neural information processing systems*.
- Guzik, E., C. Byrge, and C. Gilde. 2023. "The Creative Potential of ChatGPT." *Creativity Conference, University of Southern Oregon, May 18, 2023*.
- Hoel, Erik. 2022. "AI-art Isn't Art." *The Intrinsic Perspective, May 18, 2022*.



- Kauppinen, Antti. forthcoming. "Creativity, Spontaneity, and Merit." In *Philosophy and Art: New Essays at the Intersection*, edited by Alex King and Christy Mag Uidhir. Oxford: Oxford University Press.
- Kind, Amy. 2001. "Putting the Image Back in Imagination." *Philosophy and Phenomenological Research* 62 (1):85-109.
- Kind, Amy. 2022. *Imagination and Creative Thinking*: Cambridge University Press.
- Kosslyn, S., W.L. Thompson, and G. Ganis. 2006. *The Case for Mental Imagery*. Oxford: Oxford University Press.
- Langland-Hassan, Peter. 2020. *Explaining Imagination*. Oxford: Oxford University Press.
- Ledig, Christian, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, and Zehan Wang. 2017. "Photo-realistic single image super-resolution using a generative adversarial network." *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Moruzzi, Caterina. 2022. "Creative Agents: Rethinking Agency and Creativity in Human and Artificial Systems." *Journal of Aesthetics and Phenomenology* 9 (2):245-268. doi: 10.1080/20539320.2022.2150470.
- Nanay, Bence. 2023. *Mental Imagery: Philosophy, Psychology, Neuroscience*: Oxford University Press.
- Pavlick, Ellie. 2023. "Symbols and grounding in large language models." *Philosophical Transactions of the Royal Society A* 381 (2251):20220041.
- Pylyshyn, Z. 2003. *Seeing and Visualizing: it's not what you think*. Cambridge, MA: MIT Press.
- Radford, Alec, Luke Metz, and Soumith Chintala. 2015. "Unsupervised representation learning with deep convolutional generative adversarial networks." *arXiv preprint arXiv:1511.06434*.
- Roose, Kevin. 2023. "Bing's AI Chat: 'I Want to Be Alive!'" *The New York Times*, February 17, 2023.
- Rubin, Rick. 2023. *The Creative Act: A Way of Being*. New York: Penguin.
- Shanahan, Murray. 2022. "Talking about large language models." *arXiv preprint arXiv:2212.03551*.
- Shin, Minkyu, Jin Kim, Bas van Opheusden, and Thomas L. Griffiths. 2023. "Superhuman artificial intelligence can improve human decision-making by increasing novelty." *Proceedings of the National Academy of Science* 120:e2214840120. doi: 10.1073/pnas.2214840120.
- Stokes, Dustin, and E. Paul. 2021. "Computer creativity is a matter of agency." *IAI News*, November 11, 2021.
- Van Leeuwen, Neil. 2014. "The meanings of "imagine" part II: Attitude and action." *Philosophy Compass* 9 (11):791-802.
- Zollo, Paul. 1997. *Songwriters on Songwriting*. New York, NY: Da Capo Press.