

Modeling Generative Artificial Intelligence

by

Haochen Xiong

Department of Art, Art History, and Visual Studies
Duke University

Defense Date: November 27, 2023

Approved:

Mark Olson, Supervisor

Bill Seaman

Edward Triplett

Annabel Wharton

Thesis submitted in partial fulfillment of the requirements for the degree of Master of
Arts in the Department of Art, Art History, and Visual Studies in The Graduate School of
Duke University
2023

ABSTRACT

Modeling Generative Artificial Intelligence

by

Haochen Xiong

Department of Art, Art History, and Visual Studies
Duke University

Defense Date: November 27, 2023

Approved:

Mark Olson, Supervisor

Bill Seaman

Edward Triplett

Annabel Wharton

An abstract of a thesis submitted in partial fulfillment of the requirements for the degree of
Master of Arts in the Department of Art, Art History, and Visual Studies in The Graduate School
of
Duke University
2023

Copyright by
Haochen Xiong
2023

Abstract

The release of ChatGPT-4 has led to the prevalent use of a new term in the field of artificial intelligence (AI): generative AI. This paper aims to understand generative AI more thoroughly and place it within a broader framework of models and their relationship with knowledge. By closely examining AI's historical development, this paper will first introduce the concept of emergence to distinguish generative AI from other forms of AI. Second, by theorizing generative AI as models, this paper will evaluate their significance in human knowledge production. Third, by classifying generative AI specifically as generative models, this paper will demonstrate their unique potential, especially for art creation.

Contents

Abstract.....	iv
List of Figures.....	vii
1. Introduction.....	1
2. What is an AI?	6
2.1 Theoretical AIs.....	10
2.2 Practical AIs	11
3. Being Generative	14
3.1 Generative AIs as Total Black Boxes.....	15
3.2 Opening up the Black Box of Generative AIs.....	23
4. Generative AI as Models	30
4.1 What is a Model?.....	30
4.2 Black’s Definition of Models	32
4.3 As If vs. As It Is.....	33
4.4 A Meta-Model of Models.....	35
4.5 From Models to Knowledge.....	37
4.6 Models within the Circular Process of Knowledge Production.....	42
4.7 The Hermeneutic Loop and the Heuristic Loop	46
5. Generative Models.....	50
5.1 Procedural Modeling	50
5.2 Generative Models in PM.....	51
5.3 PM’s Impact on Artistic Knowledge Production	54

5.4 Generative Models' Potential Learned from PM	55
5.5 Generative AIs' Potential for Art Creation	60
6. Conclusion	64
Appendix.....	67
References	69

List of Figures

Figure 1: Google Trends of the Term “AI” from 22 Sep. 2020 to 22 Sep. 2023	2
Figure 2: The Initial Illustration of Neural Networks	18
Figure 3: Illustration of Second-Order Cybernetics.....	43
Figure 4: Total Randomness vs. Pseudo Randomness (Perlin Noise)	52
Figure 5: Nodes I used in Houdini for my PM Model	58
Figure 6: Screenshots of my PM model in Houdini.....	58
Figure 7: Nodes I used in Houdini for better distinguishable visualization.....	59
Figure 8: Screenshots of my PM model after rendering	59
Figure 9: Visualizations of the term Neosentience	62
Figure 10: Visualizations of the term Neosentience based on two different training models	64

1. Introduction

ChatGPT's explosive global popularity has returned artificial intelligence (AI) to the forefront of public concern (see fig. 1). Both consumers and enterprises have been using AI for years, such as engaging with voice assistants to automate daily routines and employing algorithms to identify "patterns and correlations in data" ("The Implications of Generative AI in Finance"). However, unlike previous approaches designed to solve specific tasks and often limited within particular domains, so-called generative AI reveals a disruptive potential in processing more intelligent and more inter-domain requests, which used to be believed can only be done by humans. ChatGPT is capable of generating original content, such as writing sequels, crafting jokes, and coding mini games. Although the results can sometimes appear nonsensical, a user can constantly regenerate the results to find what is arguably more creative and more satisfactory results. With the combination of Dall·E 3, the newest version of OpenAI's text-to-image model, ChatGPT is now also capable of generating detailed images based on the text descriptions that the user provides ("Dalle·3"). In a sense, generative AIs seem to show us an actual inflection point where AI is starting to correspond to our futuristic imagination of it: an intelligent agent, made by us but potentially smarter than us, bringing profound changes to every aspect of our lives.

However, the question is, what is a generative AI? As a term, it appears frequently in AI-related articles after ChatGPT became popular. Almost every big technology company declares an effort to develop a generative AI to improve the user experience of its current applications. While Adobe claims its generative AI "[having] the potential to reshape every aspect of marketing" ("Meet Adobe Sensei GenAI"), Amazon states its generative AI can "boost productivity, build differentiated experiences" with Amazon Web Services (AWS) ("Generative

AI on AWS”). Besides its ability to generate, there is no distinct definition of Generative AI in current discussions. Therefore, this paper aims to explain generative AIs: what they are, what they can do, and what futures they can bring us. It starts with the definition of AI. Clearly, there is a difference between the actual science being conducted within this field and the hyperbolic discourse of technological futurists. While a person might imagine an AI as an android as shown in science fiction films (e.g., *Westworld* (1973); *Blade Runner* (1982); *The Matrix* (1999)), current AIs, including generative ones, are all still fundamentally machine learning algorithms. Jacob Roberts emphasized that such a disconnect “is in many ways rooted in the early years of AI research and in the naming of the field itself.” Therefore, this paper first takes a historical approach to understanding AI from both its theoretical and practical development. In a sense, it attempts to facilitate a more general comprehension of AI by combining previous separate approaches.

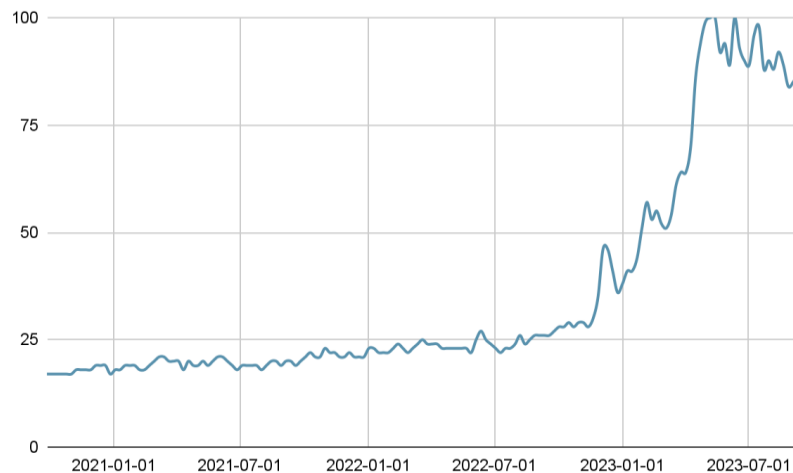


Figure 1: Google Trends of the Term “AI” from 22 Sep. 2020 to 22 Sep. 2023. Numbers on the Y-axis represent Google's “search interest relative to the highest point on the chart for the given region and time.” *Google*. Trends.Google.com. <https://trends.google.com/trends/explore?date=2020-09-22%202023-09-22&q=AI&hl=en>. Accessed 22 Sep. 2023.

This paper will then define generative AI, laying out the elements that distinguish it from the previous ones. As mentioned, an explosive use of generative AI comes with the popularity of ChatGPT, yet the term itself lacks a distinguishable definition. I argue that a generative AI is a human-made emergent machine, where accumulated quantification has led to qualitative changes, emerging new global behaviors that exceed our existing understanding of its composing agents (variables, neurons, physical units, etc.). Such a characteristic of being emergent empowers generative AIs, whereas the previous AIs are all to be understood as forms of algorithm, exploring stepwise reason - “using clearly explicable rules of logic” – induction, deduction, and abduction (Lee 1). Generative AIs are not only black boxes for their users but also for their creators. While hundreds of “the biggest names in tech, including Elon Musk,” signed an open letter to urge a six-month stop from developing generative AI, precisely because of its inexplicability as a total black box and its uncontrollability of being emergent (Perrigo), this paper encourages an unlimited development of them, yet with authored supervision and clear regulation rules when releasing to the public. Generative AI can inform new knowledge, when we recognize it as a new variety of emergence (what comes forth without clear patterns for humans to define) and study what leads to those particular emergent behaviors. In a sense, opening up such a total black box after its construction might bring us knowledge that we cannot receive through our previous ways of knowledge production.

Following such a narrative, understanding how generative AIs can change our mode of producing knowledge, is this paper’s ultimate goal. Marx Wartofsky emphasized the history of human knowledge itself, in which he stated:

By this I meant not simply that what we know has a history, or that there is a historical development of ideas or theories; but that the nature of knowing, of cognitive acquisition itself, changes historically; how we know changes with changes in our modes of social and technological practice, with changes in our forms of social organization. In effect, I argued that what we take knowledge to be is itself the subject of a historical evolution (13).

To understand generative AIs' relation to knowledge production, this paper takes a top-down approach, by first recognizing generative AIs as models - a subject that has been argued to play an essential role in our knowledge production. Undoubtedly, every existing AI can be defined as a model, or more specifically, a computational model, either digital, analog, or mixed¹. Analyzing generative AIs under such a perspective permits this paper to apply previous theories of models to review the existing knowing and learning patterns under our use of them. However, precisely because the term model itself contains pluralistic meanings, my argumentation will start by explaining what a model is. Based on the arguments and analyses from scholars like Max Wartofsky and Max Black, I propose a meta meaning of, or following its own narrative, a meta-model of models, so as to better understand generative AIs as models.

Then, by distinguishing generative AIs as generative models within the meta-model category, this paper aims to illustrate its potential ways of altering the nature of how we know and how we learn in practice. I argue a generative model is an emergent computational system constructed to at times solve problems or process tasks that potentially exceed or do not follow human computational rationality – to employ stepwise reasoning using clearly explicable rules of logic. Under such a definition, then, a model being generative is not a new story. The history of humans using computational models to generate can be traced back to the first use of computer simulation to predict the unknown, namely, the neutron problem faced by Jon Von Neumann and Stanislaw Ulam² (Balachandran et al.). Therefore, my method towards generative AIs' unique impacts on knowledge production relies on examining how previous generative models have done

¹ *Deep learning* (which will be further discussed in section 3) can be seen as an analog computational model, trying to reproduce some observed functions of human neuroanatomy, but on digital hardware and through digital methods.

² Neumann and Ulam were trying to understand the behavior of neutrons, where “[hit] and trial experimentation were too costly and the problem was too complicated for analysis” (Balachandran et al.).

so throughout history; thus, predicting their future potential. My analyses of the previous generative models are offered in the form of an outcome pushed by them: procedural modeling.

In all, this paper seeks to examine how the current generative AIs rely on the already known so as to evaluate how they can “[embody] and [promote] the never before seen” (Huhtamo and Parikka 14). Written at a time when AIs have more and more potential to become what we have imagined, this paper ultimately aims to discuss possible ways into a future where generative AIs, as generative models, would be ideally used and addressed. Considering the complexity of such a topic, there is no doubt that this paper can only cover a limited part. Yet it intends to provide an initial framework, with detailed background information, to encourage further future research.

2. What is an AI?

The daily use of the term AI more or less points to an illustration of a certain intelligence that can be equally smart or even smarter than humans. Sci-fi movies often anthropomorphize AI (offering them human appearances, human emotions, etc.) and depict a dystopian world where the androids trump human intelligence and threaten the lives of humans. Additionally, these kinds of movies always reflect the cause of such a situation back on humans, how we design those AIs in the first place and how we treat them during their “service” period. Undoubtedly, the narrative of these movies is only an epitome of how we project what AI might become. However, we cannot deny that such a dystopian view of AI from movies, novels, and daily news, constructs some pre-assumptions on our perspectives. This can be reflected in the current skepticism towards AI’s creativity. In her paper, Mingyong Cheng emphasized that, “painting produced by artists categorized with an AI identity, [...] gain lower rating on the value of their work, compared to paintings done by artists with a human identity,” due to the negative stereotypes of AI and AI-generated paintings (2). In this case, our common understandings of AI are always entangled in discourse.

There is no true AI if following the narrative as mentioned above. The term itself has been mainly thought to be coined at the Dartmouth Summer Research Project on Artificial Intelligence conference (abbreviated as the Dartmouth Conference) in 1956. Within the proposal of the conference published in 1955, John McCarthy, along with other team members, described AI as a machine that can simulate “every aspect of learning or any other feature of intelligence” (McCarthy et al. 12). While their description can arguably be recognized as the original definition of what an AI is, the journey to understand and build such a machine had begun much before that. The original concept of AI can be traced back to the starting point of cybernetics, where Norbert

Wiener, who has been considered the originator of cybernetics, theorized that “all intelligent behavior was the result of feedback mechanisms,” and therefore could possibly be simulated by machines. Again, though Wiener coined the term cybernetics as “the scientific study of control and communication in the animal and the machine,” the emergence of the concept of cybernetics can be traced back to a conference held in 1942 under the auspices of Josiah Macy Jr.

Foundation¹. This meeting focused on “some timely issues at the junction between psychology and brain science” (Conway and Siegelman 131). One year later, one of the conference members, Warren McCulloch, along with Walter Pitts, “picked up the banner of Wiener’s new ideas” (136) and created the first logical model of neurons. An artificial neuron is now the fundamental unit of a neural network, which is the current mainstream method to train generative AIs.

Summarizing this history demonstrates that the term AI has always been a highly theorized concept. Additionally, it shows that the initial goal of developing an AI focuses on building a mechanical system to mimic what intelligence can do, rather than a humanized machine. Furthermore, the underlying goal of this development is to understand humans better, and intelligence in general. This can be reflected in the arguments from McCulloch and Pitts, where they stated, “both the formal and the final aspects of that activity which we are wont to call *mental* are rigorously deducible from present neurophysiology” (114). That is to say, from a starting point, their models of neurons, their so-called theoretical neurophysiology, attempt for further insights into how intelligent brains work.

Humanity’s attempt to understand what intelligence is and what intelligence can do has a long history. The term itself derives from the Latin word *intellegentia*, which means the “action or faculty of discerning or understanding” (“intelligence, n.”). In the European Middle Ages,

¹ The conference, with those coming afterwards, were later known as the Macy Conferences on Cybernetics after Wiener coined the term (Scott 1366).

intelligence referred to “understanding in which truth offers itself like a landscape,” where God is the source of it (Piper 28). The modern study of intelligence is often traced back to the work of Charles Spearman, who proposed a general factor (g) that influences performance across various cognitive tasks. According to him, this general factor, or in other words, general intelligence, reflects a person’s overall cognitive ability, while specific abilities (s) are task specific.

Furthermore, the imagination of a manufactured intelligence always comes with our corresponding understanding of the term, as well as our corresponding technology skills at that specific time. The Greek automatons were “animate, metal statues of animals, men, and monsters,” crafted by the God Hephaestus and architect Daedalus, whereas “[the] best of them could think and feel like men” (“Automotones”). The novel “Frankenstein,” published in 1818 by Mary Wollstonecraft Shelley, revolved around Victor Frankenstein, a young scientist who successfully brought a creature (assembled from dead humans’ tissues) to life using a combination of alchemy and chemistry, but with a horrified monstrous appearance.

However, there is always a disconnect between what science is doing and our imagination. The latter can be recognized as a sociocultural response to the former. What if those scientific studies become realized? How will it change our current ways of thinking and living? From a historical perspective, Shelley’s novel can undoubtedly be seen as an exploration of the philosophical and ethical themes underlying the rise of alchemy and chemistry studies back then. The monstrous-looking creature was her reflection on scientific hubris, the consequences of humans playing God, as well as the responsibility that comes with scientific discoveries.

Additionally, a separation often takes place between scientific theories and their practices. When turning the former into the latter, the actual construction of manufactured intelligence has to follow the corresponding evaluation standards based on our existing technology skills. Though Spearman started our modern study of intelligence, the modern

measurement of it can be traced back to the work of Alfred Binet and Theodore Simon. The Binet-Simon Scale they developed can be recognized as the “forerunner of the modern Stanford-Binet Intelligence Scales,” which has become the standard measure of a human’s intelligence since 1916 (Sternberg).

Variables play an essential role in our scientific measurements. One of the key innovations of the Binet-Simon Scale was the concept of "mental age." Binet introduced the idea that intelligence could be measured by comparing an individual's performance on the test to the average performance of children at different age levels. Furthermore, although Binet did not introduce the term "IQ" (intelligence quotient), the concept evolved from his work. The IQ is calculated by dividing a person's mental age by their chronological age and multiplying the result by 100. Lewis Terman later modified this formula in the United States, resulting in the widely used IQ score we are familiar with today (Sternberg).

The same situation happened when McCarthy, along with other team members, tried to turn their theories of AI into scientific practice. Though defining the term as a machine capable of simulating every aspect of learning or any other feature of intelligence, their measurements focused on “how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.” Then, based on such measurements, they each delved into a specific aspect of what they call “the artificial intelligence problem,” such as “abstractions” and “self-improvement” (12-14). Undoubtedly, their measurements are more closely related to what science has been doing in the field of AI. In a general sense, any current generative AI, such as ChatGPT, can meet their measurements. Therefore, while our futuristic imagination of AI more or less functions as a sociocultural response to its scientific development, what an AI essentially is within the field of science can be

understood through two seemingly disconnected but, in a sense, interrelated perspectives: the theoretical and the practical.

2.1 Theoretical AIs

The theoretical understanding of AI can arguably be divided into three categories. The first category is called artificial narrow intelligence (ANI), also known as narrow AI or weak AI. Measurements from the Dartmouth Conference, as mentioned above, are now widely accepted as an original definition of ANI. Just like how McCarthy and other team members separated what intelligence can do to different particular aspects, ANI often refers to systems designed and trained for specific and well-defined tasks (Reilly). It is the dominant form of AI in use today, both in our daily lives and in many industries. Typical examples of ANI are: 1) voice assistants, such as Siri or Alexa, 2) image and facial recognition software used in social media or security systems, and 3) game-playing AI, like Deep Blue and AlphaGO.

The second category of theoretical AI is called artificial general intelligence (AGI), also known as general AI or strong AI. The concept of AGI emerges when we shift our key concern of AI, from what an intelligence can do, to what distinguishes an intelligence as it is. That is to say, instead of simulating aspects or features of intelligence, AGI focuses on achieving what the term intelligence means. The difference between ANI and AGI can be explained by reflecting back on Spearman's arguments. While the former emphasizes specific abilities (s), the latter concerns the general factor (g), the broad cognitive abilities associated with intelligence. Under such a narrative, AGI often refers to AI that possesses the ability to understand, learn, and apply knowledge across a wide range of tasks at a level comparable to human intelligence (Reilly). However, precisely because human intelligence, as well as intelligence in general, are still developing concepts, the definition of AGI often depends on the understanding of them in the

first place. The term “strong AI” was coined by John Searle in 1980, but his idea of it differs from that of other futurists like Ray Kurzweil. On one hand, Searle was against the possibility of strong AI - machines that can possess consciousness and understanding comparable to human cognition. On the other hand, Kurzweil recognizes strong AI as “human level” AGI, interested in how they can obtain the general factor (g) like humans (260). This type of school of AGI studies ignores conditions like consciousness, declaring that “as long as the program works, they do not care if you call it real or a simulation” (Russell et al. 947). That is to say, precisely because we do not know how to define and evaluate consciousness, it is unnecessary for us to examine if an AI actually has it. We can only note when an AI appears to demonstrate that it has consciousness through observable behaviors.

The third category to theoretically understand AI is named artificial superintelligence (ASI), or shortly, superintelligence. Nick Bostrom coined the term “superintelligence” in 2014, referring to “any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest” (22). While we can arguably declare that current generative AIs partially achieve the notion of AGI, ASI is certainly a theoretical concept more corresponding to our previous futuristic imagination of AI in general. Therefore, the concept of ASI functions more as a sociocultural response to the ongoing development of AGI. This can be reflected in how Bostrom deals with such a concept within his book: two-thirds of it focuses on possible ethical, societal, and existential questions brought by the emergence of ASIs. In a sense, ASI is central to AI’s present discourse.

2.2 Practical AIs

After discussing AI’s three theoretical categories, it can also be argued that its practical development contains three stages. The first stage is driven by “*machine learning*,” “where

predictive models are trained on historical data and used to make future predictions” (Bommasani et al. 3). Machine learning is the fundamental characteristic that distinguishes AI from other computer programs. However, their creation still requires domain experts to write domain-specific algorithms, as a result of which they can only be implemented in that specific domain (3-4). The second stage of AI development relies upon “*deep learning*.” Through larger datasets and the availability of more powerful computation, AI is able not only to create higher-level features, compared to machine learning, but also to learn how to learn. This is the moment when AI has been released from a specific domain. An AI model’s architecture, namely, how it learns to learn, can be applied to various fields and for various usages (4). The AI trained with text data can be used for image identification as it is driven by the same architecture but not field-specific algorithms. The third stage starts when AI can conduct “*unsupervised learning*” or “*self-supervised learning*.” The latter method can be recognized as a type of the former, with differences in goals. Both methods give AI input data without explicit instructions on what to do with it. While unsupervised learning aims to find patterns, structures, or relationships within the data, self-supervised learning aims to generate supervision signals from those data, creating its own labels or tasks to learn from. In all, the absence of human annotation essentially distinguishes third-stage AIs from the previous ones. AI is now forced to learn how to learn, through any data and only regulated by itself (4-5). Bommasani et al. define this particular kind of AI, which is currently rising in our society, as a “foundation model.” ChatGPT-4, the fourth version of ChatGPT from OpenAI, can partially fall into such a category, as it was trained through a combination of *supervised* and *unsupervised learning*. While previous AIs are still designed for specific tasks, foundation models can be recognized as a model of AI models, which is “itself incomplete but serves as the common basis from which many task-specific models are built via adaptation” (7). For example, the upcoming Microsoft 365 Copilot can be recognized as

a task-specific model leveraged on the advanced capabilities of ChatGPT-4 (Spataro). Later sections will further elaborate on AIs as models, and also such a concept of model of models.

In all, the differences between a theoretical and a practical understanding of AI originate from our different approaches. The former takes a top-down approach, starting from asking general questions like: What is intelligence? What is an intelligent machine? Then, scholars dive deeper into those questions, specifying concepts underlying those general questions and detailing steps toward a final answer. Meanwhile, the practical perspective takes a bottom-up approach, starting with those detailed steps: writing algorithms for computers to recognize words, images, and sounds, and more importantly, to improve their performances on such recognitions through those algorithms. However, though with differences, these two perspectives are interrelated as they push each other's development back and forth, creating a feedback loop for us to gradually build a clearer and clearer pattern to understand and create AIs. From a current standing point, as long as ASI is still theoretical, this feedback loop will keep going, and AI will always be a changing concept, along with our futuristic imagination, as a sociocultural response to its changes. Under such a narrative, let us move to AI's newest and currently most popular form: generative AIs.

3. Being Generative

As a term separated from any names of AIs I have mentioned, I want to classify generative AI first to avoid misunderstandings. I argue that all existing generative AIs are ANIs, meaning that they are all merely processing specific tasks. Though ChatGPT-4 can be claimed to partially be an AGI, as if it behaves to have the consciousness to understand our requests and generate original content, I still recognize it as an ANI because, fundamentally speaking, ChatGPT-4 is only good at dealing with texts, images, sounds, and codes. Under such a narrative, all current publicly available AIs are ANIs. Therefore, my use of the term AI in later sections all fall into this category. Additionally, from a practical perspective, generative AI can be trained by *supervised learning*¹, *unsupervised learning*, or a combination of both. Based on these classifications, then, the distinction between generative AIs and the others focuses on what the term “generative” represents. OED provides two meanings of it:

1. Of or relating to the generation of offspring; having the power or function of reproducing.
2. That generates, produces, or gives rise to something, or has the power or ability to do so (“generative, adj.”).

In a sense, all AIs can produce or reproduce something. OED’s definition of the term generative seems to merely explain an action or behavior shared by all AIs. To understand generative AI’s generativity, we have to dive deeper into its generating process, namely, how it produces and reproduces. I argue that it is the emergent properties of generative AI’s generating process that distinguish it from the others.

Again, let us start with one of OED’s definitions of the term “emergence,” which refers to “[the] process of coming forth, issuing from concealment, obscurity, or confinement” (emergence, n.). Such a definition reveals an emergent entity’s characteristics of being a black

¹ Opposed to *unsupervised learning*, where human annotation of the data is required.

box, as we know very little about how it comes into being. Daily use of the phrase “black box” refers to “a flight recorder which may be removed from an aircraft as a discrete unit, esp. in the event of a crash” (“black box, n.”). In the field of science, Latour defined blackboxing as “the way scientific and technical work is made invisible by its own success” (304). More generally speaking, Wharton identified a black box for which “[the] operator [...] controls its input and uses its output, but doesn’t question the means by which the data that is entered is transformed into the result that emerges” (102). However, these previous discussions of black box mainly concern the lack of understanding power of the operator. Wharton recognized black box as a “flabby” metaphor in her account. She was “beset” by them as they “[apply] loosely and personally to objects that are close to [her].” Yet, as she also pointed out, they “might be opened and easily understood by more technologically competent humans” (103-104). In this case, the creator, or a trained operator, can still more or less understand the black box as a whole. Under such a narrative, all AIs are black boxes. Or more generally speaking, all algorithms packed to form outcomes are black boxes. However, when it comes to generative AI, those who are supposed to understand them the most - their creators – might also lose an overall sense of what is happening inside the box. Therefore, to distinguish such an inexplicability of generative AIs from the general use of the phrase black box and to better understand their emergent characteristics, I want to add a prefix to the phrase, recognizing generative AIs as *total* black boxes.

3.1 Generative AI as Total Black Boxes

The Social Dilemma is a 2020 docudrama film aiming at revealing how big technology companies, such as Google, Facebook, and X (formerly known as Twitter), have applied psychological tricks and manipulation techniques to addict users. Within the film, there is a part where former senior employees talked about the algorithms employed by those companies:

JEFF SEIBERT. You are giving the computer the goal state, “I want this outcome,” and then the computer itself is learning how to do it. That’s where the term “machine learning” comes from. And so, every day, it gets slightly better at picking the right posts in the right order so that you spend longer and longer in that product. And no one really understands what they’re doing in order to achieve that goal.

BAILEY RICHARDSON. The algorithm has a mind of its own, so even though a person writes it, it’s written in a way that you kind of build the machine, and then the machine changes itself.

SANDY PARAKILAS. There’s only a handful of people at these companies, at Facebook and Twitter and other companies... There’s only a few people who understand how those systems work, and even they don’t necessarily fully understand what’s gonna happen with a particular piece of content. So, as humans, we’ve almost lost control over these systems. Because they’re controlling, you know, the information that we see, they’re controlling us more than we’re controlling them (00:48:03-00:49:02).

Undoubtedly, as a film targeting the general public, their words are metaphorical for better comprehension. Additionally, we cannot ignore the bit of their exaggerated tones as the film plays a role of warning, both to the users and to those who have created those algorithms. However, their comments on the uncontrollability of the machine learning algorithms reveal the tendency for AIs to become total black boxes. A senior programmer might design the fundamental algorithms, therefore being able to understand the framework of his AI as a whole. Yet, as his AI consumes more and more data, keeping adapting itself based on its fundamental algorithms and the feedback loop, its complexity will eventually surpass any human’s comprehension.

One might counter those employees’ statements by saying: At least humans write those fundamental algorithms, so that we are still able to understand those AIs’ operation logic, therefore, to comprehensively analyze their operations and adjust them based on such analyses. This argument is true regarding the training of what is now called “good old-fashioned AI (GOF AI)” (Haugeland), where AI “encodes knowledge as production rules, if-then-else statements representing the logical steps in algorithmic reasoning” (Lee 6). While a common understanding of the term “algorithm” refers to “a procedure or set of rules used in calculation and problem-solving,” (algorithm, n.) Lee emphasized its characteristics of being sequential when employed in computers: it is a “step-by-step [procedure] [or rule set] where each step is justified

using explicable rules of logic” (1). Therefore, by stating GOFAI can be algorithmically reasoned, Lee means that humans are capable of understanding each logical step of a GOFAI’s operation sequentially. The outcome of a GOFAI’s operation results from the accumulation of explicable rules: if A happens, then execute B, and so forth. In a sense, the operation of a GOFAI is a step-by-step conditioning process.

The operation of a generative AI is another story. Current generative AIs are all running on a machine learning method called deep neural network (DNN). While a DNN can be called an algorithm under the common understanding, it operates itself not on an algorithmic reasoning logic. That is to say, “there is no sequence of logical steps” for a human to classify (2). Referring back to the first logical model of neurons created by McCulloch and Pitts, within the same paper, they also proposed the concept of a neural network (NN), where multiple neurons connected to each other to form a global output (see fig. 2.). On one hand, their NN constructs the fundamental logic of DNN: through particular weights and thresholds under particular network layouts or rules to reproduce “observed features of human neuroanatomy and cognition” (Hardesty). On the other hand, their NN model was not a *deep* one as 1) each neuron only has a binary threshold with a binary weight², and 2) their NNs were not arranged into layers. In a sense, their NNs can still be algorithmically reasoned. The DNN model came into being when researchers figured out how to modify weights and thresholds that “were efficient enough for networks with more than one layer” (Hardesty). Furthermore, the model became a method when researchers started specifying the training mechanism of those NNs. Therefore, the disorder of DNN, as a method, originates

² Briefly speaking, a threshold can be recognized as a determinator of the final output. In the case of McCulloch and Pitts, a neuron receives multiple binary values, represented as either 0 or 1. Each input comes with a binary weight, either excitatory (positive) or inhibitory (negative). The threshold determines the neuron’s sensitivity to its inputs. If the weighted sum exceeds the threshold value, the neuron produces an output of 1; otherwise, it produces an output of 0.

from two causes: 1) the complexity of its operation process, caused by the accumulation of parameters (weights) and layers³, and 2) its corresponding training mechanism.

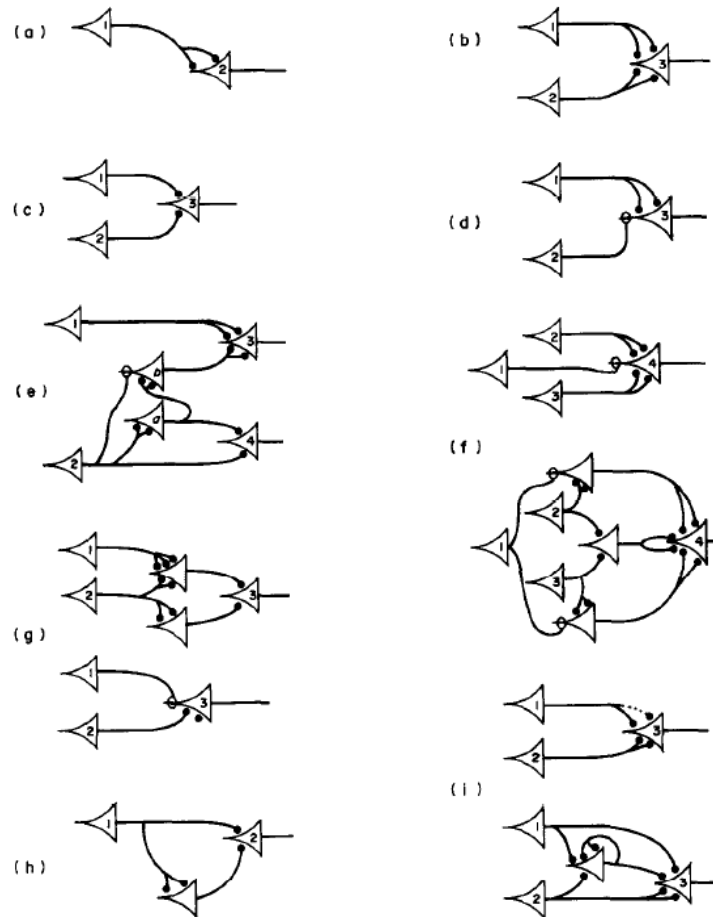


Figure 2: The Initial Illustration of Neural Networks. McCulloch, Warren S., and Walter Pitts. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *The Bulletin of Mathematical Biophysics*, vol. 5, no. 4, Dec. 1943, pp. 115–33. DOI.org (Crossref), <https://doi.org/10.1007/BF02478259>.

While the first cause follows the same logic as those earlier mentioned algorithms employed by big technology companies, I will explain the second cause based on the training of

³ The longest path of ChatGPT contains about 400 (core) layers, "with a total of 175 billion connections and therefore 175 billion weights" (Wolfram).

AlphaGo. Developed by Alphabet Inc. (formerly known as DeepMind Technologies), AlphaGo is an AI rooted in DNN. The significance of AlphaGo originates from its crack of one of the humans' oldest and hardest games - Go. "The game is a googol times more complex than chess - with an astonishing 10^{170} possible board configurations," which is "more than the number of atoms in the known universe" ("AlphaGo"). The programming team trained AlphaGo in two stages. They first introduced a "policy network" by *supervised learning* from human expert moves – about 30 million moves in total. Then, they trained a "value network" by *unsupervised learning* from "games played by the policy network against itself" (Silver et al. 2). My argumentation focuses on the second stage of AlphaGo's training mechanism. Within this stage, how the programming team trained AlphaGo was similar to the training of a Go Master – "by practicing" (Lee 2). That is to say, AlphaGo was not merely trained on existing data sets, records of those human expert moves. More importantly, once it reached a certain level of proficiency, its contained policy networks, or in another phrase, its lower-level AIs, were forced to play against each other so as to generate new collections of moves as new data. In a sense, within this stage, data was created but not mined (Lee 2). Therefore, the question is, how can we trace the algorithmic reasoning in the operation process of AlphaGo's second stage, when we train it as not sequentially in the first place?

Lee used the term "interaction" to represent the learning process of humans, as well as the operation process that happened within such DNN-trained AIs. Analogically speaking, "a human will never acquire the ability to outperform Go Masters by just watching masters playing Go." Instead, he/she has to interact with other players, especially those Go Masters, to become a Go Master (2). That is to say, algorithmic reasoning logic is merely one form of mechanized human rationality. For better clarification, I call such a power to algorithmic reason humans' computational rationality: "A rational process is step-by-step reasoning using clearly explicable

rules of logic” (1). Undoubtedly, as humans, we also solve problems and understand things through non-sequential reasoning, as well as through nonrational means, such as intuition, emotion, and imagination. However, when it comes to our operation of digital computers, or more generally speaking, computations, our primary approach has traditionally been such a stepwise reasoning process. Based on the argumentation from Alan Turing and Alonzo Church, Lee defined the term “computation” as:

- (a) algorithmic (consisting of a sequence of discrete steps, where each step is drawn from a finite set of possible operations);
- (b) terminating;
- (c) operating on discrete data (the inputs, outputs, and intermediate states are all drawn from countable sets);
- (d) non-interactive (inputs are available at the start and outputs at termination) (4).

Precisely because of how we design digital computers to follow such a definition in the first place, our understanding of any digital system they produce has been limited to algorithmic reasoning, with clear input data and output results⁴. Therefore, when we try to apply an interactive training mechanism to train AIs like AlphaGo, even the operators lose the ability to fully understand their operation processes.

How do those new collections of moves come into being? What leads an AI to generate those specific types of new data? The operators understand the operational logic of each layer of their DNNs, as well as the operation logic among layers. In a simple DNN, they might also understand the meaning of those parameters they set. That is to say, they know how the outcome will change if they modify certain parameters. Yet, when trying to build AIs with complex DNNs, they start to lose their understanding of the meanings of those parameters. The problem of generative AIs becoming total black boxes originates from the goal of the operators. In this case,

⁴ That is to say, our standard of whether a digital system is understood has been constrained to whether it has been stepwise reasoned or not. Therefore, when I only use the term “understand,” without any epithets, for any digital system, including its digital components (subsystems), in later sections, I mean the behavior of stepwise/algorithmic reasoning.

the operators only care whether their generative AIs, as ANIs, can solve the targeted problems or successfully process the targeted tasks. That is to say, such a way of creating generative AIs is fundamentally a goal-driven process. The operators will not recognize it as a problem when they start to lose their understanding of the meanings of their parameters. As long as they know what parameters can lead to the best outcome, they fulfill their goals.

A generative AI is a total black box when the operators can only understand it mathematically, yet the outcome for the users is a qualitative change. As a result, AlphaGo “defeated a human Go world champion a decade before experts thought possible” (“AlphaGo”). ChatGPT-4 can draft original and sometimes creative responses to human requests. Such an accumulation of quantified changes (by modifying parameters and layers) to qualitative results reveals the essence of generative AIs being emergent. The concept of emergence has been long discussed in various domains of science. The title of P. W. Anderson’s well-known paper can best summarize such a concept – “more is different.” Published in 1972, Anderson proposed a theory of “broken symmetry” to better break down the “constructionist converse of reductionism” in the field of science (393). He argued that the hierarchy of science is not simply a reductionist ladder, where understanding at one level can fully explain the behavior at another. Instead, there is the emergence of new principles at each level. Correspondingly, though not explicitly mentioning the term, Anderson introduced such a concept of emergence to understand complex systems in science, such as many-body physics in his case. New phenomena and behaviors will emerge at each level of our understanding of a complex system, which cannot be predicted or fully explained by the laws governing the lower levels. His theory of broken symmetry precisely points to this shift “from quantitative to qualitative differentiation –” the more complexities, the more differences.

Following Anderson's narrative, we are living in a world of emergent entities. From an ammonia molecule to the universe as a whole, they are everywhere around us. Therefore, it is necessary to distinguish generative AIs' emergent properties from those prior to our construction, as generative AIs are fundamentally created by us, but not emerging from nature. Generative AIs are emergent because of our training mechanism. Reflecting on AlphaGo, the programming team trained it by "exploiting interactions among primitive components" (Forrest 1). Previous training mechanisms, on the one hand, follow the prevailing methodology of computing, which Stephanie Forrest defined as being "parallel:" the process of breaking down larger problems into smaller, independent, often similar parts that can be executed simultaneously by multiple processors, communicating via a shared memory. Such an interactive training mechanism, on the other hand, tries to improve efficiency and increase the flexibility of AIs by offering them a "more natural representation" (Forrest 1), namely, how our brains function based on the observed features of human neuroanatomy and cognition. Therefore, the quantitative to qualitative differentiation happens when we shift our focus from algorithmic reasoning the operating process of AIs, to the interactions among their lower-level agents (variables, neurons, physical units, etc.). We still program those agents to employ stepwise reasons, each following explicit instructions. Yet, their interactions with each other form "implicit global patterns at the macroscopic level" (Forrest 2).

Based on these distinguishments, hence, I conclude my definition of generative AI. It is a human-made emergent machine (digital systems based on hardware), where accumulated quantification has led to qualitative changes (through agents' interactions), emerging new global behaviors that exceed our understanding of its agents. Its emergent properties come from how we train them interactively in the first place, an approach that reflects our current understanding of human brains, aiming to surpass the limitations of algorithmic reasoning, so as to build more powerful AIs.

3.2 Opening up the Black Box of Generative AIs

Many critics take a negative perspective on generative AIs. Within the open letter, Elon Musk, along with hundreds of big names in tech, urged a six-month stop from developing them, as “no one – not even their creators – can understand, predict, or reliably control” (“Pause Giant AI Experiments”). I hold a different opinion as I recognize their worries mainly focusing on generative AIs being a total black box, rather than their emergent properties. That is to say, those opponents have been concerned that an incomplete understanding of generative AIs might bring unexpected, and even terrible changes to our society. They do not deny our current training mechanism, as it has to be proven fruitful in constructing AIs. Therefore, they spurred to “jointly develop and implement a set of shared safety protocols for advanced AI design and development that are rigorously audited and overseen by independent outside experts” (“Pause Giant AI Experiments”). Combining their concern with my argumentation, I suggest an unlimited development of generative AIs, yet with authored supervision and clear regulation rules when released for public use. In a sense, within the laboratory, we should not change our training mechanism as generative AIs’ emergent properties might bring us new knowledge.

My treatment of a black box differs from that of scholars like Annabel Wharton. While they recognize it as a problem to solve, I consider it a possibility to explore. I hold the same belief in the black box compared to that of Ranulph Glanville: “[It] is such a powerful device, that it is time to explore it seriously, in its own right; for it allows us that most magical of tricks, a way of acting confidently with/from the unknown/unknowable” (189). By saying opening up the black box of generative AIs, I am not referring to algorithmically reasoning it, so as to make its operation process interpretable to humans. Instead, I suggest recognizing what emerges from our construction of generative AIs as new varieties of emergence – what comes forth without clear patterns for humans to define –to inform us of the unknown/unknowable by studying what leads

to their emergent properties. It is a way of opening up the black box as that newly learned knowledge might help us understand them better.

Many scholars have advised the development of interpretable machine learning algorithms to solve AI's black box problem. Cynthia Rudin and Joanna Radin defined interpretable AI models as those that “provide a technically equivalent, but possibly more ethical alternative to black box models [...] – they are constrained to provide a better understanding of how predictions are made.” Undoubtedly, interpretable AIs are better for public usage as they can be fully understood and controlled. Yet, how to define interpretable AIs as being technically equivalent needs further elaboration. To prove such an argument, Rudin and Radin mainly relied on the equivalent, and in some situations, better technical efficiency of an interpretable AI model called Certifiably Optimal Rule Lists (CORELS) for criminal risk assessment. According to the developing team⁵, while testing CORELS on “a number of publicly available data sets,” it is “able to achieve better or similar out-of-sample accuracy on these data sets compared to the popular greedy algorithms⁶” (Angelino et al. 3). More importantly, when testing CORELS on the ProPublica datasets of Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)⁷, it is “as accurate as COMPAS” and is much easier to debate about its fairness as a completely transparent model (42).

However, CORELS is not a generative AI; neither are the machine learning algorithms it was compared to. CORELS was developed based on “rule lists,” “predictive models composed of if-then statements.” It was designed following algorithmic reasoning logic. It is interpretable

⁵ Rudin was one of the members of CORELS' developing team.

⁶ Greedy algorithm is “an algorithm that always takes the best immediate, or local, solution while finding an answer” (Black, “greedy algorithm”)

⁷ COMPAS is a widely used criminal risk assessment tool in the United States. It has been used to “access more than 1 million offenders” since its development in 1998 (Dressel and Farid). ProPublica acquired two years' worth of COMPAS scores from the Broward County Sheriff's Office in Florida through a public records request. The data encompasses all 18,610 individuals who were scored in 2013 and 2014 (Larson et al.).

because those explicable rules “provide a reason for each prediction” (Angelino et al. 1). Though the algorithm of COMPAS is commercially confidential, a test conducted by Julia Dressell and Harry Farid showed that it is “equivalent to a simple linear classifier.” In a sense, “despite the impressive sounding use of 137 features,” a linear classifier “based on only 2 features – age and total number of previous convictions – is all that is required to yield the same prediction accuracy as COMPAS.”

The urge for an interpretable AI from Rudin and Radin focuses on better regulations when releasing AI for public usage, especially when those AIs play an essential role in “socially-important decision-making” (Angelino et al. 1). The concept of an interpretable AI is valuable as a person has the right to not only know what variables lead to the results of his criminal risk assessment, but also understand the weights of those variables, as well as how the algorithms process those weights. Especially when COMPAS has been verified to have racial biases⁸, an interpretable AI is necessary in the criminal justice system. However, when constraining AI to provide a better understanding of how predictions are made, it also limits its potential to solve tasks that can otherwise be stepwise reasoned, namely, tasks that can be interpretable for humans in the first place. This can be reflected in the limitations of CORELS. Angelino et al. mentioned that it could have difficulty dealing with “problems with many possibly relevant features that are highly correlated.” Additionally, CORELS is not designed for problems “where the features themselves are not interpretable,” such as raw image processing (Angelino et al. 57).

Constraining and clarifying clear variables is crucial for training AIs under an algorithmic reasoning logic, as we do not want them to learn patterns that are unnecessary for the tasks we

⁸ For example, COMPAS’ predictions frequently indicated that Black defendants have a higher risk of recidivism than they actually exhibited. Research from Larson et al. showed that “black defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts (45 percent vs. 23 percent).”

aim to process. Yet, we also limit our understanding of those tasks to those variables we set.

Ultimately, no matter how a criminal risk assessment AI can be, what causes a person to conduct or re-conduct a crime can certainly be more than our abstracted variables.

The disadvantage of interpretably developing a generative AI, following the narrative of Rudin and Radin, relies on the goal of interpretable AIs. The design of CORELS was aimed at the optimization of existing machine learning algorithms: “[finding] a transparent model that is optimal within a particular predetermined class of models and produces a certificate of its optimality, with respect to the regularized empirical risk” (Angelino et al. 2). That is to say, CORELS was designed for a better solution to an already interpreted task. In their case, a better solution means a more accurate prediction and a more transparent operation process. Yet, the construction of generative AIs attempts to find *new* solutions to previously interpreted, stepwise reasoned tasks, or tasks that are not interpretable to humans in the first place. As illustrated earlier in the example of AlphaGo, only through emergent computing, forcing policy networks (interpretable lower-level agents) to interact with each other to emerge new collections of moves (new global behaviors), can AlphaGo be empowered to defeat a Go Master. How can we interpretably develop a generative AI when it works primarily from the interaction of its training data, rather than explicable rules?

One approach to developing interpretable generative AIs was making them learn algorithmic reasoning, “where the results of the training phrase is a set of explicable production rules” (Lee 7). However, Lee also emphasized that these interpretable generative AIs have proven to underperform those being trained through DNN:

Wilson et al. created a program that could write programs to play old Atari video games credibly well. Their program generated random mutations of production rules, and then simulated natural selection. Their technique was based on earlier work that evolved programs to develop certain image processing functions (Miller and Thomson). The Atari game-playing programs that emerge, however, are far less effective than programs based

on DNNs. Wilson et al. admit this, saying that the main advantage of their technique is that the resulting programs are more explainable. The learned production rules provide the explanations (7).

Therefore, based on my argumentation of interpretable AIs and Lee's emphasis, I first suggest an unlimited development of generative AIs in the laboratory, yet under authored supervision. Reflecting on the open letter mentioned above, clear supervision rules must be established to avoid them being totally out of control. Under what conditions can we develop an unlimited generative AI? What should be clearly documented to evaluate its development stages? Who should jointly supervise its progress? Clarifying these types of questions can be a good start. When released to the public, clear regulation rules need to be established, and the interpretability of a corresponding generative AI needs to be clearly defined: which parts are interpretable to whom and to what degree those parts are interpretable. Reflecting on the goal of Rudin and Radin, when a generative AI is planned to be used for critical social decisions, it must first be transformed into an interpretable one for transparency and justice.

Undoubtedly, as a paper fundamentally emphasizing the potential of generative AIs, my discussion of its possible authored supervision is only an initial theoretical assumption that needs further and more explicit elaboration from various scholars and practitioners in the field. Yet, my argument is that, if such an assumption is qualified, the emergent properties of generative AIs might bring us new knowledge. Starting from AlphaGo, reversely studying what causes its new collections of moves might offer us new strategies in playing Go. In a sense, by recognizing what generative AIs emerge as new varieties of emergence and studying the causes of their emergent properties, we can be empowered to know what is previously unknown/unknowable. Meanwhile, such a methodology can also be recognized as a way of opening up the black box of generative AIs. Reversely studying their emergent properties can help us better understand those generative AIs themselves, so as to inform better ways of control.

In a one-day meeting within the Physics Department of the Massachusetts Institute of Technology (MIT), Max Tegmark illustrated such a way of opening AI's black box⁹:

How can we [(physicists)] help? I feel that, first of all, we obviously should. [...] I think we really can help by opening up the black box and getting to a place where we're not just using ever more powerful systems that we don't understand, but where we are instead able to understand them better. This has always been the tradition in physics when we work with powerful things. If you want to get a rocket to the moon, you don't just treat it as a black box and you fire: That one went a little too far to the left. Let's aim a little farther to the right next time. No. What you do is you figure the laws of ... You figure out Einstein's laws of gravitation. You figure out thermodynamics, et cetera. And then you can be more confident you're going to control what you do (03:24 - 04:20).

Named Mechanistic Interpretability (MI) for AI Safety and Control, this field of study, led by scholars like Max Tegmark, aims to keep AIs under control by MI. While the interpretability of interpretable AIs emphasizes stepwise explicable rules, MI focuses on understanding parameters and operations acting on those parameters. It can be recognized as one form of reverse engineering, where we "break neural network activations into independently understandable pieces" (Olah) to form a coherent picture of the whole. Tegmark called such a breaking down process "extracting learned knowledge from the black box in the MI spirit" (07:57 - 08:02).

Undoubtedly, building a reductionist ladder on generative AIs is not an easy task. For example, neurons might correspond to features that humans cannot interpret in the first place. On the one hand, adversarial examples in the training of DNNs can be explained as "specialized inputs created with the purpose of confusing a neural network, resulting in the misclassification of a given input" ("Adversarial Example Using FGSM"). On the other hand, they can also be recognized as the results of "our model's sensitivity to well-generalizing features in the data," yet

⁹ Tegmark's use of the term AI more likely referred to one form of generative AI – the large language model (which will be further explained in section 4). This can be reflected in his examples of AI and the title of the meeting, which is "The Impact of ChatGPT and Other Large Language Models on Physics Research and Education."

incomprehensible to humans (Ilyas 1). Nevertheless, reflecting back on Anderson's arguments, MI's philosophy more or less corresponds to the constructionist converse of reductionism in science. As an approach that has proven to be fruitful in understanding complex systems, such as how to send a rocket to the moon, implementing MI might be a good start. In the end, as Tegmark emphasized, if we can "use AI to actually mechanistically extract out the knowledge that is learned," we can then "re-implement it in some other kind of architecture which is not a neural network," creating interpretable AIs that is potentially smarter than us, yet under our control (07:27 - 07:51).

4. Generative AI as Models

Following my discussion on generative AI's potential to bring us new knowledge, I argue that it also has the ability to change the mode of producing it. To understand generative AI's relation to knowledge production, I recognize them as models. Programmers often use this term to abstract their generative AIs. For example, ChatGPT has been defined as a large language model (LLM)¹. However, their use of this term merely focuses on classifying what they have created into either an existing or a new domain. That is to say, they name their generative AIs as models to distinguish their functionalities – what they can do or what tasks they are designed for. The name LLM clearly shows that ChatGPT is an AI targeting processing language in a large amount. My recognition of generative AIs as models differs from their creators, as I focus on a more general and interdisciplinary meaning of our use of models in knowledge production. In a sense, models have always been the only medium for humans to produce knowledge. Under such a narrative, generative AIs are hardly new. However, if we empower generative AIs to construct new models themselves; we might achieve new pathways for knowledge production. To prove my arguments, I will start by defining what a model is.

4.1 *What is a Model?*

Typical examples of models in a daily life setting are three-dimensional replicas, either physical or digital, “more or less ‘true to scale,’ of some existing or imagined material object” (Black, “Models and Archetypes” 219). The Oxford English Dictionary (OED) provides three primary definitions: 1) “A representation of structure, and related scenes.” 2) “An object of imitation.” 3) “A type or design” (“model, n. & adj.”). Despite the simplicity of these definitions,

¹ According to Nvidia, a leading company in AI computing, LLMs are “deep learning algorithms that can recognize, summarize, translate, predict, and generate content using very large datasets” (“Large Language Models Explained”).

the term is more complicated when used in different contexts. For example, an analogue model of deoxyribonucleic acid (DNA) might both satisfy all three definitions proposed by OED and go beyond. Imagining a physical double-helix DNA model placed in a biology laboratory, it can be interpreted not simply as a representation (of the DNA structure), an imitation (of human DNA or of the original analogue one created by James Waston and Francis Crick) (“Francis Crick”), and a type or design (of all human DNA). If we recognize such a model as a representation of the DNA structure, then more important is the information or knowledge it conveys, which, historically speaking, “marked a milestone in the history of science and gave rise to modern molecular biology” (“The Discovery of Double Helix”). One might counter this view by saying his understanding is more related to the latter two, and there is a distinction between this model and the original one which changed history. Yet the key here is not to argue an accurate and definite definition of such a particular model set in the laboratory. It is to demonstrate our pluralistic use of the term “model,” by showing how the exemplary DNA model can cover meanings that exceed OED’s definitions, as well as possible debates on determining those meanings.

The same situation happens when we try to define AIs as models. Following OED’s definition, ChatGPT, as a model, can be recognized to be 1) a representation of AIs, 2) an imitation of humans, as it aims at mimicking what humans can do especially in relation to language, and 3) a type or a design, to process language in a large amount. Clearly, programmers’ use of this term falls into the third category. However, under such a narrative, we also limit the knowledge it can produce merely to the field of building AIs. Therefore, to better understand generative AIs as models, it is necessary to first investigate the term from a more general and interdisciplinary perspective. Finding a meta meaning of our use of models can help us better understand generative AIs’ potential for knowledge production in a broader sense.

4.2 Black's Definitions of Models

In 1962, Max Black proposed four types of models to understand the “presuppositions and the implications of [models’] practice” (“Models and Archetypes” 219). 1) “Scale models” tend to cover “all likeness of material objects, systems, or processes, whether real or imaginary, that preserve relative proportions.” They are asymmetrically relational to the original, such as miniatures (220). 2) “Analogue models” refer to “material object, system, or process designed to reproduce as faithfully as possible in some new medium the *structure* or *web* of relationships in an original.” While scale models rely heavily upon “identity,” functioning more as imitations, analogue models focus more on “isomorphism.” The versatility of abstract structure makes it suitable for a vast array of content, thereby offering limitless potential for constructing analogue models (222-223). 3) “Mathematical models,” as its name implies, are mathematical treatments “to extrapolate to testable consequences in the original field” (224-225). Though Black distinguished mathematical models from analogue models, the former will still be categorized under the latter in this paper. The reason is not only for more straightforward classification, but also because mathematical treatments can indeed be treated as certain types of abstract structure of the original consequences they aim to address, even though sometimes “ethereal” (223). 4) “Theoretical models” are hypothetical entities proposed mainly in the scientific field. Unlike scale models and analogue models, theoretical models might lack known referents. Black used the modeling of ether as an example to illustrate the function of theoretical models. To describe ether “as it is,” the aim is not to verify the existence of such a hypothetical entity itself, but to “[talk] in a certain way.” Utilizing theoretical models attempts to introduce “a new language or dialect, suggested by a familiar theory but extended to a new domain of application.” The dynamic behind such an attempt is a need, which seeks “further scientific mastery of the original domain” (228-230). Though the physics community started to deny the existence of ether with the

publication of Einstein's Special Theory of Relativity in 1905, ether theory had been dominant in physician's study of light and encouraged much research (e.g., Boyle; Newton; Young), so as to expand the scientific understanding of the field ("Searching for Light's 'Ether'"); thus demonstrating the power of theoretical models.

4.3 As If vs. As It Is

While not explicitly stated, it is evident that Black's definitions of models are rooted in his interpretations of metaphors. Scale models are "substitution[s]," where metaphorical expressions are "used in place of some equivalent *literal* expression[s]" ("Metaphor" 279). Analogue models refer to "comparison," which "consists in the *presentation* of the [original's] underlying analogy or similarity" (283). Theoretical models emphasize "interaction," where new meanings are produced through the interplay between the principal subject (the original field) and the subsidiary subject (the metaphorical field) (286-287). While the significance of recognizing models as metaphors will be further addressed in later paragraphs, here the emphasis is on how Black himself was creating theoretical models, applying his theory of metaphor to the domain of models, so as to expand our understanding of the term. When presenting the example of ether, Black distinguished two ways of thinking towards the use of it, in which he stated:

There is certainly a vast difference between treating the ether as a mere heuristic convenience, as Maxwell's first remarks require, and treating it in Kelvin's fashion as "real matter" having definite - though, to be sure, paradoxical - properties independent of our imagination. The difference is between thinking of the electrical field *as if it were* filled with a material medium, and thinking of it *as being* such a medium. One approach uses a detached comparison reminiscent of simile and argument from analogy; the other requires an identification typical of metaphor ("Models and Archetypes" 228).

Black's treatment of theoretical models clearly falls into the second category. Under his narrative, a theoretical model acts as it is meant to be, just like a metaphor. As Black emphasized "the metaphor itself neither needs nor invites explanation and paraphrase," his theoretical model of

theoretical models focuses on expanding the original domain as mentioned above, which, in this case, is the domain of models.

On the contrary, authors like Morgan and Morrison took the first perspective, treating models as bare heuristic convenience, as if they were qualified. While theoretical models in Black's arguments are the theories, Morgan and Morrison argued that we use models as "instruments" to build theories (18). Black highlighted Maxwell's insistence, whereas the price paid for the "as if" thinking, is the "absence of explanatory power" (228). Correspondingly, Morgan and Morrison argued that "we do not learn much from looking at a model - we learn more from building the model and from manipulating it" (11-12). Such a methodology is prevailing in our current treatment of constructing generative AIs. As mentioned in my discussion of AGI, scholars who believe in its existence have argued their capacity to provide insights into cognition through the approach of "understanding by building" (Pfeifer, Rolf, et al. 21). Their reasoning follows that by comprehending the design and construction of intelligent embodied systems, we can build a better understanding of intelligence in general. However, precisely because the original domain lacks clear understanding in the first place, the methodology of learning by building and manipulating can also be reflected in their own methods. For example, despite the examples Morgan and Morrison used, such a methodology can also be found in their own methods towards their study of models, as the original domain. By building a model of models' four basic elements ("construction," "functioning," "representing," and "learning"), and manipulating such a model with different contexts, Morgan and Morrison demonstrated their arguments not only by argumentations within the paper but also by the paper itself.

In this case, regardless of their differences, Black, as well as Morgan and Morrison, were all trying to "model" models. That is to say, their methods of approaching models precisely followed their analysis of how models function. Additionally, to view models as metaphors or

instruments is analogical (the analogue model) and hypothetical (the theoretical model). Both terms can be regarded as symbolic representations that function to “[reproduce] the structure” of models, which, in a sense, can never be fully understood and copied (Black 222). By explaining these terms’ formal uses and to what degree they can be applied to explain models, Black, Morgan, and Morrison simplified models to concepts from other fields that the audience would find familiar. In the end, Black, as well as Morgan and Morrison, did not walk out of what I call the “hierarchy of models.” A hierarchical modeling process can be found within their argumentation, from the research topic (the modeling of models), to detailed structures of specific subtopics (the modeling of sub-models), so and so forth, until reaching concepts that require no further explanation, either analogical or not. As for the former, Black defined such a system of concepts, which is “used analogically” but with “no question of a definite explanation of given phenomena or laws,” as “conceptual archetypes,” or more briefly, archetypes. Again, the significance of archetypes will be discussed later, and our focus here is how Black, Morgan, Morrison, and even OED (by using terms like representation, imitation, and design to represent a model) heavily relied on creating models, so as to model the use of models.

4.4 A Meta-Model of Models

Not only for producing knowledge in understanding models, the same hierarchy of models can be found in any kind of human knowledge production. Reflecting back on Anderson’s arguments, the constructionist converse of reductionism in the field of science can certainly be recognized as a way of building hierarchy of models. For AI construction, Stephen Wolfram emphasized the inexistence of a “model-less model.” Every model possesses “some particular underlying structure,” meaning that every model incorporates internal sub-models (*What Is ChatGPT Doing*). Therefore, the first meta-meaning of models relies on our hierarchical use of

them. However, such a meta-model contains one pre-assumption: everything can be recognized as a model in the first place.

In 1979, Wartofsky stated that anything is a model as “[anything] (in the strongest and most unqualified sense of anything) can be a representation of anything else.” Though not explicitly explained, Wartofsky’s use of the term “representation” was literal, referring to “the action of standing for, in the place of, a person, group, or thing, and related senses” (“representation, n.¹”). Seeming implausible, yet Wartofsky’s statement has two presumptions: 1) “There are no intrinsic or relational properties which mark one thing off as a representation of something else; or everything has infinitely many properties in common with everything else.” That is to say, no matter how unrelated two objects seem, similarities can always be found under certain perspectives. A model-referent relationship can then be built as long as their similar properties are asymmetrical. 2) “It is *we* who constitute something as a representation of something else.” In other words, once the correlations are found, a representation is “taken to be one,” under our will. Nothing is a representation unless “we make it, or take it to be” (20-21).

Compared to Black, Morgan, and Morrison, Wartofsky’s treatment of models was more subjective. Rather than stating that models can be categorized into specific types or can serve particular purposes, as if models are prior to human intervention, Wartofsky emphasized models being “deliberately constructed representational artifacts” (28). In this case, the second meta-meaning of models is that they are all either physically or metaphysically created by us, by purposes and to serve purposes. On the one hand, a model can be argued to be a metaphor, an instrument, or “an autonomous agent” (Wharton), thus empowering us to analyze its usage and the corresponding impact afterward. On the other hand, such a meta-meaning of models, following Wartofsky’s statements, permits us to question the reason for creating models in the first place.

4.5 From Models to Knowledge

Why models? Wartofsky argued that “in science, as in much of art, human knowledge is achieved by means of representation” (25). In the field of science, using anatomical dissection as an example, as Catherine Waldby mentioned, “to anatomize is to analyze, to partition, to reduce a whole to its constituent parts.” It is by simplifying, systemizing, and translating the human body into representations like “graphs, indices, preserved body fragments, micrographs,” that we form coherent understandings of the human body as a whole (55-56).

As for the field of art, David Novitz proposed three basic types of knowledge claims. The first refers to the knowledge acquisition of the artwork itself, “just by reading and interpreting a novel or a poem, or by viewing and constructing a painting” (990). A reader can claim to know about certain relationships between characters within a novel or poem; or he can claim to know what techniques, such as glazing, stippling, or dry brushing (Thomas), the painter used to create his artwork. The second knowledge claim in art concerns the “appropriate emotional responses” to a work of art. Certain emotions would normally be raised when we face or try to understand an artwork. While which one is appropriate could be a question worth discussing, such emotional responses suggest that there is essentially something within the artwork that deserves responding to. The third knowledge claim highlights “the aspects of the world external to the work.” A painting might give a certain degree of meaning to one’s life and a novel might convey insights for a reader to understand things in the real world. Overall, in the first and the third circumstances, the artwork itself is the representation for understanding, as appropriate emotional responses in the second act the same.

In a sense, what Wartofsky argued is not simply that we create theoretical models for further scientific mastery or that we build models to learn. Under his narrative, representations are the foundation for human knowledge production. Language, text, and “the play of signs”

(Nietzsche, Vol. II, 180), following Friedrich Nietzsche and Michel Foucault, can all be seen as means of them for us to know and understand. Nietzsche and Foucault stated that our claims to truth and knowledge can be “deconstructed,” essentially not rooted in the inherent nature of things but depends on the drive for control and power behind, the desire to prioritize our selected “perspectives,” or in this case, representations, as they serve us well. Correspondingly, the second layer of meaning contained in Wartofsky’s statements of representations being the foundation of knowledge is that these representations are not simply models for use. Referring to his statements about models being human created, these representations also represent “the mode of activity in which they are used, or the mode of their own production” (23). Hence, models for use are also models of our own cognition. We create “cognitive artifacts,” namely, models, as “representations to ourselves of what we do, of what we want, and of what we hope for.”

Following such an argument, the question is, are we limiting our ability to know and to understand within the models we create? On the one hand, how we know and understand, whether it is ourselves, the world, or the things around us, are essentially conditioned by our “biologically evolved and genetically inherited modes of perceptual and cognitive activity” (Wartofsky 25). It has been proven that different animals can have different perceptions of the world, such as a dog seeing the world like a person who is red-green color-blind (Ling). That is to say, our innate physiological structure has already shaped the scope of our sentience and rationality. On the other hand, the models we create are designed initially to go beyond that limitation. A thermometer can quantify the temperature so that we can decide what to wear for a day based on a number before going out and feeling it through our bodies. Further up, a computational climate model can be fed with pure data to predict weather up to centuries (Lyon), though sometimes the results are inaccurate. Undoubtedly, we live in a world of models. Models are everywhere, from a toy car in the store to the theories I am quoting in this paper. Recognizing

models as our cognitive artifacts, in a sense, we indeed limit our ability to know and understand within those models. However, the initial goal of creating them is to help us surpass our inherent limits, thus achieving goals that we can never attain by our own bodies. In many situations, those models can reversely guide us to better understand ourselves, therefore even helping us to break some of the limitations. Recognizing the building of generative AIs as a way for us to learn cognition is one of these situations.

In all, as Wartofsky emphasized models “[altering] the very nature of learning,” a model of the history of human knowledge production we can propose here is that, to understand is to model, and to claim knowledge is to find good models. As the former has been illustrated in previous paragraphs, for the latter, in this case, Wartofsky recognized models as “proffered truth.” Clearly, “[to] proffer truth” is one of “the human means of acquiring knowledge,” but not every model is a proffered truth (28).

Annabel Wharton used the terms “strong” and “weak” to evaluate the usefulness of models. Firstly, the strength or weakness of a model depends “in part on the status of its referent” (12). An officially authorized model of a Lamborghini sports car might be expensive as a genuine one has already been set at a high price. A model of the human brain holds significance because the human brain is the center of our nervous system. Additionally, a model’s strength or weakness is also determined by the conditions of the relationship between a model and its referent. As Wharton argued, “‘strong’ describes a model that acts as a dominant subject that determines its weak object” (12). A blueprint for a mechanical machine is the archetype for the machine to be built. A new algorithm can be applied to the system to make performance 30% faster than the previous. Under such conditions, weak models are more like “copies,” often “subordinate to their archetypes.” Wharton also emphasized the oscillation of a model’s usefulness. That is to say, in different situations or with different perspectives, a model may “shift

between [its] weak and strong potentials” (12). The Lamborghini toy car is a weak model of a genuine one, yet at the same time, it can be recognized as a strong model, when a teenage boy has saved for months to buy it for his collections.

Applying Wharton’s arguments to the process of knowledge production, a model can certainly be evaluated as being strong or weak. A programmer can always utilize multiple methods when he tries to code for certain effects on a computer. Using JavaScript (JS), a programming language for web design, as an example, to assign some properties to every object within an array, a programmer can do it 1) manually by assigning those properties one by one, 2) through default functions, such as the for-loop, 3) by calling built-in functions within the library he uses, 4) by writing his own customized functions. In this case, a strong model might either be the most concise method, which increases the readability of the programmer’s codes, or be the most efficient one, which saves more power for other computations. Such uncertainty also reveals the oscillation of a model being strong or weak, referring back to Wharton’s arguments.

However, even if the methods that the programmer uses can be evaluated according to the strong-or-weak measurement, we cannot deny that they are all helpful in terms of the results of knowledge production. Every method can lead to the programmer's desired effects, though with different costs. As Wharton focused on the relationship between the model and its referent, as well as the conditions of such a relationship, when measuring the usefulness of a model based on the consequences of its usage, the value of the knowledge it produces, the terms “good” or “bad” might be added into our discussion. For scientific knowledge production, a good model is a promising one, whether being tested as a good fit for a particular problem or domain, or being proven to be fruitful in discoveries, with “implications rich enough to suggest [new] novel hypotheses and speculations” (Black 233). Under such a narrative, a model is bad when it 1) cannot solve the targeted problem or even crash the original environment, such as a computer

system or a petri dish, 2) or is recognized as less efficient or less beneficial for the future development of a specific research or project. The programmer might use a JS library, which contains pre-built functions designed by others, to save time. He can then add other functions from that library, some of which he probably cannot write by himself. However, by doing so, he might also limit himself within that library, thus preventing him from developing more customized and more creative functions in the future.

Again, oscillation happens when labeling a model as good or bad. It is not simply because there are different situations. Being deliberately constructed, a model's value fundamentally depends on our perspectives, whether they correspond to what we do, satisfy what we want, or have the potential to achieve what we hope for. In a sense, back to the arguments of Nietzsche, Foucault, and Wartofsky, a relatively good model is one that serves us well, often during a specific period and under certain circumstances. The Ether theory was a good model before the Special Theory of Relativity as it permitted the scientists to answer their theoretical questions and propose more hypotheses for further scientific mastery. In the field of science, such a need fundamentally pushes the construction of scientific models, leading to the standard mentioned above, under which a good scientific model has to be a promising one, in order to fulfill this need.

4.5 Models within the Circular Process of Knowledge Production

To emphasize our dominant role in terms of the creation and interaction with models is to demonstrate we, as the observer, with our interference, cannot be excluded from the process of knowledge production. As Edward Lee emphasized:

In the prevailing philosophy of science, observation trumps interaction. We are taught that the best science is objective, not subjective. Let the data speak for itself. Design your instruments to minimally disrupt what you are observing. But science also teaches us that observation without interaction is impossible (2).

In a sense, Wartofsky's and Lee's arguments both correspond to the philosophy of second-order cybernetics, where von Foerster emphasized observing systems instead of observed ones (first-order cybernetics). Under his narrative, the environment of observing systems "contains no information; [it] is as it is" (259). There is nothing without the "properties of the observer," namely, to observe and to describe (289). The existence of an observer is prior to the distinction of a system. It is the observer who "enters the system" and is "allowed to stipulate his own purpose." An observer is "autonomous" (286). Therefore, the production of knowledge is a subjective experience rather than an external objective reality, an *adaptive function* that "cognitive efforts have the purpose of helping us cope in the world of experience, rather than the traditional goal of furnishing an 'objective' representation of a world as it might 'exist' apart from us and our experience" (von Glasersfeld 24).

Scholars of second-order cybernetics (e.g., Margaret Mead; Ranulph Glanville; Gordon Pask) highlighted this epistemology of the observer – the feedback and "circular causality" within the process of knowledge production (Von Foerster, *Cybernetics: circular causal and feedback mechanisms*). The figure created by Bernard Scott can best summarize their arguments (See fig. 3). It contains three parts. Firstly, an observer begins with the "first-order study of observed systems." However, secondly, being one "organizationally closed" system, his observations are not direct ones of a "reality" but "constructions based on particular sets of assumptions" (1374-1375).

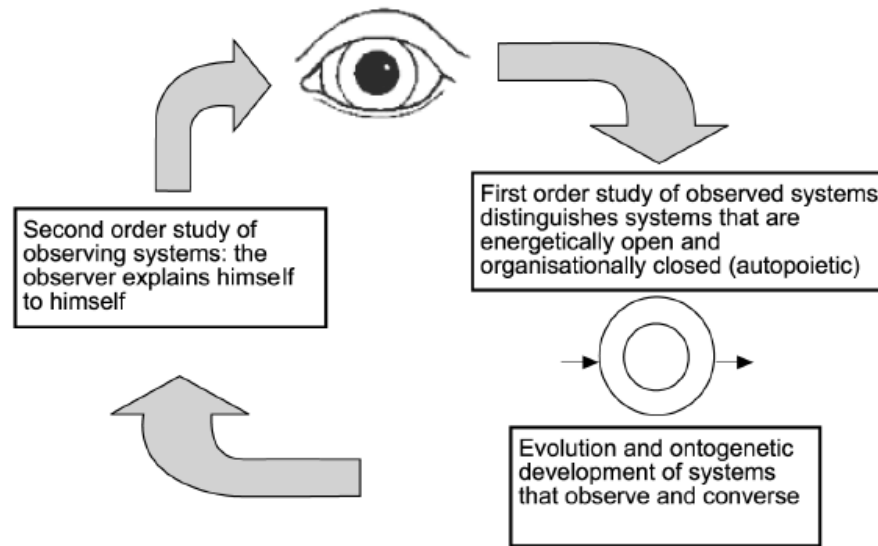


Figure 3: Illustration of Second-Order Cybernetics. Scott, Bernard.
“Second-order Cybernetics: An Historical Introduction.” *Kybernetes*, vol. 33, no. 9/10,
Oct. 2004, pp. 1374.

Before diving deeper into the third part, a further explanation of what an organizationally closed system is might be necessary. Maturana and Varela defined an organizationally closed system (an autopoiesis or an autopoietic machine) as “a machine organized (defined as a unity) as a network of processes of production (transformation and destruction) of components” (78-79). Corresponding to cybernetics² characteristics of being “metadisciplinary” (a “discipline about disciplines”) (Scott 1368), they aimed to find “the organization of the living,” or more precisely speaking, a shared organization of all the living being (Maturana et al. 22). Based on the recognition of the living as a machine, which refers to one of the philosophies of cybernetics as a formal study of “all possible machines” (Ashby), the focus of Maturana and Varela was on the

² By using “cybernetics” separately, I am referring to all the concepts related to the term, including its first-order and second-order.

“processes and relations between processes realized through components.” Under their narrative, it is those processes and relations “define a machine as a unity and determine the dynamics of interactions and transformations which it may undergo as such a unity;” thus constituting the organization of a machine (which includes the living as they recognize it as a type of living machine) (77).

However, Maturana and Varela also emphasized that an explanation of a machine is “always a reformulation of a phenomenon showing how its components generate it” through its organization. Additionally, it is always generated by the observer. Corresponding to such an argument, as Von Foerster mentioned, “a brain is required to write a theory of a brain.” That is to say, any interpretation or explanation produced by the observer has to account for his own activity (*Understanding Understanding*, 289). Back to Scott’s figure, thirdly, the observer is then always a part of what he is observing, therefore trapped in a hermeneutic cycle of interpretation and explanation. In this case, to know and to claim knowledge becomes a circular self-referential process.

While second-order cybernetics aimed to challenge traditional objective thinking given its interest in the subjective participation of an observer in the system, particularly in Western science, applying its philosophy here emphasizes the role models play within such a circular process of knowledge production. When explaining his models of archetypes, as mentioned earlier, Black quoted Abrams’ theory to explain our use of “a dominating system of concepts to describe a new realm of application by analogical extension” (240). Abram stated that, when investigating an area without prior concepts to characterize and structuralize, there is a tendency for humans to “describe the nature of something in similes and metaphors.” One might get confused here as metaphors are commonly used as “device[s] of the poetic imagination and the rhetorical flourish” in daily life (Lakoff 3). Black applied his interpretation of metaphors to his

model of models, yet his original discussion on metaphors also stopped at their semantic level. George Lakoff and Mark Johnson recognized metaphors as “metaphorical thoughts” and scrutinized them at a conceptual level. In this case, they argued that “[the] essence of metaphor is understanding and experiencing one kind of thing in terms of another” (5).

Under such a narrative, metaphors are models. They can also be recognized as constructed artifacts for representation. Therefore, Abram’s stated tendency can be translated into what has been argued earlier in this paper: to understand is to model. However, what has not been further explored is the asymmetrical relationship between a model and its referent. Wharton emphasized that “a model is not a model of an object if it is identical with it – rather, it would be a clone, a simulacrum, a double” (10). Wartofsky proposed a mathematical expression to model the asymmetry of the modeling relation:

$$M(S, x, y) \& R(x) < R(y)$$

where S takes x as a model of y, with a certain range or richness (R) of relevant properties of y (6-8). However, as the referent, y has to be richer in properties than its model x; otherwise, their relationship becomes a “negative analogy” (Hesse 27), which would place “a limit on models taken as intended factual descriptions” (Wartofsky 7).

Referring back to the tendency, if a model represents its referent, as “the better known,” then, when understanding by modeling, we are clearly using such better known to “elucidate the less known” (31-32). Regarding second-order cybernetics, how we know heavily relies on what we already know – the observers’ existing sets of assumptions. However, what Abram, Black, and Wartofsky have not asked is: after being tested as good fits or proven to be fruitful in discoveries (the standard of evaluating a scientific model as mentioned above), when a good model has been recognized as a knowledge claim, a proffered truth, what will humans intend to do?

Such a proffered truth then becomes the new better known to elucidate more less known. The original model is now an archetype, which can be used to create new models. A verified medical theory can lead to the creation of a new pill, which should theoretically solve the illness that the theory addresses. After being tested effective, the methods of the pill's development can be applied to relative fields, helping research other pills. Therefore, not only is there a deconstructive hierarchy of models, where, as illustrated in the first meta-meaning of models, a model, when not being an archetype, can be deconstructed into low-level models, until reaching archetypes. At the same time, there is a constructive loop of models, where we, as humans, are using the better known (archetypes or good models that have become archetypes) to elucidate the less known. After tested useful or with fruitful discoveries, when our understanding of the less known is rich enough, it can then be used as a new good model, a new better known for us to inform new less known (new models), and so on. Such use of models for constant knowledge production is the third meta-meaning of models.

4.6 The Hermeneutic Loop and the Heuristic Loop

Combining all three meta-meanings of models, hence I conclude my meta-model of models. We are not simply living in a world of models. As the observer, as we draw on ourselves to know and to claim knowledge, we find ourselves in both a hermeneutic loop and a heuristic loop. The hermeneutic loop refers to the epistemology of the observer, where we explain ourselves to ourselves through the hierarchy of models. The heuristic loop stands for our constant construction of models, so as to keep expanding our understanding of the world and ourselves. As humans, we produce knowledge through these two loops, as a result of which models have always been the only medium for our knowledge production.

In this section, my interests in models might have no difference from those of Black, Morgan, Morrison, Wartofsky, and Wharton. In a sense, we all conduct both interdisciplinary and metadisciplinary investigation of models. In other words, we all theorize models from a broader and more general perspective. Additionally, our theories can all be reflected back on themselves: They are all models, yet trying to model the concepts of models. By doing so, undoubtedly, as essentially a model, my argumentation of a meta-model of models can never surpass the richness of the original term as the referent, just like how Wartofsky's mathematical expression shows. However, it is sufficient enough to help us better understand generative AIs as models. On one hand, generative AIs are barely old wine in a new bottle. They are essentially models created by us to serve purposes. While those purposes vary in different perspectives, such as a programmer might use it for solving specific problems or processing specific tasks, a general understanding of them, through my meta-model of models, reveals the same human tendency for using them to produce knowledge through our hermeneutic loop and our heuristic loop. On the other hand, generative AIs have the potential to change our current ways of knowledge production as we design them as new organizationally closed systems. As mentioned in section 3, it is the interactions and transformations its agents undergo that define an AI to be generative. That is to say, generative AIs are new organizationally closed systems that function representationally based on our current understanding of the human body as a living machine. We trust their power because they are all fundamentally computers – a tool that has helped us the most in understanding and creating representations since its development. By transferring our cybernetic circular causal loops to generative AIs, they hold the potential to help us jump out of the epistemology of the observer. Generative AIs are empowered to construct new models themselves, thus informing us of new ways of knowledge production.

“Neosentience,” proposed by Bill Seaman and Otto Rössler, is a theoretical inquiry that aims to develop more powerful generative AIs. Such an inquiry focuses on the properties of sentience, recognizing them as the core interactions occurring both 1) within a complex organizationally closed human body system and 2) between this system and its nested larger environment. By abstracting the operation process that enables these properties of sentience to arise, Seaman and Rössler discussed the possibility of developing a more powerful generative AI – *the benevolence engine* – which seeks to “exhibit sentience of a new variety” (2). Undoubtedly, their benevolence engine holds the potential to change our mode of knowledge production as it is designed to be a new sentient entity that aims at jumping out of the epistemology of human observers. However, their approach also reveals a seeming impossibility of building an ultimate generative AI. How can we define and abstract all the properties of sentience in the first place? This is an underlying question shared by all approaches towards building generative AIs. As a model, its richness of interactions can never surpass its referent – human intelligence – because we currently have no methods to fully define and abstract all interactions within our body system, not to mention those connected to the environment. This is a paradox rooted in the difference between knowledge production as an adaptive function and our practical approaches towards it.

Under the narrative of second-order cybernetics, knowledge is a resolution of alternatives. While von Foerster suggested to “act always so as to increase the number of choices,” it seems impossible to define and abstract all the alternatives if knowledge production is essentially an adaptive function. Especially towards building AIs as forms of computation, our abstracting process has a pre-assumption that what we abstract must be computable in the first place. That is to say, while we recognize reality varying in forms of alternatives in theory, we hold the faith that everything is computable in practice. Therefore, following our theoretical assumption, an ultimate generative AI is something you only do on faith. However, while we can

never reach an ending point under such a narrative, building generative AIs can also be recognized as a constant knowledge production process corresponding to our hermeneutic and heuristic loop. Building AIs that function representationally (unpacking the functional entailment structures at operation in human body) can inform new insights into our understanding of those representations. We can then use those new understandings to create new AIs, and so on. This is a different approach to that of the goal-driven AIs like LLMs that currently are being explored. By repeating such a process within the loops, eventually, we can at least build better and better understandings, not only of AI and its construction, but also of humans ourselves and the world around us.

5. Generative Models

To more clearly illustrate generative AIs' potential to provide new knowledge and new modes of producing knowledge in practice, I recognize them as generative models under the meta-model category. I argue that a generative model is an emergent computational system constructed to at times solve problems or process tasks that potentially exceed or do not follow human computational rationality. While section 4 can be recognized mainly as a theoretical approach toward generative AIs' potential in knowledge production, classifying them into such a specific category provides more practical understanding. Through the examination of existing changes brought by the use of generative models throughout history, a better picture of generative AI's potential in future practices can be formed. My discussion of historical generative models concentrates on one outcome pushed by them: procedural modeling.

5.1 Procedural Modeling

While there is no distinct definition of what procedural modeling (PM) is, it always functions as an umbrella term to “[encompass] a wide variety of generative techniques that can (semi-)automatically produce a specific type of content based on a set of input parameters” (Smelik et al. 1). Compared to traditional linear modeling, where a 3D model has to be explicitly and manually constructed and specified, PM has been argued to include three advantages: *abstraction* (Ebert), *data amplification* (Smith), and *data compression* (Smelik et al.). *Abstraction* stands for PM's ability to abstract complex modeling scenes or sequences into “a function or an algorithm (i.e., a procedure)” (Ebert 2). In Houdini, a well-known PM software, every modeling action is abstracted into a node. Each node can then be connected to others, therefore wired into a node-based network to produce composite results. Such an accumulative process, from simple nodes to complex models, is recognized as *data amplification*. A few sets of input parameters do

not simply yield a large number of details. More importantly, a slight tweak on the parameters permits a wide variety of distinct model generations. Precisely because details and varieties can be stored in simple nodes and parameters, a model's actual geometry is generated only when necessary. In other words, while the creator has to remember which nodes he uses and what parameters he sets, the computer stores the data of the nodes and the parameters. Such ability to save and run "instruction" data for generation is defined as *data compression*. It empowers PM to vastly reduce the amount of modeling effort required, both for the creator and the computer, to produce digital content. As a result, PM is most commonly used for large-scale projects, such as the creation of terrain, architecture, and city layout. It holds to its promise to create unexpected visual effects, mostly for application in films and video games.

5.2 Generative Models in PM

Models become generative in PM when their *data amplification* process exceeds or does not follow human computational rationality. Stochastic methods are often involved when using PM to amplify more seemingly realistic features or more unexpected effects. Most stochastic methods in PM are based on noise generators – pseudo-random functions – as, in most cases, artists might not want what they eventually create to be a totally random result without any clear meanings. That is to say, employing noise generators aims to bring complexities in detail, yet the result is still interpretable to humans. A certain degree of randomness is necessary for more realistic representations or more unexpected effects, yet the final features need to be smooth for naturalness. Such a concept of *pseudo randomness* is essential as it allows humans to extract interpretable patterns from graphics that a digital computer generates. Fig. 4 illustrates the difference between *total randomness* and *pseudo randomness*. In some situations, meanings can be generated only when we constrain features' randomness to a certain degree. *Pseudo*

randomness is achieved by applying mathematical functions to approximate noise generators' outputs if their inputs are near each other. Using the latter image in Fig. 4 as an example, its *pseudo randomness* roots in a procedural generation function named Perlin noise¹. In this case, if two input coordinates are near each other, then the results of inputting them in Perlin noise (pixels' colors) will be approximate. Precisely because I set the inputting process to follow a left-to-right, top-to-bottom sequence², the final image shows a subtle difference among each pixel on the same horizontal line, with a more distinct, but still smooth transition across those lines.

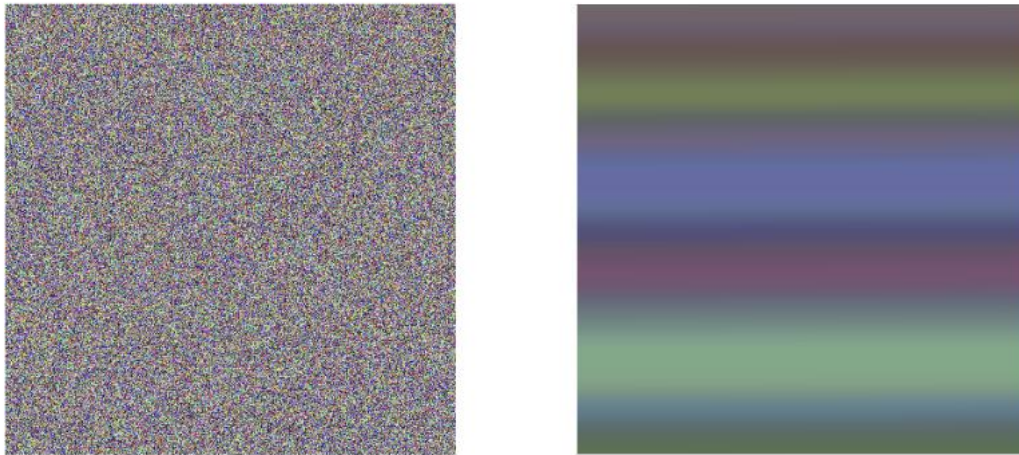


Figure 4: Total Randomness vs. Pseudo Randomness (Perlin Noise). Both are created in the p5.js editor. The only difference between their codes is the function to generate each pixel's red, green, and blue (RGB) value. The former uses the *random()* function, and the latter uses the *noise()* function with particular seeds and changing levels. Further explanation of these two functions can be found on the p5.js website: <https://p5js.org/>.

¹ For further explanation of Perlin noise's mathematical formulas, see Raouf's blog: <https://rtouti.github.io/graphics/perlin-noise-algorithm>.

² That is to say, the Perlin noise will process each pixel from left to right with the same y-axis. From the top to the bottom, it will process each y-axis following the same left-to-right logic.

A creator can understand his noise functions when he manually writes them. He might also understand the meaning of the parameters he set for his noise functions when input data are not complicated (small amount, two dimensions, etc.). However, when those noise functions are abstracted into an agent, such as a node in Houdini, a creator who does not understand them in the first place loses the meaning of the agent's parameters (noise seed, noise weight, etc.).

Furthermore, a creator who does understand might encounter the same situation when input data becomes too complicated. Especially in Houdini, as a PM software targeting physics-related simulations (e.g., how a model performs itself to gravity and time), understanding the meaning of noise generator's parameters becomes an almost impossible task.

Houdini's characteristic of being simulative reveals the goal of PM: to model *dynamically* through *data amplification*. In traditional linear modeling, a modeler creates a 3D model following explicable rules of logic, just like how he creates a physical model in real life. On one hand, to make a sculpture of a bare tree, a sculptor might first choose the material he wants, either wood, stone, or metal. Then, he uses his physical tools, such as a hammer and a knife, to carve an initial prototype. Finally, he works on the details, polishing them into a completed artwork. On the other hand, to make a 3D model of a bare tree in a traditional linear modeling environment, a modeler might start by clicking the corresponding button and create a cube, as the initial prototype, on the screen. Then, he sets the parameters of his cube (width, height, numbers of polygons, etc.) and adjusts their primitives' positions to make it more like a bare tree. Finally, within the rendering process, he chooses the material he wants and applies it to his 3D model. Precisely because his creation process is a stepwise reasoned one in the first place, a 3D modeler understands the parameters he sets. PM is different. Interactions among agents, such as gravity and time in Houdini, empower PM to be an emergent computational system that does not follow human computational rationality. A modeler can algorithmically model a bare

tree in Houdini, but it is not what Houdini is potentially designed for. As a PM software, Houdini encourages a modeler to generate a bare tree through the interaction of nodes (though in most cases, following a top-down sequential order). Furthermore, by providing users with variables like gravity and time, Houdini encourages a modeler to grow a realistic or unexpected tree, or even a corresponding forest, by manipulating the interactions among those variables as agents. *Data amplification*, or more specifically, *agent amplification*, in this case, empowers PM to be a dynamic task. New global behaviors emerge from PM's *agent amplification* process, thus making its models to be generative.

5.3 PM's Impact on Artistic Knowledge Production

Reflecting back on PM's three advantages, its impact on artistic knowledge production relies on changing how to create digital art, from graphics recognition, to be under data or data agent (such as a node in Houdini) format. Traditional linear 3D modeling only digitizes models in computer's operation process: how their information can be saved as data and how that data can be decoded to be shown as graphics on a computer screen. How a user interacts with those graphics is similar to how he interacts with visual patterns on a real-life physical model. Undoubtedly, differences still exist between these two types of interaction. A 3D modeler is now using digital tools with different manufacturing orders as mentioned above. Additionally, interacting with objects in a graphics mode empowers artists to tailor their works more accurately. Meanwhile, they might get more freedom to express what they want, by changing the position of the camera, the parameters of the lights, or the color of their materials. Yet, by doing so, they also lose the power of feeling their works' texture in person, as well as other possible physical interactions with them. My argument of a similar cognitive interaction focuses on how a 3D modeler still deals with the graphics on the screen in patterns. Though the computer handles

those graphics in the form of data, he still manages them as visual patterns like in a physical setting.

Again, a 3D modeler can create a model in Houdini based on graphics recognition, but it is not what Houdini is potentially designed for. Houdini encourages a modeler to deal with graphics on the screen as data agent formats – nodes with input parameters and output results. This can be reflected in how Houdini documents the use of each node. On every node documentation page, Houdini shows the data type of input parameters and output results, as well as how those input parameters are operated to form the corresponding output results. Furthermore, by providing a modeler with tools to examine the data information of graphics, such as a “Geometry Spreadsheet” to view a geometry’s data information (point coordinates, scales, normals, etc.), Houdini encourages a modeler to deal with graphics on the screen as data format. A modeler can do so by using Vector EXpressions (VEX) codes to write his own algorithmic functions. Under such a narrative, therefore, artistic knowledge in PM shifts from techniques for visual patterns to agents and algorithms for graphics. While a modeler acquires knowledge for building a traditional linear 3D model by learning corresponding techniques, he knows how to generate a PM model by studying which agents or algorithms to use, as well as studying setting related parameters.

5.4 Generative Models’ Potential Learned from PM

Precisely because PM aims to parameterize our approach to graphics, generative models’ potential in the field of art relies on using algorithms, mainly mathematical functions, to imagine the unimaginable. In traditional linear 3D modeling, a modeler has a pre-imagination of what he tries to model. Then, he uses different digital tools to manually specify it based on the direct and precise feedback of his changes on the model. Such a pre-imagination can be unnecessary for the

generation of a PM model. A modeler can purely have an intent on algorithms (which to use or what to write) and then imagine the unimaginable by following their narrative. To illustrate such a methodology, I will show my process of creating a PM model in Houdini, which purely concentrates on the visualization of algorithms.

My PM model starts with one intent: since I have illustrated the principle of Perlin noise in a 2D setting (in my discussion of fig. 4), how can I visualize it dynamically in a 3D space? As mentioned earlier, Perlin noise targets complexities in details but a smooth final result for humans to interpret. Therefore, in Houdini (See fig. 5)³:

1. I created a sphere, by adding a “Sphere geometry node” (sphere1) as a general final shape for the audience to interpret.
2. I employed an “Add geometry node” (add1) to delete the sphere’s geometry but keep its polygons as points.
3. I created a cuboid by adding a “Box geometry node” (box1) and adjusted its input parameters.
4. I added a “Copy to Points ^{2.0} geometry node,” (copytopoints1) to copy my cuboid to every point of the sphere. These copied cuboids are the sources of Perlin noise’s complexity.
5. I employed an “Attribute Wrangle geometry node” (pointwrangle1). Within it, I wrote VEX functions to change the direction of each copied cuboid based on Perlin noise⁴.
6. I added a “Normal geometry node” to fix the normals of my copied cuboids⁵.
7. Within pointwrangle1, I created a vector value named “Offset.” I tied its value to the time agent in Houdini and then added it as a variable to the noise function, so that the copied cuboids can change their directions based on the changing results of the noise function through time.

After these steps, by clicking on the play button, my PM model, targeting dynamically visualizing Perlin noise in a 3D place, comes into being (See fig. 6). It has a general sphere shape, yet with cuboids as its body. The direction of each cuboid will change based on the changing results of the noise function through time. Precisely because the results of the noise function (directions of cuboids) will be approximate if input parameters (coordinates of cuboids) are near each other, the

³ My description only shows general moves of my creation process for simplification. For example, there is a lack of information on parameters as their discussion might be verbose and cause misunderstandings.

⁴ I used the “curlnoise VEX function,” which can compute divergence free noise based on Perlin noise. For further details, see: <https://www.sidefx.com/docs/houdini/vex/functions/curlnoise.html>.

⁵ Without this node, the normals of copied cuboids will be messed up by the sphere’s point normals.

model as a whole generates a new global behavior – a smooth flow of cuboids – for the audience to interpret.

Then, my next question is, how can I better distinguish such a flow for the audience by adding agents like materials and colors into the interaction? Through my discussion with Mark Olson, I realized that I have to set a new variable – a new agent – that stores a certain type of data information based on each cuboid’s direction, so as to make agents like materials and colors join the play. Therefore, in Houdini (see fig. 7):

1. I used a new Add geometry node (add2) to create a new point at the sphere's center.
2. I created a new float variable named “ang” in pointwrangle1 to calculate the angle between each cuboid’s normal (a vector value) and its vector value towards the new point.
3. I added an “Output geometry node” (Output_Sphere_One) to pass my model to the “stage” network, where I can add a camera, lights, and materials for the final rendering process.
4. Within the stage network, I set a cuboid’s metalness⁶ and color (white or purple) to be decided by the ang variable.

After these steps, by rendering the model, it shows a more distinguishable visual pattern for the global flow happening on the sphere (see fig. 8). Precisely because the noise function offers a smooth change among the directions of cuboids, their ang values also change smoothly. By tying such an ang variable to cuboids’ metalness and color, as a result, these two agents also emerge a smooth transition among cuboids – from non-metal to metal and from white to purple – for the audience to interpret.

⁶ Metalness is a float value in Houdini, from 0 to 1.

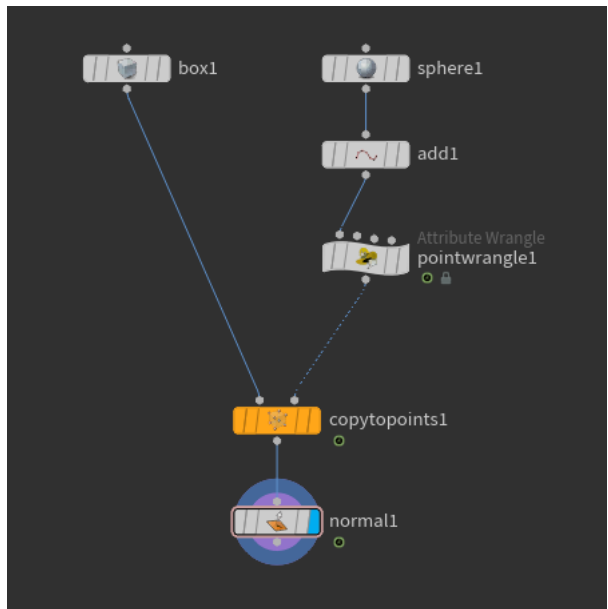


Figure 5: Nodes I used in Houdini for my PM Model. Houdini version: 19.5.773.

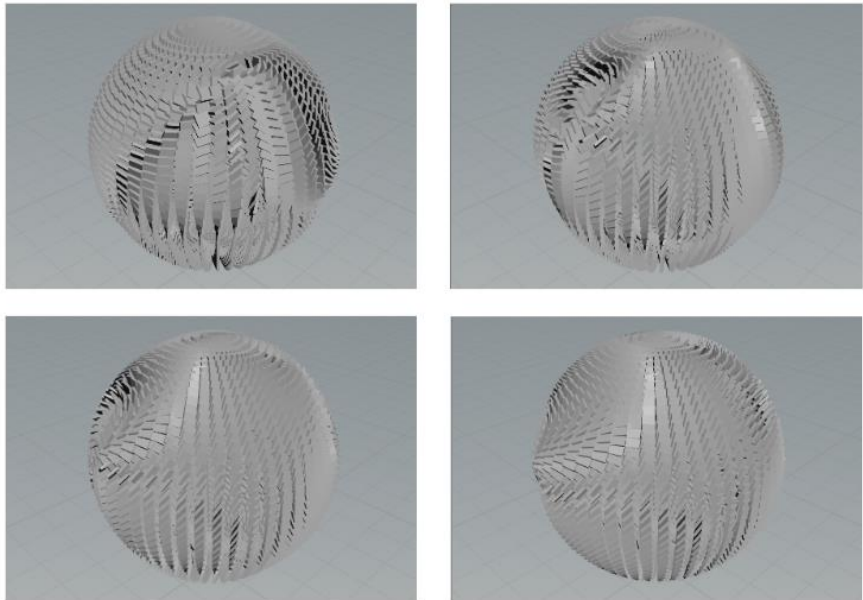


Figure 6: Screenshots of my PM model in Houdini. From top-left to bottom-right, frame numbers: 24, 48, 72, 96.

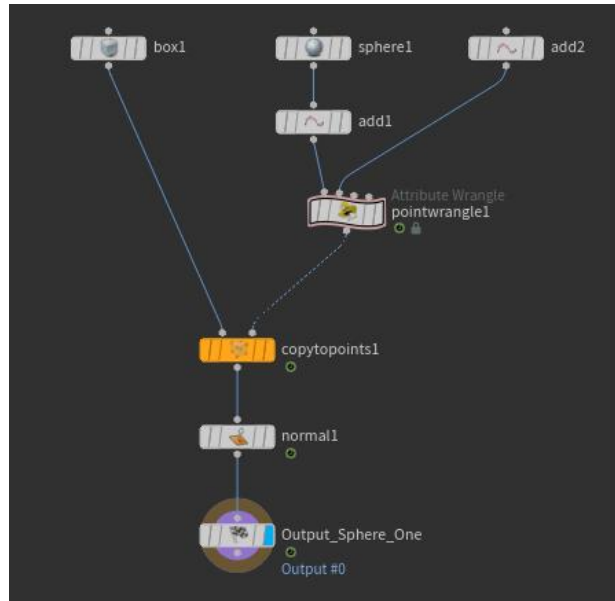


Figure 7: Nodes I used in Houdini for better distinguishable visualization.
Houdini version: 19.5.773.

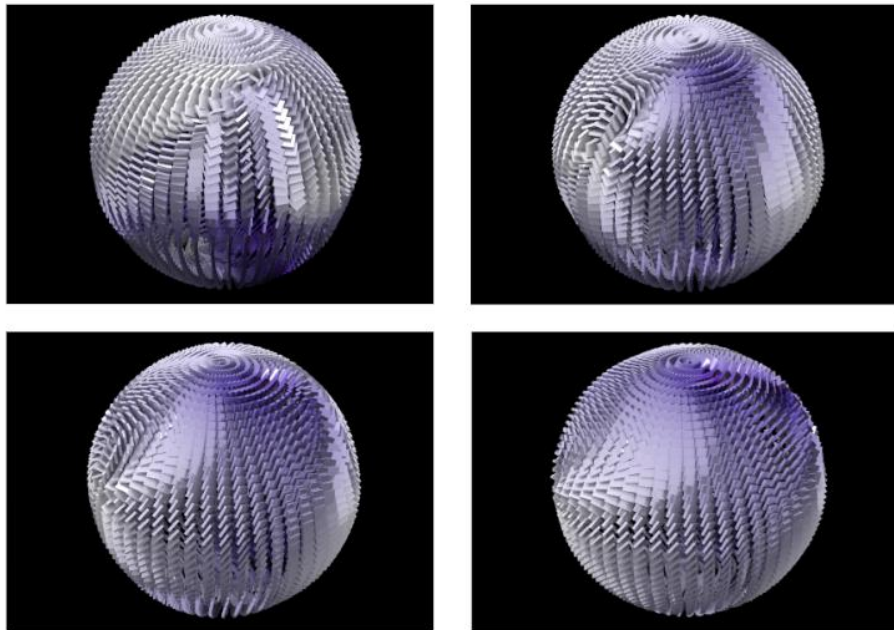


Figure 8: Screenshots of my PM model after rendering. From top-left to bottom-right, frame numbers: 24, 48, 72, 96.

In all, my creation process of a dynamic PM model demonstrates generative models' potential in the field of art. By treating artistic knowledge as algorithms for parameterized graphics, our imagination of visual patterns can be expanded when we transform non-visual algorithmic functions into digital visualizations. Furthermore, by manipulating the interactions among agents, we can imagine the unimaginable when new global behaviors emerge. In my case, I concentrate my intent on visualizing Perlin noise dynamically in a 3D place. Such a characteristic of being dynamic cannot be achieved without the interactions among agents. I can pre-imagine a flowing effect of cuboids because I know the principle of Perlin noise, and I have already visualized it in a statical 2D setting. That is to say, I still obtained a high degree of control over my model as I was barely trying to visualize an already known and widely used noise function. What about an algorithm that I have not visualized yet? What about a mathematical function that I do not understand in the first place? Even if I understand Perlin noise and know how to use Houdini, my creation process is still a manipulation of agent interactions. It is till the end, when their interactions are complete and new global behaviors emerge, that I come to a complete picture of what I initially aimed for. One might counter my arguments by declaring my model is not emergent enough, thus still being predictable. I intentionally designed it to be simple for better comprehension. It can be more emergent by adding more stochastic methods into the agents' interactions. Additionally, instead of using time as the indicator, cuboids' movements can be tied to another agent or multiple agents for more unpredictability. Internal functions can be applied to make cuboids' movements totally based on their own operation status, thus being independent of my variables' control. In a sense, models in PM can become more generative and more unimaginable by adding more randomness and more interactions.

5.5 Generative AIs' Potential for Art Creation

Classifying Generative AIs as generative models, therefore, one of their potentials for art creation is precisely what has been argued above: to imagine the unimaginable. While PM focuses on visualizing algorithms, currently available artistic generative AIs rely on visualizing the meaning of our language. Stable Diffusion (SD) is a text-to-image AI model “capable of generating photo-realistic images given any text input” (“Stable Diffusion Online”). While it has an online version for easier accessibility, a user can also deploy the model on his own device as its code and model weights have been released publicly on GitHub. SD is famous not only because of being open source but also because it claims no rights to the images it generates. Images generated on the online version are licensed under the “CC0 1.0 Universal Public Domain Dedication” (“Stable Diffusion Online”), and a user can claim rights to images if they are generated in a local environment (Coles)⁷.

Reflecting on the concept of Neosentience mentioned in section 4, while Seaman and Rössler explained such a term thoroughly by language, it is hard to imagine visually. Therefore, I tried using SD to visualize Neosentience (see fig. 9). While I have no clue why and how SD generated such two images, they help express Neosentience in visual patterns. Meanwhile, exploring such why and how can be my further attempts to generate new knowledge. We can ask ourselves: How can such images help us understand Neosentience? Or more generally speaking, how can visual patterns help? Artistic generative AIs can bring new questions into discussion by providing what is unimagined/unimaginable before.

Undoubtedly, uncontrollability can be recognized as a problem not only for artistic generative AIs, but also for all artistic generative models. Even if SD provides “ControlNet” to

⁷ However, copyrights for generated images in general is still a topic under debate. For example, Stable Diffusion might create “unauthorized reproductions in training” (Newman).

constrain generated images by adding extra conditions (“ControlNet Online”), it is a black box where a user can only adjust input parameters and do so after actually seeing the output results. Their understanding of the parameters has been limited to the optimality of the output – what they want on those generated images. They lose the meaning of parameters, but in a sense, such meanings are essential for the generation process. The parameters of an SD’s training model and its corresponding training data heavily influence what a user can generate eventually. A slight tweak in the parameters, or a change of training data, can cause a significant difference in output (see fig. 10), while a user has no idea what happens inside the black box. He can only keep playing with his parameters and keep generating outputs to reflect, so as to gradually reach what he aims for.



Figure 9: Visualizations of the term Neosentience. Both are trained through the base model of SD (version 1.5): v1-5-pruned-emaonly.safetensors. Both generated images only used Neosentience as the prompt. The second one also has a negative prompt – text – which forces the model to generate images without text.

The same works for Houdini. When a user relies on its nodes to build 3D models, the model’s generation process is a black box that exceeds the creator’s control. *Data amplification* allows PM to generate complex models accumulatively, but the results are often unpredictable.

Unlike traditional linear modeling where every precise change can be made during the construction process, the creator who utilizes PM has to repeatedly adjust the inputs based on his perception of the outputs. In this case, PM not only abstracts and compresses the model into procedures. It spontaneously shapes and limits the creator's understanding of the model merely to the inputs he sets and the outputs he sees.

This is the curse of using generative models for art creation: They can never accurately fabricate an existing imagination. However, modelers always carry imagination to build models in PM. As essentially a commercial 3D software, Houdini is often used for creating concrete models for use. Because of its power to deal with data and create dynamic objects, Houdini holds its fame in modeling complicated scenes and visual effects, as mentioned earlier. In this case, the use of PM becomes a way of control: How a modeler can adjust parameters, apply nodes, and write customized algorithms to limit the generated model to what he aims for in the first place. My argumentation of PM's potential is counter to such a common usage. Only by getting rid of the goal of building a concrete model can PM allow us to imagine the unimaginable.

Undoubtedly, generative models' uncontrollability is a double-edge sword. It is a strength toward the unimaginable, but it might become a problem when such models are released to the public as a total black box, where the creators also lose the understanding of their parameters. Again, as discussed in section 3, further discussions on regulating the practical use of generative AIs, such as SD, are necessary. Only by establishing clear and thorough regulation rules can we realize their potential safely.



Figure 10: Visualizations of the term Neosentience based on two different training models. The former has been trained mainly based on anime images, and the latter has been trained mainly based on human model images.

6. Conclusion

This paper seeks to provide an initial ground to understand generative AI through two approaches. First, it looks at history to distinguish generative AI and predict its future potential. Through closely examining AI's historical development, this paper defines generative AI based on its emergent properties. Generative AI is an emergent machine. Its new global behaviors offer significance for its construction and for us to study. Furthermore, by proposing a meta-model of model's historical uses and recognizing generative AI as models, this paper evaluates their potential in human knowledge production. On the one hand, as deliberately constructed artifacts, generative AI follows humans' existing knowing and learning patterns – the hermeneutic and heuristic loop – to produce new knowledge. On the other hand, designed as new organizationally closed systems, generative AI holds the potential to be out of our loops, gradually helping us improve our knowledge production as an adaptive function. Though the vision of a true AI is more likely a faith, constantly building AIs that function representationally can enable us to better understand the world and ourselves.

The attempt to build a bridge between theories and practices is this paper's second approach. The understanding of generative AI, as well as AI in general, varies in terms of differing perspectives and is often entangled in discourse. This paper aims to form a more general comprehension, a meta-model, by combining the discussions of theories and practices. Undoubtedly, as eventually a model, many aspects are missed and remain to be further explored. Furthermore, details are lost by abstracting different perspectives under such a meta-discussion, and misunderstandings might appear due to the lack of full explanations. Therefore, readers are encouraged to return to my sources and closely examine those theories or practices. Again, as eventually a model, this paper encourages challenges and further extensions to improve its richness towards its referent – generative AI.

In all, while many critics have been keen on discussing the possible benefits and drawbacks of generative AI (e.g., Bell; Hughes), my two approaches demonstrate that their developing philosophy, illustrated as generative models, have already long infiltrated into our daily lives and left a profound impact on how we understand the world and ourselves. Therefore, situating at a time when the rise of computing power makes AI evolve rapidly, this paper advocates an interdisciplinary investigation of AI throughout history, re-examining what we have already neglected, so as to build better regulation in practice and for the future.

Appendix

Glossary

1. Knowledge

My definition of the term knowledge follows Wartofsky's narrative: knowledge is what we know.

In practice, what knowledge is relies on what it represents specifically in various fields, such as means to draw in art creation. By claiming that generative AIs have the ability to produce new knowledge, I mean they have the potential to bring us new types of specific knowledge in their corresponding fields.

2. Model

Following Wartofsky's narrative, I define a model as a deliberately human-made representational artifact, by purposes and to serve purposes. Anything can be a model if we offer an asymmetrical relationship with its referent. The referent does not need to be a concrete entity. It can also be intangible, such as a concept or a theory. By claiming a model cannot surpass its referent richness, I mean it cannot fully obtain the related characteristics or properties of its referent. Otherwise, their relationship becomes a negative analogy. However, a model can be richer outside of the relationship we construct. For example, a toy car can obtain more functions for play. A designer can add a remote-control system to it, but such a property is not included in the original relationship we built.

3. Emergence

Based on OED's definition, my treatment of emergence refers to the process of coming forth, where the process itself and its outcome lacks defined patterns for humans to understand. Particularly in computation, under Forrest's narrative, emergent computing stands for the

appearance of implicit global patterns based on the interactions among agents, which follow explicit instructions offered by humans.

3. Generative

My definition of models being generative focuses on whether they have emergent properties or behaviors. Their generativity relies on emerging new global behaviors which do not follow the instructions of their composing agents. In computation, it is always a shift from quantitative changes to qualitative changes, as we program those agents in data format and under stepwise algorithms.

4. Interpretability

I argue two forms of interpretability towards understanding computational systems: 1) if we can understand their operation processes through stepwise reasoning, 2) if we can understand the meaning of their parameters and the operations processes acting upon those parameters. While the AI tool CORELS falls into the first category, Tegmark's methodology towards understanding AI corresponds to the second (both examples can be found in section 3.2).

5. Black Box

My treatment of generative models as black boxes falls into the second categories of my definition of interpretability. They become total black boxes when their creators lose the meaning of the parameters they set. In this case, creators can only keep adjusting parameters based on the outputs, to gradually reach what they want in the first place. They cannot set a parameter and immediately know what they can get from it, without seeing the actual results.

References

- “algorithm, n.” *Oxford English Dictionary*, Oxford University Press, July 2023, <<https://doi.org/10.1093/OED/1019775631>>
- “Automotones.” *Theoi Greek Mythology*, www.theoi.com/Ther/Automotones.html. Accessed 10 Nov. 2023.
- “AlphaGo.” *Google DeepMind*, Google, deepmind.google/technologies/alphago/. Accessed 12 Nov. 2023.
- Anderson, P. W. “More Is Different: Broken Symmetry and the Nature of the Hierarchical Structure of Science.” *Science*, vol. 177, no. 4047, Aug. 1972, pp. 393–96. DOI.org (Crossref), <https://doi.org/10.1126/science.177.4047.393>.
- Ashby, W. Ross. *An Introduction to Cybernetics*. Martino Publishing, 2015.
- Angelino, Elaine, et al. “Learning Certifiably Optimal Rule Lists for Categorical Data.” *Journal of Machine Learning Research*, vol. 18, no. 1, June 2018, pp. 1–78.
- “Adversarial Example Using FGSM.” *TensorFlow*, TensorFlow, www.tensorflow.org/tutorials/generative/adversarial_fgs. Accessed 16 Nov. 2023.
- “black box, n.” *Oxford English Dictionary*, Oxford University Press, July 2023, <<https://doi-org-s-a865.shfd.findres.net/10.1093/OED/1091899133>>
- Balachandran, Amar, et al. “Introduction to Modeling and Simulation Systems.” *Simulation & Modeling*, <https://uh.edu/~lcr3600/simulation/historical.html>.
- Bell, Elyse. “Generative AI: How It Works, History, and Pros and Cons.” *Investopedia*, Investopedia, 26 May 2023, www.investopedia.com/generative-ai-7497939.
- Black, Max. “Metaphor.” *Proceedings of the Aristotelian Society*, vol. 55 (1954-1955), pp. 273-294.
- Black, Max. “Models and Archetypes.” *Models and Metaphors: Studies in Language and Philosophy*, Ithaca: Cornell University Press, 1962, pp. 219-243.
- Binet, Alfred and Theodore Simon. *The Development of Intelligence in Children*. Williams & Wilkins, 1916.
- Bommasani, Rishi, et al. “Introduction.” *On the Opportunities and Risks of Foundation Models*, pp. 3-12. <https://doi.org/10.48550/arXiv.2108.07258>.
- Boyle, Robert. *The Works of the Honourable Robert Boyle*. Edited by Thomas Birch, 2nd ed., 6 vols., London, 1772.

- Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Reprinted with corrections 2017, Oxford University Press, 2017.
- “ControlNet Online.” *Stable Diffusion*, stablediffusionweb.com/ControlNet. Accessed 20 Nov. 2023.
- Conway, Flo, and Jim Siegelman. *Dark Hero of the Information Age: In Search of Norbert Wiener, The Father of Cybernetics*, Basic Books, 2006. ProQuest eBook Central, <https://ebookcentral.proquest.com/lib/duke/detail.action?docID=876408>.
- Cheng, Mingyong. “The Creativity of Artificial Intelligence in Art.” *The 2021 Summit of the International Society for the Study of Information*, MDPI, 2022, p. 110. DOI.org (Crossref), <https://doi.org/10.3390/proceedings2022081110>.
- Coles, Gloria. “Can You Sell Stable Diffusion Images?” *PC Guide*, 30 Mar. 2023, www.pcguide.com/apps/can-you-sell-stable-diffusion-images/.
- “Dall·E 3.” *OpenAI*, OpenAI, openai.com/dall-e-3. Accessed 6 Nov. 2023.
- Dressel, Julia, and Hany Farid. “The accuracy, fairness, and limits of predicting recidivism.” *Science advances* vol. 4,1 eaao5580. 17 Jan. 2018, doi:10.1126/sciadv.aao5580
- “emergence, n.” *Oxford English Dictionary*, Oxford University Press, July 2023, <<https://doi.org/10.1093/OED/1146385283>>
- Ebert, David S. “Introduction.” *Texture and Modeling: A Procedural Approach*, AP Professional, Boston, 1994, pp. 1–5.
- Foucault, Michel. *Power / Knowledge*. Edited by Colin Gordon, Harvester Wheatsheaf, 1980.
- “Francis Crick, Rosalind Franklin, James Watson, and Maurice Wilkins.” *Science History Institute*, Science History Institute, 28 May 2023, sciencehistory.org/education/scientific-biographies/james-watson-francis-crick-maurice-wilkins-and-rosalind-franklin/.
- Forrest, Stephanie. “Emergent Computation: Self-Organizing, Collective, and Cooperative Phenomena in Natural and Artificial Computing Networks.” *Physica D: Nonlinear Phenomena*, vol. 42, no. 1–3, June 1990, pp. 1–11. DOI.org (Crossref), [https://doi.org/10.1016/0167-2789\(90\)90063-U](https://doi.org/10.1016/0167-2789(90)90063-U).
- “generative, adj.” *Oxford English Dictionary*, Oxford University Press, September 2023, <<https://doi.org/10.1093/OED/6584721568>>
- “Generative AI on AWS.” *AWS*, Amazon, 2023, aws.amazon.com/generative-ai/.
- Glanville, Ranulph. “A (cybernetic) musing: Ashby and the Black Box.” *Cybernetics and Human Knowing*, Vol. 14, 2007, pp. 189-196.

- Huhtamo, Erkki, and Jussi Parikka, editors. *Media Archaeology: Approaches, Applications, and Implications*. University of California Press, 2011.
- Haugeland, John. *Artificial Intelligence: The Very Idea*. The MIT Press, 1989. DOI.org (Crossref), <https://doi.org/10.7551/mitpress/1170.001.0001>.
- Hesse, Mary B. *Forces and Fields: The Concept of Action at a Distance in the History of Physics*. Philosophical Library, 1962.
- Hardesty, Larry. “Explained: Neural Networks.” *MIT News*, MIT, 14 Apr. 2017, news.mit.edu/2017/explained-neural-networks-deep-learning-0414.
- Hughes, Owen. “Generative AI Defined: How It Works, Benefits and Dangers.” *TechRepublic*, TechRepublic, 7 Aug. 2023, www.techrepublic.com/article/what-is-generative-ai/.
- “intelligence, n.” *Oxford English Dictionary*, Oxford University Press, September 2023, <<https://doi.org/10.1093/OED/2404969105>>
- Ilyas, Andrew, et al. “Adversarial Examples Are Not Bugs, They Are Features.” *Advances in Neural Information Processing Systems*, edited by H. Wallach et al., vol. 32, Curran Associates, Inc., 2019, https://proceedings.neurips.cc/paper_files/paper/2019/file/e2c420d928d4bf8ce0ff2ec19b371514-Paper.pdf.
- Kurzweil, Ray. *The Singularity Is near: When Humans Transcend Biology*. Viking, 2005.
- “Large Language Models Explained.” *NVIDIA*, NVIDIA, www.nvidia.com/en-us/glossary/data-science/large-language-models/. Accessed 13 Oct. 2023.
- Latour, Bruno. *Pandora’s Hope: Essays on the Reality of Science Studies*. Harvard University Press, 1999.
- Lyon, Christopher, et al. “Our Climate Projections for 2500 Show an Earth That Is Alien to Humans.” *The Conversation*, The Conversation, 3 Nov. 2022, theconversation.com/our-climate-projections-for-2500-show-an-earth-that-is-alien-to-humans-167744.
- Lee, Edward A. “What Can Deep Neural Networks Teach Us About Embodied Bounded Rationality.” *Frontiers in Psychology*, vol. 13, Apr. 2022, p. 761808. DOI.org (Crossref), <https://doi.org/10.3389/fpsyg.2022.761808>.
- Lakoff, George, and Mark Johnson. *Metaphors We Live by: With a New Afterword*. University of Chicago Press, 2011.
- Larson, Jeff, et al. “How We Analyzed the Compas Recidivism Algorithm.” *ProPublica*, 23 May 2016, www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

- “model, n. & adj.”. *Oxford English Dictionary*, Oxford University Press, September 2023, <<https://doi.org/10.1093/OED/3984201854>>
- “Meet Adobe Sensei GenAI.” *Adobe Experience Cloud*, Adobe, business.adobe.com/products/sensei/adobe-sensei-genai.html. Accessed 18 Nov. 2023.
- Maturana, Humberto R., et al. *Autopoiesis and Cognition: The Realization of the Living*. D. Reidel Publishing Company, 1980.
- Morgan, Mary S., and Margaret Morrison. “Models as Mediating Instruments.” *Models as Mediators: Perspectives on Natural and Social Science*, Cambridge University Press, New York, 1999, pp. 10–36.
- McCarthy, John, et al. “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955”. *AI Magazine*, vol. 27, no. 4, Dec. 2006, p. 12, doi:10.1609/aimag.v27i4.1904.
- McCulloch, Warren S., and Walter Pitts. “A Logical Calculus of the Ideas Immanent in Nervous Activity.” *The Bulletin of Mathematical Biophysics*, vol. 5, no. 4, Dec. 1943, pp. 115–33. DOI.org (Crossref), <https://doi.org/10.1007/BF02478259>.
- Newton, I. *Opticks: Or, A Treatise of the Reflections, Refractions, Inflections and Colours of Light*. William Innys, 1730, <https://books.google.com/books?id=XXu4AkRVBBoC>.
- Novitz, David. “Knowledge and Art.” *Handbook of Epistemology*, edited by Ilkka Niiniluoto et al., Springer Netherlands, 2004, pp. 985–1012. DOI.org (Crossref), https://doi.org/10.1007/978-1-4020-1986-9_27.
- Newman, Benjamin. “Stable Diffusion and Copyright: Wading into Uncharted Legal Waters.” *LinkedIn*, 19 May 2023, www.linkedin.com/pulse/stable-diffusion-copyright-wading-uncharted-legal-waters-newman#:~:text=Stable%20Diffusion%20could%20violate%20two,the%20creation%20of%20derivative%20works.
- Nietzsche, Friedrich Wilhelm. *The Complete Works of Friedrich Nietzsche*. Translated by M. A. Muggle, 18 volumes, Allen & Unwin, London.
- Olah, Chris. “Mechanistic Interpretability, Variables, and the Importance of Interpretable Bases.” *Transformer Circuits Thread*, transformer-circuits.pub/2022/mech-interp-essay/index.html. Accessed 16 Nov. 2023.
- Perrigo, Billy. “Elon Musk Signs Open Letter Urging AI Labs to Pump the Brakes.” *Time*, 29 Mar. 2023, <https://time.com/6266679/musk-ai-open-letter/>.
- “Pause Giant AI Experiments: An Open Letter.” *Future of Life Institute*, 6 Nov. 2023, futureoflife.org/open-letter/pause-giant-ai-experiments/.

- Paul E. Black, "greedy algorithm", in *Dictionary of Algorithms and Data Structures* [online], Paul E. Black, ed. 2 February 2005. Available from: <https://www.nist.gov/dads/HTML/greedyalgo.html>
- Pfeifer, Rolf, et al. *How the Body Shapes the Way We Think: A New View of Intelligence*, MIT Press, 2006. ProQuest eBook Central, <https://www.proquest.com/legacydocview/EBC/3338629?accountid=10598>.
- "representation, n.¹". *Oxford English Dictionary*, Oxford University Press, September 2023, <<https://doi-org-s-a865.shfd.findres.net/10.1093/OED/8146709541>>
- Roberts, Jacob. "Thinking Machines: The Search for Artificial Intelligence." *Science History Institute*, 1 June 2023, www.sciencehistory.org/stories/magazine/thinking-machines-the-search-for-artificial-intelligence/.
- Reilly, Jon. "What Is Narrow Ai? Understanding the Differences between the Three Types of Ai." *Akkio*, Akkio, 5 Mar. 2021, www.akkio.com/post/what-is-narrow-ai-understanding-the-differences-between-the-three-types-of-ai.
- Russell, Stuart J., et al. *Artificial Intelligence: A Modern Approach*. 3rd ed, Prentice Hall, 2010.
- Rudin, Cynthia, and Joanna Radin. "Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson from An Explainable AI Competition." *Harvard Data Science Review*, vol. 1, no. 2, Nov. 2019. DOI.org (Crossref), <https://doi.org/10.1162/99608f92.5a8a3a3d>.
- Scott, Bernard. "Second-order Cybernetics: An Historical Introduction." *Kybernetes*, vol. 33, no. 9/10, Oct. 2004, pp. 1365–78. DOI.org (Crossref), <https://doi.org/10.1108/03684920410556007>.
- "Searching for Light's 'Ether.'" *American Museum of Natural History*, American Museum of Natural History, www.amnh.org/exhibitions/einstein/light/a-new-view-of-light. Accessed 4 Oct. 2023.
- Smith, Alvy Ray. "Plants, Fractals, and Formal Languages." *ACM SIGGRAPH Computer Graphics*, vol. 18, no. 3, 1984, pp. 1–10., <https://doi.org/10.1145/964965.808571>.
- Spearman, Charles. *The Abilities of Man*. Macmillan, 1927.
- Shelley, Mary. *Frankenstein*. Dover Publications, 1994.
- Smelik, Ruben M., et al. "A Survey on Procedural Modeling for Virtual Worlds." *Computer Graphics Forum*, vol. 33, no. 6, 2014, pp. 31–50., <https://doi.org/10.1111/cgf.12276>.
- Sternberg, Robert J. "Intelligence." *Dialogues in clinical neuroscience* vol. 14,1 (2012): 19-27. doi:10.31887/DCNS.2012.14.1/rsternberg.

- Searle, John R. "Minds, Brains, and Programs." *Behavioral and Brain Sciences*, vol. 3, no. 3, 1980, pp. 417–424., doi:10.1017/S0140525X00005756.
- Spataro, Jared. "Introducing Microsoft 365 Copilot – Your Copilot for Work." *The Official Microsoft Blog*, Microsoft, 16 May 2023, blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/.
- Silver, David, et al. "Mastering the Game of Go without Human Knowledge." *Nature*, vol. 550, no. 7676, Oct. 2017, pp. 354–59. DOI.org (Crossref), <https://doi.org/10.1038/nature24270>.
- "Stable Diffusion Online." Stable Diffusion, stablediffusionweb.com/. Accessed 17 Nov. 2023.
- Seaman, Bill, and Otto Rössler. "Neosentience – a New Branch of Scientific and Poetic Inquiry Related to Artificial Intelligence." *Technoetic Arts*, vol. 6, no. 1, May 2008, pp. 31–40. DOI.org (Crossref), https://doi.org/10.1386/tear.6.1.31_1.
- Thomas, Carys. "10 Essential Painting Techniques for Artists." *UAL*: 9 Dec. 2021, www.arts.ac.uk/study-at-ual/short-courses/stories/10-essential-painting-techniques-for-artists.
- The Social Dilemma*. Directed by Jeff Orlowski, performances by Jeff Orlowski et al., *Netflix*, 9 Sep. 2020.
- "The Discovery of The Double Helix, 1951-1953." *National Library of Medicine: Profiles in Science*, National Library of Medicine, profiles.nlm.nih.gov/spotlight/sc/feature/doublehelix. Accessed 3 Oct. 2023.
- "The First Climate Model." *NOAA*, NOAA, 12 Dec. 2006, celebrating200years.noaa.gov/breakthroughs/climate_model/welcome.html#vision.
- "The Impact of Generative AI in Finance." *Deloitte*, Deloitte, 6 July 2023.
- Tegmark, Max. "The Impact of chatGPT Talks (2023) - Prof. Max Tegmark (MIT)." *YouTube*, YouTube, 4 Aug. 2023, <https://www.youtube.com/watch?v=RheZlFj3Zp8>. Accessed 16 Nov. 2023.
- Von Foerster, Heinz. *Understanding Understanding: Essays on Cybernetics and Cognition*. Springer, 2003.
- Von Foerster, Heinz, et al. *Cybernetics: circular causal and feedback mechanisms in biological and social systems*. Josiah Macy, Jr. Foundation, 1953.
- Von Glasersfeld, Ernst, editor. "Introduction." *Radical Constructivism in Mathematics Education*. Springer Netherlands, 2002. DOI.org (Crossref), <https://doi.org/10.1007/0-306-47201-5>.

- Wharton, Annabel Jane. *Models and World Making: Bodies, Buildings, Black boxes*. University of Virginia Press, 2021.
- Waldby, Catherine. “Theaters of Violence” *The Visible Human Project: Informatic Bodies and Posthuman Medicine*. Biofutures, Biocultures. London: Routledge, 2000, pp. 51-80.
- Wartofsky, Marx William. “Introduction.” *Models: Representation and the Scientific Understanding*, Boston: Dordrecht, 1979, pp. 13-26.
- Wolfram, Stephen. *What Is ChatGPT Doing ... and Why Does It Work?* First edition, Wolfram Media, Inc, 2023.
- Wilson, Dennis G., et al. “Evolving Simple Programs for Playing Atari Games.” *Proceedings of the Genetic and Evolutionary Computation Conference*, ACM, 2018, pp. 229–36. DOI.org (Crossref), <https://doi.org/10.1145/3205455.3205578>.
- Wiener, Norbert. *Cybernetics or Control and Communication in the Animal and the Machine*. 2. Ed., 10. Print, MIT Press, 2000.
- Young, Thomas. “II. The Bakerian Lecture. On the Theory of Light and Colours.” *Philosophical Transactions of the Royal Society of London*, vol. 92, Dec. 1802, pp. 12–48. DOI.org (Crossref), <https://doi.org/10.1098/rstl.1802.0004>.