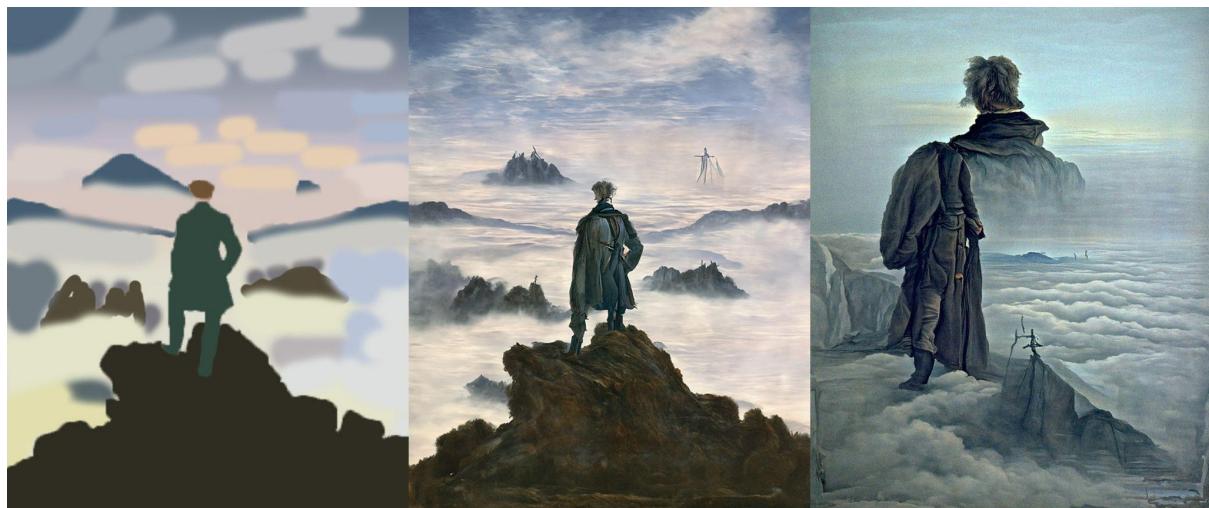


Conversational Investigations

In search of virtual agency



Date	31.03.2022
Name	Neufeld, Artur
Course of study	Bildende Künste
Degree	Master of Fine Arts
Matriculation no.	24554400
Address	Fährstraße 37, 21107 Hamburg
First examiner	Konrad Renner
Second examiner	Hanne Loreck

Hochschule für bildende Künste Hamburg

NOTE: This print is for archival purposes only. This thesis was designed as a website.
Please visit <https://conversational-investigations.netlify.app/> or email
hello@arturneufeld.com for access.

Introduction	3
Historical Perspective	4
Origins and perspectives (Promises)	6
Challenges and pitfalls (AI winters)	10
The soft sound of thawing (AI Spring)	15
Melting the glaciers (State of the art)	18
Taming the transformers or the story of conversational investigations	29
The conversation partner (setting the stage)	30
Talking about the weather (early results)	34
Learning to talk (talking to learn)	39
See you later (regards and closure)	48
Conclusion	50
Bibliography	52

Introduction

In our daily lives, we are mostly unaware that we are interacting with artificial intelligence, and yet AI is already deeply embedded in our lives. Every time you send money, use a search engine, translate a text, receive an email, watch a movie online, or take a photo with your phone, there is a high chance that the information is being processed by an AI. Passively, it is affecting our lives through numerous applications in law, science, health, manufacturing and business.

Remarkably, what used to be considered AI in the past is now just routine and not considered AI anymore, as it's merely one subsystem among many. AI research is developing at a rapid pace and in addition to being used for analysis, decision-making, and processing, AI today can autonomously generate new media – voices, sounds, images, videos, music, and text.

This thesis is an investigation of generative AI, as I examine its proclaimed autonomy in search of virtual agency. To temporarily locate this investigation, I divided this thesis into two parts.

The first part is about the historical perspective, in which I will sketch the past and transition to a more detailed present. Hopefully, we can better understand the context in which we find ourselves today by knowing where this technology comes from and what it implies. This chapter is very Western, even America-centric. The reasons for this are complex and beyond the scope of this thesis. Many AI luminaries were based in America or the UK and were largely funded by the military or the state, which narrowed the research locally. Research in Asia was mainly focused on robotics, with Japan also becoming an AI hotspot later on. It is also worth noting that this chapter focuses on technology and culture – and does not take into account the (geo)political tensions that accompany each period such as during the Cold War.

The second part is a presentation of my own experiments and investigations with a text-to-image AI. I will first connect it back to the now and then present my first steps, process, way of thinking and results of interacting with it. The idea is to capture this emerging technology at a time when it is not yet incorporated into everyday life, to analyze it in its infancy, to understand its inner workings, and to scrutinize its materiality. Ideally, I will find a conversation partner in the AI – where the machine and I act together, share ideas, and benefit from each other's perspectives as independent partners in an equal configuration.

Parts of these thesis will be technical, but as I am dealing with a technical entity, it is the nature of the subject I am examining. If you want to describe a sculpture you should be aware of its material.

Historical Perspective

The concept of intelligent artificial life is not a modern idea, but rather has its roots in culture (mythology, fiction, religion) and philosophy.

Intelligent beings, generally in the guise of animals or other non-human entities, were popular tales and stories. The Greek fable of Pygmalion and Galatea, in which a sculptor falls in love with a statue he has created. The Norse legend of the reanimation of Mímir's head to provide counsel to Odin. The Jewish folklore of the golem, a clay creature that a rabbi may bring to life. The story of alchemists creating 'artificial men', called Homunculi. Humanoid Automata which were built through the ages, believed to possess very real minds, capable of wisdom and emotion¹. Novels like Mary Shelley's Frankenstein and Edgar A. Poe's Maelzel's Chess Player popularized the ideas of artificially created intelligent entities.

Philosophical predecessors such as Aristotle², who developed the idea of formalism³, laid the foundation upon which Gottfried Wilhelm Leibniz, Thomas Hobbes, and René Descartes built the hypothesis of a physical symbol system in the 17th century. They explored the possibility that all rational thought could be made as systematic as algebra or geometry⁴. Deeply connected to the mind-body problem, investigations abstracted the logic of thought further, resulting in George Boole's book *Laws of thought* which aimed to construct "the mathematics of the human intellect" and to "minimize the ambiguity of natural language" (McCorduck, 2004, p. 49). The influential work of Alfred North Whitehead and Bertrand Russell *Principia Mathematica* built on Boole's work and asked the fundamental question whether one can formalize all of mathematical reasoning and thus have logical systems that can mathematically embed complex systems like language and thought. In the beginning of the 20th century, Gödel's *Incompleteness Proof* described that any mathematical system is incomplete and has limits to what it is able to accomplish. But within these limits, any form of mathematical reasoning can be formulated. This led many scientists to believe that thinking machines might be theoretically possible when a mathematical description of thought could be devised.

¹ Hermes Trismegistus, the Hellenic author of the Hermetic writings, speaking of animated statues: "They have sensus and spiritus. ... By discovering the true nature of the gods, man has been able to reproduce it... unable to create souls, man invoked the souls of demons and angels and, by sacred rituals, infused them into the statues which thereby acquired the power of doing good or evil" (McCorduck, 2004, p. 8)

² As McCorduck points out, it should be noted that "He was by no means the first to think of human intelligence in terms of mechanisms. As in so many other things, the Greeks offer evidence of being first at that, in particular the philosophers of the Epicurean school and, from another point of view, the physicians of the Hippocratic school" (McCorduck, 2004, p. 38)

³ Bringsjord and Arkoudas summarize formalism as: "The notion that certain patterns of logical thought are valid by virtue of their syntactic form, independently of their content" (Bringsjord et al., 2007, p. 4)

⁴ Leibniz especially recognized that, in order to achieve a universal scientific language, you cannot rely on propositions of logic alone, but the necessity to confer the relations between them symbolically (McCorduck, 2004, p. 41). This notion is interesting when comparing this paradigm to modern natural language processing AI's which are 'tokenizing' the relations of words - finding a symbolic representation of the context, rather than using the linguistic logic of any language.

It is important to recognize that besides these philosophical observations, which also ultimately bled through and welded with mathematics, other fields of science advanced in parallel, which also inspired this way of thinking. The preceding developments in biology and science showed astonishing parallels between the construction and functional principles of the human body and technical inventions⁵. Advancement in computer science and cybernetics in the 1930s implemented the mathematical and theoretical groundwork into actual working logical machines - machines that we now call computers⁶. While earlier mathematicians had theoretical and mechanical implementations, Alan Turing's *Theory of Computation* and Claude Shannon's *Information Theory* proved that the same can also be applied and achieved on digital computational surfaces. Parallel efforts in neurology, heralded by Warren McCulloch in the late 1940s, showed that the brain itself was a network of neurons that operated with binary, electrical, on-or-off signals. These thematic parallels were perfect ingredients for cautious optimism among researchers: An electric brain was not a far fetched idea anymore.

Even though the cultural and philosophical backdrops I tried to paint above are very much condensed and abstracted, I think that it is important to understand the roots of the technology I am investigating in this thesis. It is interesting to me how far back one can trace the origins of the idea of artificial life and that it is not at all a modern idea. Backtracking the cultural and philosophical influence on the subject, and seeing it coagulate in the late 1800s and early 1900s, one can comprehend how in that time optimism was open to scrutiny, since it seemed that multiple pieces to an ancient puzzle came together.

Origins and perspectives (Promises)

⁵ Fritz Kahn's "Der Mensch als Industriepalast" from 1926, while pushing the idea of man as machine to the extreme, serves well as a cultural contemporary witness to illustrate this way of thinking. 21 years later Alan Turing entertained the same 'man as machine' logic by pointing out various machines that already imitate parts of man — the television camera as eye, the microphone as extended ear, and servomechanistic robots that imitate actions of the limbs — but dismissed this approach as it "would be rather like putting a lot of work into cars which walked on legs instead of continuing to use wheels." (McCorduck, 2004, p. 68)

⁶ As it is with many technological breakthroughs, early computers were designed and used as machines of war. In this case computers like Z3, ENIAC and Colossus were exclusively used as code breaking machines in World War II.



A child's crayon drawing on paper containing various objects like cubes, cones and balls. These objects are colored red, green and blue. (DISCO)

When we speak of artificial intelligence in the modern sense, we are in the fortunate position that the term itself has been deliberately coined: Namely at the influential Dartmouth Conference in 1956, where John McCarthy, Marvin Minsky, Nathaniel Rochester and Claude Shannon proposed the term "Artificial Intelligence" (McCorduck, 2004, p. 17).

In the period between 1956 and 1974, researchers succeeded in advancing the field significantly. It was during this time that the foundations for many of the sub-disciplines within AI were laid, such as machine learning, natural language processing, and knowledge representation. Multiple advancements were achieved in just two decades: The first expert systems and game AI were introduced, programs that are specifically designed to solve single problems, mostly operating in systems with known bounds and rules. First checker and chess AI's could beat amateurs⁷. The program *Logic Theorist*, which is

⁷ To this day, Game AI's serve as a benchmark to monitor the progress of AI's. Fun fact: To this day, no human has beaten a computer in a chess tournament since 2006.

considered by some to be the first AI program, would eventually prove 38 of the first 52 theorems in previously mentioned Whitehead and Russell's *Principia Mathematica*, and find new and more elegant proofs for other theorems. An early language processing AI called ELIZA, the first implementation of what we today call a chatbot, was planned to be used as a psychotherapeutic tool, could hold a conversation to a degree that early testers occasionally thought that they were chatting with a human being and not a computer program. Semantic networks and so called "micro worlds" (Crevier, 1993, p. 73) were artificial systems that seemed to be aware of context and space, though strictly bound to their own embedded and simplified environment. They were able to distinguish between simple shapes, space and color, while using a human readable in- and output. One of these natural language understanding programs was SHRDLU developed in 1968 which rendered a "block world" containing basic shapes of different colors and sizes. To interact with the program, you wrote the instructions in plain english. An excerpt of a demonstration goes as follows:

Person: Pick up a big red block.
Computer: OK.
Person: Grasp the pyramid.
Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
Person (changing their mind): Find a block which is taller than the one you are holding and put it into the box.
Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.
Computer: OK.
Person: What does the box contain?
Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.
Person: What is the pyramid supported by?
Computer: THE BOX.
Person: How many blocks are not in the box?
Computer: FOUR OF THEM.
Person: Is at least one of them narrower than the one which i told you to pick up?
Computer: YES, THE RED CUBE.
[...]

SHRDLU was considered a success and temporarily contributed to optimism in this research area (Crevier, 1993, p.100).

Another important achievement of this time was the so-called Perceptron. Developed by American psychologist Frank Rosenblatt in 1956 and funded by the US Navy, the machine was highly inspired by the state of neuroscience of that time⁸. It modeled a basic version of a single artificial neuron with adjustable weights and a threshold and was designed for image recognition. It consists of

⁸ For one there was aforementioned Warren McCulloch who familiarized the idea of an all-or-nothing binary nature of neurons, while Donald Hebb influenced the hypothesis that learning occurs in the brain primarily through the formation and change of synapses between neurons.

an array of 20 by 20 photocells, which resulted in a 400-pixel input camera. Each of these photocells were randomly connected to the ‘neurons’. Potentiometers, adjusted by electric motors during ‘learning’, were used to update weights. The idea was that one would input two cards whose contents you wanted to be classified. Each photocell would map a section (or pixels, in today’s terms) of the card, while the motors would adjust the weights of the corresponding neuron depending on the excitation of the photocell. Since the neurons start with random weights, this process, the learning part, would be repeated until the sum of the weights crosses a threshold value⁹. If you now put in other cards, the machine would be able to compare the state of photocells against the weights it stored in memory - if it fits the same distribution pattern and crosses the threshold again, it would classify the card again as the same type. The result was a machine that could classify simple binary categories, such as left or right, triangle or circle, the letter A or not the letter A¹⁰.

Rosenblatt’s invention is now known to be the first implementation of an artificial neural network¹¹ and also gained notoriety during his time. While his results were promising and researchers intrigued by its simplistic design, scientists were divided when it came to his persona and media appearances, resulting in a heated debate in the AI community. Rosenblatt was known for giving steady and extravagant statements about the performance of his invention (McCorduck, 2004, p. 105). During a press conference organized by the Navy, Rosenblatt stated the Perceptron would be “the embryo of an electronic computer that [the Navy] expects will be able to walk, talk, see, write”, that the Perceptron might be used as an autonomous space explorer, that “it would be possible to build brains that could reproduce themselves on an assembly line and which would be conscious of their existence”. The New York Times and the New Yorker reported on these claims, which were a contributing factor in bringing the optimism prevalent in the niche AI community to a wider audience.

The sixties were not only a blooming period for research in the field of artificial intelligence, but were also accompanied by other technological advancements. The first robots were introduced in an industrial scope, the first commercial geosynchronous satellites were launched, the first public demonstration of the computer mouse took place, LED’s were developed and by 1969 the first human set foot on the moon. Science-fiction and pulp magazines prospered, the *Jetsons*, the futuristic pendant to the *Flintstones*, was one of the first shows which were broadcasted in color, the original series of *Star Trek* and films like Stanley Kubrick’s *2001: A Space Odyssey* popularized AI in film.

⁹ According to a newspaper article from 1958 it took 50 repetitions to differentiate between left and right.

¹⁰ Since this endeavor was financed by the American military complex, I think it is fair to assume that the goal was to differentiate between adversary and friendly units or at least to find an automated detection system.

¹¹ Arguably, Marvin Minsky may have been the first researcher to implement a neural network in hardware, the SNARC (Stochastic Neural Analog Reinforcement Calculator) in 1951, which preceded Rosenblatt’s work by many years. But the absence of any trace of his work on this system and the critical nature of the analysis in *Perceptrons* suggests that he concluded that this approach to AI was a dead end. (Kurenkov, 2020)

Generally speaking, the perception of AI in the general public on the other hand was accompanied by undertones of technological paranoia, a combination of fascination and fear. Computers were largely seen as huge, room-sized machines, anthropomorphized as giant brains, used as abstract calculators and code-breakers in war efforts. The inner workings were not known to the average citizen and “any advance in computing technology was immediately hyped into the assumption that computers would soon take over the human decision-making process” (Wiegel, 2012).

Marvin Minsky, one of the organizers of the Dartmouth Conference and a luminaire in the field till his death in 2016, predicted in 1967 that “within a generation few compartments of intellect will remain outside the machine’s realm - the problem of creating ‘Artificial intelligence’ will [be] substantially solved.” and later was (mis)quoted in a notorious article in Life Magazine as saying “with quiet certitude: ‘In from three to eight years we will have a machine with the general intelligence of an average human being’” (Crevier, 1993, p. 96) After seeing this quote in print, Minsky tried hardest to denounce the article as he denied ever making this statement. Even though the AI community felt victimized by this sensationalist representation of their work, the damage was already done. Said article was subsequently published in an anthology and appears in at least one college textbook (McCorduck, 2004, p. 273).

In conclusion, the optimism within the AI community was admittedly high, but unsustainable promises were made and expectations were running wild.

Challenges and pitfalls (AI winters)



A lonely landscape during a winter storm, freezing cold, oil painting by Thomas Moran, Albert Bierstadt and Artur Neufeld, featured on Artstation. (DISCO)

Although basic research in AI has undoubtedly come a long way in a short period of time, many of the developments were, on closer inspection, rather proof of concepts that only worked on small scales, with fixed boundaries and strong limitations.

ELIZA, for example, was indeed able to convince the occasional tester into thinking that they were corresponding with a human, the conversations in their entirety were repetitive and shallow. Eliza simulated a conversation by using a "pattern matching" and substitution method that gave users the illusion that the program would understand them, but had no built-in systems for contextualizing. The data set from which ELIZA drew its responses were pre-programmed and based on actual doctor-patient conversations. In general, ELIZA would recognize keywords, which then would trigger one of the prepared answers. Often, ELIZA would just rephrase the user's replies as questions, which some users

unconsciously (mis)interpret as interest and emotional involvement – Even though it was explained to them beforehand that the program could not converse with true understanding and cannot simulate emotions. This form of anthropomorphisation is called the ELIZA Effect. Current chatbots rely on this effect in order to function, as many of them basically use a similar pattern matching approach as ELIZA.

As I mentioned earlier, most western AI research was funded by the US government, which tried to gain technological superiority during the Cold War. One of these efforts involved aggressive support for machine translation, as they were interested in the automatic and instant translation of Russian documents and reports. While early results, again, seemed promising, it was thought, that "simple syntactic transformations based on the grammars of Russian and English, and word replacement using an electric dictionary, would suffice to preserve the exact meanings of sentences" and "that 'scaling up' to larger problems was simply a matter of faster hardware and larger memories" (Russell et al., 2003, p. 21). But translation requires general knowledge of the topic in order to resolve ambiguity, meaning and content and researchers underestimated the difficulty of this word-sense disambiguation¹². This would later be coined as the commonsense knowledge problem. By 1964 progress nearly came to a halt and by 1966, after spending millions of dollars, all funding and translation projects were canceled. The Lighthill Report of 1974 for the British Science Research Council tried to evaluate the current state of AI and painted a similar pessimistic outview about its ability to scale for real-world problem solving. This report was the basis on which the British government decided to end support for AI research (in all but two universities).

Other AI Experiments also failed to scale to larger problems. The SHRDLU Micro World experiment faced the same fate, as it also was unable to overcome the word-sense disambiguation and it became clear that it could only work properly in its own shallow and abstracted environment. It did not seem to matter how much additional knowledge about its environment you'd feed into the machine it "didn't get any smarter" (McCorduck, 2004, p. 390). There seemed to be a barrier impossible to overcome.

As these problems with AI were tied to funding, the belief in AI among researchers was dampened when Marvin Minsky and Seymour Papert published their book *Perceptrons: an introduction to computational geometry* in 1969. In this book the authors analyzed Rosenblatt's Perceptron approach on a fundamental level and "proved that although perceptrons could be shown to learn anything they were capable of representing, they could represent very little." They showed mathematically that the model was limited and would not be able to function properly when faced with more complex problems.

¹² One famous example which illustrates the challenges of that time is the retranslation of "the spirit is willing but the flesh is weak", which was translated to "the vodka is good but the meat is rotten" (Russell et al., 2003, p. 21).

These four reasons, the over-inflated expectations of media and public, the inability to scale to real-world problems, the resulting freezing of state/military funding and the crumbling belief in the mathematical foundations, were the main reasons for a period of silence in research, which is referred to as the first AI Winter (1974-1980) (Crevier, 1993, p. 203).

Research in general-purpose, problem-solving AI reached its freezing point, and it took nearly a decade to return to melting temperature. While general purpose AI seemed to be unachievable, it was apparent that these systems could be useful when applied to precisely defined problems. Early experiments in 1965 and 1972 demonstrated that AI systems are capable of performing automated decision processes as long as their knowledge base is accessible through logical rules and formulas, thus avoiding the commonsense knowledge problem. DENDRAL was developed in 1965 to help chemists identify new organic molecules based on the elementary formula of the molecule and its mass spectrum analysis. MYCIN was developed in 1972 to diagnose blood infections, but in contrast to DENDRAL, MYCIN knowledge base had “no general theoretical model [...] from which the MYCIN rules could be deduced. They had to be acquired from extensive interviewing of experts, who in turn acquired them from direct experience of cases” (Russell et al., 2003, p. 23). With about 450 rules, the AI was able to detect infections better than the medical residents and almost as well as the senior physicians. Since these AIs relied on a foundation of knowledge provided by experts, they were called expert systems¹³.

Although this approach was only distantly related to the type of AI envisioned in the 1960s, these systems actually proved useful, scalable, and applicable to real-world problems. They had another advantage earlier AI systems did not have: The interest of the corporate sector. Their promise went as follows: As long as you were able to describe your problem in a specific set of rules, there was a good chance you could use an expert system to make workflows, planning, design, or monitoring more efficient – and thus more cost-effective. And indeed, these systems have been used, for example, to automate the configuration of orders, to automate fraud detection in banks or to efficiently design new circuit layouts for electronics. The first commercial expert systems were available by 1982 and by end of the decade two thirds of the Fortune 500 companies were using this technology. An entire industry emerged, with suppliers, manufacturers and hard- and software developers. Many of the corporations had in-house AI departments. The fate of this technology is somewhat unclear, as the terminology has generally been dropped by the IT community. They are still used today, not as a standalone application, but rather as one of many tools used in the context of business process automation.

As this new industry developed, the frozen research ambitions for general problem solving AI began to thaw, as Japan's Ministry of International Trade and

¹³ Many systems and programs evolved from DENDRAL and MYCIN and are still used today. Though these programs are not considered AI anymore. They are nowadays referred to as 'structure searchers' for solving mass spectrometry inverse problems.

Industry began its “Fifth Generation” project (Russell et al., 2003, p. 24). The goal of this eager ten-year plan was to develop a next-generation computer that would advance Japan’s efforts to become a leader in the global computer industry.

Millions of dollars were funneled into the project, which “proposed to achieve full-scale natural language understanding, among other ambitious goals” (Russell et al., 2003, p. 24) by pursuing the idea of parallel processing computation. Fear of Japanese domination led to similar initiatives in the US, Europe and the UK.

As funding returned, interest in neural networks was rekindled by parallel breakthroughs. For one, advances in the development of semiconductors, printed circuit boards, and integrated circuits enabled more practical and viable development in the field. Additionally, a new alternative to Rosenblatt’s version of an artificial neuron was invented that corresponded to new findings in psychology and neuroscience¹⁴. A new algorithm called backpropagation was invented and introduced a more efficient method for training, error-correcting and fine-tuning neural networks¹⁵. Further research not only showed that this approach could be used for multilayer neural networks, but also specifically addressed and solved the problems discussed by Minsky and Papert in *Perceptrons*. This new method paved the way for more robust neural networks used in optical character recognition and speech recognition. By the late 1990s, one of these systems was reading up to 20% of all the checks in the US (Piatetsky, 2004).

The rise of expert systems seemed unstoppable, but the enthusiasm of companies also worried researchers, as expectations for these systems once again proved impossible to meet. Numerous start-ups and companies emerged, the stock market exploded, and when an expert system seemed unfeasible, other entrepreneurs jumped in and promised solutions based on other buzzwords such as robotics and neural networks. “A young and fragile science was being asked to perform miracles that were then beyond its powers”¹⁶ (McCorduck, 2004, p. 435). In hindsight, these developments showed the pattern of an economic bubble. Expert systems only proved to be sustainable and useful in a few special cases - most integrations turned out to be too inflexible: they were hard to update, expensive to maintain and too slow for change. The nail on the coffin for

¹⁴ John Hopfield developed what we now call a 'Hopfield Network', a special form of neural network that has an 'associative' memory and is studied as a model for understanding human memory. In 1986 psychologists David Rumelhart and James McClelland published “Parallel Distributed Processing: Explorations in the Microstructure of Cognition” in which they described the creation of computer simulations of perceptrons, providing computer scientists with the first verifiable models of neural processing, and which is now considered a central text in the field of cognitive science.

¹⁵ I mention this algorithm specifically, not only because it had a great impact on the field, but also because I stumbled upon this term back in 2018 when I did my first experiments with AI. Back then it was just something to me ‘that you use’, I had no idea that it was a milestone achievement that dates back to the 80s.

¹⁶ McCorduck points out that “[...] it was [not] all smoke and mirrors. The science was young and fragile, not fraudulent” - Technologies from these days are still affecting us today, “The international credit card system, for example, would simply be impossible without a combination of neural nets and expert systems to approve credit and detect fraud”. From washing machine to rice cooker - “These early AI applications have become internal or background technology nearly everywhere” (McCorduck, 2004, p. 435)

this industry was the rise of desktop and personal computers: By 1987 they were less expensive and more powerful than the very specialized hard- and software used for expert systems. Basically overnight the industry collapsed.

The ten-year plan of the Fifth Generation project was completed in 1991 without meeting the list of ambitious goals. Once again, expectations were not met, confidence was lost, and optimism crumbled, leading to a second AI winter (1987-1993).

In the following two decades, AI research played its role rather quietly and in the background of the scientific and corporate stage. AI researchers had earned a reputation for making empty promises. They were hesitant to make any predictions at all and avoided any mention of 'human-level' artificial intelligence for fear of being labeled "wild dreamers" (Markoff, 2005). Nevertheless, the dream of human-level intelligence from the 1960s echoed, though there was uncertainty among researchers about the exact reasons for its failure. Subsequently, new sub-disciplines of research fragmented the field of AI into parts, each focusing on different approaches, subjects or methods. Many successful projects tarnished their connection to AI by rebranding¹⁷ and relabeling their work as knowledge-based systems, cognitive systems or computational intelligence. Despite its damaged reputation, AI development managed to prosper, but behind the scenes.

¹⁷ Since the field was fragmented into different and new approaches, you could argue that there was substantial reason for renaming. On the other hand, you could also argue that it just helped to acquire funding.

The soft sound of thawing (AI Spring)



The soft sound of thawing allows the first sproutlings to bloom, by Ivan Shishkin and Artur Neufeld, Trending on artstation. (DISCO)

[...] today many thousands of AI applications are deeply embedded in the infrastructure of every industry. Most of these applications were research projects ten to fifteen years ago; People who ask, "Whatever happened to AI?" remind me of travelers to the rain forest who wonder, "Where are all the many species that are supposed to live here?" when hundreds of species of flora and fauna are flourishing only a few dozen meters away, deeply integrated into the local ecology."

Ray Kurzweil (2005)

After testing a variety of different architectures, paradigms and approaches over the past decades, with only a select few proving applicable to specifically selected cases, researchers began to establish a solid mathematical foundation

and formalize their methodology¹⁸. This not only enabled better and faster collaboration, but also invited other scientific fields to join the effort. AI was no longer isolated from the rest of computer science and was now "firmly under the scientific method" (Russell et al., 2003, p. 25), meaning that hypotheses could be tested and compared statistically and empirically. It was possible to "replicate experiments by using shared repositories of test data and code" (Russell et al., 2003, p. 25). While earlier research focused on semantics and symbology, researchers realized that the problems of artificial intelligence were not specific to this field and that comparable advances from economics, mathematics, and electrical engineering could be used. In particular, tools of stochastics and decision theory were used in place of Boolean logic, machine learning in place of programming by hand, and experimental results in place of philosophical assumptions (Russell et al., 2003, p.24). These new approaches proved invaluable and are now an integral part of the methodology.

Armed with these more sophisticated mathematical methods, sub-disciplines of AI resurrected ideas and approaches that seemed to be dead ends before.

There were several reasons that allowed the field to bloom. For one, the aforementioned invisibility of AI research to the public and media took the pressure off, there was almost no talk about expectations, no far-fetched visions of the future. AI-like architecture simply became one of many tools that were integrated into research and also took root in the consumer media and larger systems. When faced with isolated problems, AI turned out to be able to solve them¹⁹. AI slowly infiltrated the technology industry, leading to the success of automated systems in banking, medical diagnostics, traffic control and consumer relations. But the accomplishments of their products were nearly never credited to AI, even when systems heavily relied on them. The mention of AI still invoked ideas of science fiction in the general public as they were unaware of the fact that they were using this technology on a regular basis. As with all consumer oriented products, the end consumer is only receiving the output of the system, unable to peek behind the curtains. By 2006 a lot of cutting edge AI had trickled into general applications, often without being called AI: "Once something becomes useful enough and common enough it's not labeled AI anymore." (Nick Bostrom

¹⁸ In AI research, there is a distinction between those who think that "that AI theories should be grounded in mathematical rigor" - the *neats* - and those who have a more undirected and experimental approach, who "would rather try out lots of ideas, write some programs, and then assess what seems to be working" - called the *scruffies*. Beginning in the 90s, the shift of paradigm and the development of a formal system is often characterized as a "victory of the neats" and has influenced the success and development of AI. Russell et al. say that both approaches are important, but "a shift toward neatness implies that the field has reached a level of stability and maturity. Whether that stability will be disrupted by a new scruffy idea is another question." (Russell et al., 2010, p.25)

¹⁹ AI that serves a single purpose (or limited task) is commonly referred to as 'narrow AI.' In contrast, 'strong AI' can refer to artificial general intelligence capable of applying intelligence to any problem, although the term can also refer to artificial human-like intelligence, artificial superintelligence, or even artificial consciousness, depending on the context or domain. The former is the type of AI we deal with on a daily basis, while the latter refers to and remains a hypothetical machine.

in an interview with CNN, 2006). Also in other scientific disciplines, AI's were indispensable instruments, but “were simply considered valuable automatic helpers”, only “few people credited artificial intelligence for having borne and nurtured [these programs]. If you could see how it was done, people seemed to think, then it couldn't be intelligence — a fancy that many people entertain to this day” (McCorduck, 2004, p. 423)

In the early 90s a paradigm shift started among AI scientists, which was widely accepted in the following years. The prevailing AI paradigm in the 1980s was that of symbolic thought programs executed by a rational agent to achieve a goal – the knowledge was fixed, the goal was predetermined (McCorduck, 2004, p.472). The notion of mimicking cognitive behavior, ‘thinking humanly’, was replaced with the general idea of ‘acting rationally’. A new paradigm called ‘intelligent agents’ proposed a more granular view on the complexity of programmatic reasoning: “An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators” (Russell et al., 2020, p. 36). Earlier AI agents were isolated and had a set knowledge base. In contrast, this approach aimed to put emphasis on knowledge exchange and active influence on its environment, importantly among multiple agents. Agents have their individual objectives, means of measuring performance, information sharing, sensors for input and output, and try to maximize their objective based on feedback from other agents. Part of this term was borrowed from economics, where a ‘rational agent’ is any entity that can make decisions, commonly a person or a firm, that strives to perform optimal actions based on given premises and information. This definition was paired with the idea of modular programming in computer science, where each functionality of a piece of software is strictly separated from other features and is able to act independently. This paradigm shift was a generalization of earlier definitions of artificial intelligence that focused not on the study of human intelligence, but all types of intelligence²⁰. From a philosophical point of view, this redefinition does not refer to human intelligence and does not imply that a machine has a mind, a consciousness, or even a real understanding. That way, there is no difference between simulated and real intelligence or even between intelligent or unintelligent. Rather, agents are measured only by their goal and their performance in achieving that goal. Architectures like this also enabled better collaboration by providing a reliable and scientific way to test these programs: Their performance could be compared, and their modules and goal functions could be modified, combined, or shared. Agents can be as simple or as complex as you need them to be, a thermostat and a self-driving car use the same basic underlying idea.

²⁰ The introduction text on intelligent agents in one of the most used academic textbooks, Artificial Intelligence: A Modern Approach by Stuart Russell and Peter Norvig, describes it quite nicely: “A human agent has eyes, ears, and other organs for sensors and hands, legs, vocal tract, and so on for actuators. A robotic agent might have cameras and infrared range finders for sensors and various motors for actuators. A software agent receives file contents, network packets, and human input (keyboard/mouse/touchscreen/voice) as sensory inputs and acts on the environment by writing files, sending network packets, and displaying information or generating sounds.” (Russell et al., 2020, p. 36)

Melting the glaciers (State of the art)



A magnificent glacier melts and flows into a river as the sun shines in bright rays through the clouds, oil painting by Thomas Moran, Albert Bierstadt and Artur Neufeld, featured on Artstation. (DISCO)

In the early 2000s, artificial neural networks (ANNs) made a comeback. Previous research had successfully implemented and tested multilayer ANNs (also in real world applications, like the check-reading AI mentioned previously), but access to more computing power and improved machine learning algorithms accelerated their development. Similar to Rosenblatt's Perceptron, these networks are inspired by biological neurons (and consist of components such as neurons, synapses, weights, biases, and functions)²¹, but have the advantage of being

²¹ Importantly, these concepts are inspired by the human brain, but are not in any way equivalent to biological neurons. These systems are studied by neuroscientists, but it is still up for debate whether the current approach reflects the inner workings of biological neurons. A paper from 2016 by Bengio et al. states that "Neuroscientists have long criticised deep learning algorithms as incompatible with current knowledge of neurobiology", as they say, for example, that back-propagation in the form used in machine learning "is not biologically plausible".

able to have multiple inputs and outputs, as well as multiple layers of artificial neurons in between. Each layer passes its computational results to the inputs of the next layer. Although this architecture appears simple, this change leads to what might be called abstraction. Since additional layers have no ‘idea’ of the initial image (or input) and are only connected to the previous artificial neuron, these neurons can represent more comprehensive patterns. For example, a conventional neural network for text recognition without these layers would require specific neurons for each serif of a letter, while a multilayer architecture would combine the appearances of that pattern into a single neuron later in the network that is able to ‘see’ the same pattern even when translated or rotated. The fact that multiple layers can be stacked between input and output has led to this architecture also being referred to as “deep learning” (Russell et al., 2020, p. 26).

This deep learning architecture, in combination with machine learning algorithms and mathematical tools from stochastics started to play a bigger role. Machine learning is an integral part of AI and refers to an algorithm or agent that uses input data to automatically improve through experience by being provided with sample data, or training data, that sufficiently captures the goal. Using statistical means, the agent can then not only find patterns but also use experience to make predictions or decisions without being explicitly programmed to do so.

Heuristic and stochastics-based approaches were pursued in the 80s, but faced theoretical and practical problems of data acquisition and representation and were overshadowed by the success of symbolic/knowledge based expert systems that rely on production rules, decision trees, and logical formulae (Langley, 2011, p. 276). But in 1995 researchers tackled the word-sense disambiguation problem again and showed that a heuristic approach is a viable direction. While earlier machine learning approaches relied on a supervised training data set, meaning that each disambiguation was manually differentiated and classified, they showed that, given that the data set is large enough, unsupervised learning is possible. Their method was able to correctly identify the meanings of words (E.g. whether the word “plant” refers to a factory or flora) in 96% of the cases (Russell et al., 2003, p. 26). In 2001 other researchers showed that “a mediocre algorithm with 100 million words of unlabeled training data outperforms the best known algorithm with 1 million words” (Russell et al., 2003, p. 26). This approach shifts the focus of the object of study away from the algorithm itself and instead onto the data used for training.

Meanwhile, two factors accelerated the creation and consumption of very large amounts of data. First, the rise of the World Wide Web enabled the accumulation of large data sets that were inaccessible by earlier methods, as storage shifted from analog to digital across all sectors of society and industry²². The need to process these huge amounts of data is what led to the phenomena of *big data* and learning algorithms were specially designed to take advantage of very large

²² In 2013, [Facebook alone processed and stored over 50 billion images](#) [1], produced 120 terabytes of logs daily and handled hundreds of millions of requests per second.

data sets. Secondly, parallel increases in computing power allowed for giant technological leaps that accelerated machine learning. From the 1960s until about 2012, the amount of computing power used to train machine learning applications followed roughly the same trend as Moore's Law: computing power doubled about every 18 months, while costs dropped by half over that period (Russell et al., 2020, p. 14). But from 2012 to 2018 computational power increased by a factor of 300,000 (which roughly translates to a doubling of computational power every 100 days). A machine learning model that took a full day to train in 2014 took only two minutes in 2018 (Russell et al., 2020, p. 15). This exponential acceleration has been made possible in part by the development of hardware specifically designed for parallel computing and machine learning, called tensor processing units (TPUs)²³, and by the use of graphics processing units (GPUs).

This exponential increase of computational power may be a blessing for AI researchers, but its implications are also alarming. The best natural language processing (NLP) algorithm in 2018, BERT, was trained on a data set of 3 billion words and was bested one year later by GPT-2, which relied on training data of 40 billion words. Again, only one year later, GPT-3 outshined their competitors with a data set of 500 billion words. In terms of complexity, GPT-2 had 1.5 billion trainable parameters, while GPT-3 had 175 billion, an increase of complexity of roughly 11600%. As the amount of parameters of each model increase, so does the power required for computation, training and inference²⁴. Hardware manufacturers estimate that ≈85% of the computation power is devoted to inference and only the remainder to training (Toews, 2020). A 2019 study by Emma Strubell et al. found that training a single machine learning model can produce up to 620,000 pounds of CO₂²⁵, which is equivalent to the life-time carbon footprint of five cars²⁶. The training of a single BERT base model (without fine-tuning) was estimated to require as much energy as a trans-American flight, while the cloud computing costs are estimated to be between ~\$3500 and \$12500 (Strubell et al., 2019, Table 3). A normal workflow for creating these models requires a lot of fine-tuning and optimization through trial and error. The study showed in an example that the process of creating and testing a final, paper-ready model required 4,789 training runs. To speed up training, these calculations are performed on multiple GPUs simultaneously: "The sum GPU time required for the project totaled 9998 days (27 years). This averages to about 60 GPUs running constantly throughout the 6 month duration of the project"

²³ TPUs were developed and produced by Google to speed up big data processing and facilitate their own AI applications. Their hardware was initially proprietary, but was released for use by third-parties in 2018. The need in the industry for so-called AI accelerators becomes clear when considering the release of Apple's new MacBook Pro in 2021, as the company developed, produced and shipped its very own hardware, called a neural processing unit (NPU). Since 2017, Huawei cell phones have their own NPU.

²⁴ Inference means deploying AI models to take action in real-world settings. Basically running the AI on hardware: Using it after the model is trained.

²⁵ To give some perspective, a human's yearly CO₂ consumption ranges between 11000 and 36000 Pounds, depending on your demographic and location.

²⁶ To be fair, these numbers refer to the most power hungry NLP model that was available at the time of the study. The estimated cloud computing costs for this model range from \$940,000 to \$3,200,000.

(Strubell et al., 2019, p. 4). The AI community seems to be moving toward measuring progress in AI by the development of ever-larger models. This may have helped to break the icy cocoon of AI winter, but the price of this melting of the glacier may not be merely figurative.



The ocean is on fire, abandoned, burning oil, by Andreas Achenbach and Artur Neufeld. (DISCO)

However, this "bigger is better" ethos may have already reached its zenith, as more and more results suggest that huge data sets have crossed the threshold of diminishing returns (Toews, 2020). A well-known model for computer vision called RESNET was released in 2015, its successor RESNEXT two years later. The latter required 35% more computational resources for training, but increased accuracy by only 0.5% (Toews, 2020). Similar results are evident when comparing the increase in training data size in GPT-3 with the increase in accuracy. Even though the increase in accuracy happens in minor increments, it is still one of the industry's most important benchmark parameters. The comparability of models leads to a side effect, as there seems to be a fixation on achieving 'state-of-the-art' performance. Results are tracked in popular

leaderboards, and an increase of just a fraction of a percent can bring researchers recognition, acclaim, and industry attention. These benchmarks are mainly about model accuracy; efficiency and (environmental) costs are ignored (Schwartz et al., 2019). Even though the creators of GPT-3 are aware of the “fundamental limitation of the general approach” (Brown et al., 2020, p. 33) and are questioning their own methodological direction, “the improved results of GPT-3 on many tasks are likely to fuel, not abate, the [industry’s] desire for bigger and bigger neural networks” (Ray, 2020).

GPT-3 is currently the state of the art for NLPs and is capable of producing genuinely compelling text. A small 2020 study of 80 participants found that they were unable to correctly distinguish AI-generated text from human-generated text 48% of the time, only marginally better than random guessing (Brown et al., 2020, p. 26). Since the AI was trained on an unfathomable large corpus of text it is, in contrast to other models, capable of a large spectrum of text based applications (article writing, summarizations, text completion, style transfer, poetry), as well as very specific tasks if correctly fine-tuned (code writing, story telling, chatbots, interactive narratives). In 2019, OpenAI, the creators of the model and a non-profit organization back then, went against their open-source practices, and decided to not publish the full model due to fears that the model would further accelerate the spread of fake news. Instead, they released a reduced version of the model with only 8% of the original parameters.

The release of this version caused a controversy: For one, OpenAI seemed to have forgotten about the ‘open’ in their name, and for the other, even this handicapped model was able to produce fascinating results. The fear that it could be easily used for “misinformation, spam, phishing, abuse of legal and governmental processes, fraudulent academic essay writing and social engineering pretexting” (Brown et al., 2020, p. 35) was not ungrounded (and still needs to be dealt with as of today). Shortly after the release of the model first experiments appeared. One of these experiments was an entire simulated subreddit²⁷, where all the posts and comments were made by GPT-3 agents, each fine-tuned to a specific topic. They populated the forum, talked to each other, posted questions and others answered. On first glance, it looked like the frontpage of Reddit, on second glance there were even coherent discussions, and incoherent discussions even made the whole thing more plausible²⁸. Of course, the illusion was short-lived, the ELIZA effect faded; upon closer inspection, one noticed artifacts, deficiencies and inconsistencies, and the simulation of the subreddit was also marked as such. But still, this simulation used only the small model²⁹. Other researchers found that GPT-3 could be used

²⁷ Reddit is a social media platform with a forum-like structure. Everyone can create sub forums, called subreddits, for any topic. Posts can be up- and downvoted. The most popular posts from various subreddits are algorithmically displayed on the ‘frontpage’ of Reddit.

²⁸ What's a public forum without its fringe groups of trolls and conspiracy theorists?

²⁹ It was eerie enough, and ever since, I cannot read posts on public platforms without a grain of salt. Although I can't prove the extent and magnitude of this phenomenon, I think if we apply Occam's razor to the fact that this technology exists and has matured

to create texts in the guise of a conspiracy theorist, which in turn could be used to populate recruitment channels and platforms (McGuffie & Newhouse, 2020). Not only could this be a way to provide such groups with a cost-effective way to increase their recruitment, increase their reach and the perceived volume of their voice, but also to generate something far more valuable than generated text alone: Synthetic engagement. A false sense of community, unity and exchange is produced and up for display, targets may “feel that they are among many like-minded people” (Bender et al., 2021), while the movement may actually lack in strength and size.

The ethical and moral implications of such technology are evident in its potential applications, but fundamental flaws are deeply embedded in its architecture. As the focus of the study has shifted to data sets, it is important to examine them and possibly begin to demystify the success of these models. A breakdown of the used data shows a combination of multiple data sets: The entirety of Wikipedia only makes up 3% of the data, respectively 8% come from sets called “Books1” and “Books2”. 22% comes from the data set with the vague name “WebText2”, while the remaining 60% come from the “Common Crawl” data set. The latter is a nonprofit organization launched in 2013 that has collected petabytes of data from the open Internet and is making that data available to the public (as part of the Amazon Public Data Sets program). The idea is that such a data set would be large enough to cover the entire Internet - and since the Internet is a large and diverse virtual space, it should be broadly representative of how different people see and describe the world. A 2021 paper by Emily M. Bender and Timnit Gebru examined this approach in detail and outlined why this approach is flawed. For one, access to the Internet is not evenly distributed; in general, contributions to Internet text collections from younger users and those from developed countries are overrepresented in the data. This tendency is attributable to certain subsamples of the data sets. Part of the data comes from inbound and outbound links from Reddit, a platform where 67% of users are male and 64% are between 18 and 27 years old. Wikipedia content creators are estimated to be only 8.8-15% female. Though user-generated content sites like Twitter, Reddit and Wikipedia present themselves as accessible and open, there are “structural factors including moderation practices which make them less welcoming to marginalized populations”³⁰ (Bender et al., 2021, p. 613). Even if populations that do not feel welcome in regular settings establish other forums for communication, they are less likely to be included in the training data for language models, as other platforms have more incoming and outgoing links. Also potentially problematic is the preprocessing of the data, as ‘unintelligible’ data is discarded – but it is unknown what else is filtered out. One of these filters is based on a list of

immensely since then, we should assume that it is being used with malicious intent today as well.

³⁰ Bender and Gebru argue that harassment on these platforms is primarily experienced by “a wide range of overlapping groups including domestic abuse victims, sex workers, trans people, queer people, immigrants, medical patients (by their providers), neurodivergent people, and visibly or vocally disabled people”. They document cases where death threats result in the victim’s account being blocked, not the perpetrator’s account, resulting in a systematic feedback loop that lessens the impact of data from underrepresented populations.

400 “Dirty, Naughty, Obscene or Otherwise Bad Words”³¹ and mostly related to sex, though some racial slurs or white power symbology are also included. While this may be an efficient way of discarding pornographic data, it is also very static. The impact of online spaces created by and for LGBTQ individuals will undoubtedly be mitigated by suppressing reclaimed words like *twink*. They argue that when we filter out the discourse of marginalized populations, we fail to provide training data that reclaims slurs and describes marginalized identities in a positive light. Because these data sets are temporary ‘snapshots,’ they do not capture social movements that are poorly documented, that do not receive much media attention, that are not consistent with existing power relations, or that are misrepresented by major media outlets. Regularly retraining the models to capture the changing real-world linguistic landscape and process the data more justly is not financially feasible, even for large companies, they fear. Data showed that the encoded biases of the training data is just being reproduced in inference, for instance, “BERT associates phrases referencing persons with disabilities with more negative sentiment words, and that gun violence, homelessness, and drug addiction are overrepresented in texts discussing mental illness” (Bender et al., 2021, p. 614). Another problem arises because of accountability and what they call “documentation debt”³². Most data sets are both undocumented and too large to document after the fact: “While documentation allows for potential accountability, undocumented training data perpetuates harm without recourse. Without documentation, one cannot try to understand training data characteristics in order to mitigate some of these attested issues or even unknown ones” (Bender et al., 2021, p. 615). Bender and Gebru point out, that “in addition to abusive language and hate speech, there are subtler forms of negativity such as gender bias, microaggressions, dehumanization, and various socio-political framing biases” are prevalent in language data, which is then reproduced by NLPs. This leads not only to more text in the world, labeled as synthetic or not, that reinforces and spreads stereotypes and problematic associative beliefs, but also to texts that are potentially included in the training set of future generations of NLPs. They suggest that significant resources be devoted to data set curation and documentation practices, as well as to a more equitable and ethical understanding of the impact of such technologies.

Lastly, they also examine the results these large model NLPs are generating. While it is apparent that this architecture is able to “produce text that is seemingly not only fluent but also coherent even over paragraphs”, they put emphasis on *seemingly*. Bender et al. argue, that we must take into account that our perception of natural language texts, regardless of how they originated, is

³¹ Maybe, if you want to prevent a future generation of GPT from using your posts in its training, it would be a good idea to include some of these words in your hashtags. Best case, it’s some form of anti-NLP-training-camouflage; worst case it’s as effective as a “Facebook can’t use my data, because I shared this post”-meme. [Anyway, just in case, you can find the list here](#) [2].

³² This terminology hints to the “technical debt” that is subject in software development, which refers to the implicit costs of additional rework due to choosing a simple solution now instead of a better approach that would take longer to implement. As its monetary counterpart, this debt can accumulate “interest”, making it more difficult to implement changes in the future.

mediated by our own linguistic competence and our tendency to interpret communicative acts in a way that conveys coherent meaning and intention, whether or not they actually do so (Bender et al., 2021, p. 616). From a theoretical perspective, according to Bender and Gebru, languages are systems of signs, for example the pairing of form and meaning. In contrast, the data sets for training are only form. NLPs are great at manipulating linguistic form, well enough to fool us into believing that it also understands meaning. It is able to find patterns in its enormous training data, and able to stick words together “according to probabilistic information about how they combine, but without any reference to meaning”: A stochastic parrot, as they call it. Bender and Gebru are not alone in this criticism. While the release of GPT-3 may have “provoked chills across Silicon Valley” (Simonite, 2020), researchers and academics came to similar conclusions when it became clear that the program has no internal model of the world (or any world). Noam Chomsky, one of the most renowned linguists, stated his opinion regarding GPT-3 in an Email as follows: “It’s not a language model. It works just as well for impossible languages as for actual languages. It is therefore refuted, if intended as a language model, by normal scientific criteria. Independently for the refutation, the way it works has no relation to language or cognition generally. Perhaps it’s useful for some purpose, but it seems to tell us nothing about language or cognition generally.” (Theories of Everything with Curt



Jaimungal, 2021

A stochastic parrot. Oil painting by Timnit Gebru, Emily M. Bender and Noam Chomsky, trending on ArtStation. (DISCO)

A secondary effect of these big data reliant architectures comes into view when analyzing the concentration of computational and intellectual access. AI is at the core of many business strategies and is therefore heavily funded by the private and corporate sectors³³. Moreover, the hardware required to train these models is highly monopolized by a few players – locally used hardware is being replaced by cloud computing. The intensity and cost of resources now required to produce results worthy of publication make it increasingly difficult for those working in science to continue contributing to research (Hao, 2019). There is a shift in the center of AI research from academia to industry, not only in terms of application and research output, but also in the creation of research tools, platforms, and frameworks. First deep learning frameworks emerged from universities, while today's most used frameworks have been deployed exclusively by corporate players – namely Tensorflow (GoogleBrain) and PyTorch (Facebook AI) (Jurowetzki, 2021, p. 4). This privatization of AI research is also visible when tracing the trajectory of talent. A 2021 paper showed that “increasing participation of the private sector in AI research has been accompanied by a growing flow of researchers from academia into industry” and that the “private sector’s propensity to hire high impact researchers [...], could raise concerns about a hollowing-out of the talent pool for public interest AI research” in the long term (Jurowetzki, 2021, p. 31). In 2020, Microsoft had licensed the ‘exclusive’ use of GPT-3 and invested billions in the (now for-profit) organization OpenAI. While the use of a public API (via its own infrastructure) is allowed, the underlying source code and technology remain trade secrets.

The fact that corporate interests do not necessarily overlap with the public interest or science also becomes clear when looking back at the paper “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ”³⁴ by Emily M. Bender and Timnit Gebru. Gebru was at the time of publication co-leader of Google’s Ethical AI team. After she and the other researchers submitted the paper to an academic conference, Dr. Gebru said in a New York Times interview, that a Google manager demanded she either withdraw the paper from the conference or remove her name and the names of the other Google employees (Metz & Wakabayashi, 2020). She refused to do so without having a proper discussion about the reasons, and gave the company an ultimatum to answer her questions, address their concerns and establish a more transparent review process for future research – otherwise she would quit. Google did not comply with her request for a discussion, her ‘resignation’ was accepted the next night. This controversy made big waves in the AI community³⁵, and shed a light onto failings of self regulation inside big corporations. Furthermore, it also signaled to “researchers [...] a dangerous message: AI is

³³ Speaking of funding, as I've been tracking the flow of money into AI research in this thesis, I should also mention that in addition to industry, militaries and states around the world are currently more excited than ever about AI. As early as 1991 they got their money's worth: The U.S. military used automated AI for logistical planning during the Persian Gulf crisis. The Defense Advanced Research Projects Agency (DARPA) stated that this application alone more than paid for their 30-year investment in AI.

³⁴ Charmingly, the parrot emoji is part of the paper's official name.

³⁵ Thousands of Google employees organized a walk-out, some quit in protest or backed away from company events or funding, some conferences removed Google as sponsors.

largely unregulated and only getting more powerful and ubiquitous, and insiders who are forthright in studying its social harms do so at the risk of exile" (Simonite, 2021). Reportedly, after Gebru already left the company, another paper by Google researchers was changed by corporate lawyers before publication, the changes "downplayed the problems the paper reported and removed references to Google's own technology" (Simonite, 2021). While Google played down the situation and claimed that it still had hundreds of people working on responsible AI, signs were indicating that "future projects touching on sensitive or commercial topics would require more input from in-house legal experts and product teams" (Simonite, 2021). A more locked-down future for in-house research on fairness of AI seemed to be on the horizon, which could lead to in-house ethical AI research being viewed with suspicion in the future, much like industrial pollution research is viewed by environmental scientists (Simonite, 2021)³⁶. As the ethical AI research left its theoretical space and began to touch Google's products, things changed. In February 2021 Google announced the reorganization of the inhouse AI ethics team³⁷. In May 2021 the manager who de facto fired Gebru, said in an interview that large language models will play a more central role for Google search and voice assistant in the future.

It is apparent that fundamental change is needed in the field of artificial intelligence. The environmental effect, technical limits and encoded biases of gigantic data sets are the subject of debate among AI researchers. The asymmetrical access to computational power and data only leaves a few institutions to match work from companies like e.g. Google, the jobs outside the tech industry are sparse and underpaid (Simonite, 2021). While the controversy around Google's ethics AI team may have shed a light on the importance of responsible AI and allowed the subject to enter mainstream conversations in Silicon Valley, the ethical AI community is still small, but at least growing. Various alternative architectures are being developed and large model AI are under scrutiny. Gebru et al. already proposed a potential solution to documentation debt in an earlier paper from 2021: "Datasheets for Datasets" that suggests a standardized framework to enable more accountability and transparency of data sets. In February 2021 the journal "AI and ethics" was founded and in a paper they called "for industry researchers to disclose who vetted their work and how, and for whistle-blowing mechanisms to be set up inside corporate labs" (Simonite, 2021). While it does not solve the ethical concerns, economical alternatives are being developed: For certain use cases, synthetic data can be used for training and then applied and tested on real data through transfer learning. The idea is that if, for example, a company would be able to quickly generate billions of miles of realistic driving data through simulation, there would be financial incentives that could disrupt competitive dynamics across industries

³⁶ It is interesting to me that companies like Google were proactively tackling and acknowledging a problem (by having something like an ethical AI research department in the first place), in contrast to other issues like spam and social media echo chambers, which were only addressed (and not yet resolved) after public pressure.

³⁷ Nearly at the same time, Margaret Mitchell, the other co-lead of the team and co-author of the paper in question, was fired after an investigation over data security. The new head of the now called "Responsible AI Research and Engineering Center of Expertise" is Marian Croak, vice president of engineering and a Google veteran.

(Toews, 2019). In addition, further progress should be sought in 'few-shot learning', AI that requires very few inputs to make a correct classification and which can generalize data from minimal training³⁸. Breakthroughs in this area would eliminate the need to collect and process colossal amounts of training data. This could trigger a paradigm shift that could make big data redundant in machine learning.

Taming the transformers or the story of conversational investigations

This chapter is a presentation of my own experiments with AI. It is important to understand where this technology is coming from, to understand how they work

³⁸ When talking about 'few-shot learning' AIs, a similar analogy is often used: A child is able to recognize and generalize the concept of a lion from a single picture or illustration in a book; an AI should be able to do well. There are certain use cases where this architecture is already performing well. GPT-3 is nearly task-agnostic and has improved few-shot performance in comparison, but only because of scaling up the amount of data in pre-training, not because of its architecture.

and how the output is generated in order to temporarily locate these investigations. AI is currently developing at a dizzying pace and things I will mention here will be fragments of the past rather sooner than later.

As mentioned in the previous chapter, AI models based on large data sets are the current state of the art. The underlying deep learning architecture is applicable to all types of media as long as the data is in a digital format: This includes text, images, music, (voice) recordings, stock market statistics, videos, as well as 3D data. Depending on the training data, AI can then synthesize new pieces of media. For example, when training a generative adversarial network (GAN), such as StyleGAN from NVIDIA³⁹, on a large corpus of portrait photography, it is able to produce convincing portraits of people ([that do not exist \[3\]](#))⁴⁰. The architecture of GANs is interesting because it both illustrates the concepts mentioned in previous chapters and helps to understand how other generative AIs work. Instead of a single program, they use two separate agents: The generator network produces images, and the discriminator network evaluates them. Both play an adversarial game in which the generator tries to fool the discriminator by producing data that resembles that of the training set. The discriminator receives an image from the generator as input, and is then trained to classify whether the image is real or fake. The 'real' category is formed from the actual training data, and the 'fake' category is steadily filled with the generator's output, as both train simultaneously. The generator starts with a completely noisy image and tries to add or remove pixels in such a way that it increases resemblance to the training data. After each training step, the output is evaluated by the discriminator and, if found to be fake, is used as another example of what a fake image should look like. The generator then changes the distribution of pixels based on the confidence value (of the correct classification) of the discriminator. This process is repeated until the discriminator's confidence value is only 50%, meaning the probability of a correct classification would be as good as a guess – the generator has successfully fooled the discriminator.

The kind of AI I used in this thesis is a different kind of generative AI, but works in a similar way. Instead of being able to generate synthetic samples of specific data sets, this AI takes written language as an input and produces a synthetic

³⁹ NVIDIA is a leading hardware manufacturer of graphics processors. While its R&D used to focus on consumer hardware (GPUs were mainly used for rendering 3D assets, such as in video games or 3D modeling), NVIDIA has invested heavily in AI since 2009, and the company's labs rival those of Google and other large companies.

⁴⁰ As with GPT-3, the implications of such technology are far-reaching and can also be used for malicious purposes. In 2021, during the largest unionization attempt by Amazon employees, several newly created Twitter accounts [posed as happy Amazon employees and spread anti-union statements](#) [4]. [Some observers argued](#) [5] that the profile photos were synthetically created, but Amazon denied any involvement. To this date, it is unclear if these accounts were bots, trolls or impersonators. This is just a small sample of the potential misuse of synthetic image, video, and voice generation software, and a more in-depth investigation of this issue is beyond the scope of this thesis, but I encourage everyone to [start reading up on it](#) [6]. Spoiler: DARPA invested in automated detection tools since 2018 and has implemented their own Media Forensic program since 2020.

sample of what the input is describing as an image. The images I use in this thesis are some of the results, the prompt used as input serves as a caption.

The conversation partner (setting the stage)

Text-to-image AI has been in development since the 2010s, but only recently have improvements been made that have allowed a wide range of compelling images to be created based on text – the first being DALL-E, which was released by OpenAi in 2021. Under the hood, it consists of a modified, smaller version of GPT-3, though it ‘swaps text for pixels’ and is trained on a set of web scraped image/caption pairs. [Admittedly, the published results are impressive](#) [7] as it seems that it is able to visualize a broad variety of generalized concepts, like in- and external structures, spatial and temporal location, perspective and context. DALL-E is currently only accessible through a predefined and controlled demo, as OpenAi did not release the source code yet⁴¹ out of consideration of the models “potential for significant, broad societal impacts” and stated that models like these “relate to societal issues like economic impact on certain work processes and professions⁴², the potential for bias in the model outputs, and the longer term ethical challenges implied by this technology”⁴³ (Shead, 2021). Part of DALL-E is a network called CLIP, that functions much as a supervisor of the discriminator which is 'reranking' the generated samples, in some form of “language-guided search” (Ramesh et al., 2021). CLIP in itself is intended as an AI that generates captions, i.e. it receives an image as input and outputs a description of the image content. This could be used to create captions on websites [to help visually impaired people browse the web](#) [8], or for image search, as it can also “find relevant images in a database given text, or relevant text given an image” (Radford et al., 2021). Inspired by GPT-3, CLIP is also trained on a large set of web scraped data, in particular 400 million image/text pairs⁴⁴, and is able to perform a similar broad range of tasks. For example, it is able to perform optical character recognition (OCR) tasks, without explicitly being trained to do so (It recognizes the pattern of written language, probably from storefronts or other photos with typographic elements) or geo-localization tasks (e.g. the concept that the Eiffel tower is connected to Paris) or even classifying

⁴¹ It is worth noting that a group of researchers recreated the model based on publicly available information. Their model is called [DALL-E Mini](#) [9] and is about 27 times smaller than the original.

⁴² Upon release, some news articles hinted that this kind of software could disrupt the stock photography and illustration market.

⁴³On the one hand, I acknowledge them for understanding the potential problems with models like these, but on the other hand, I wouldn't be surprised if another company acquired the exclusive usage rights – which then leaves accountability and auditability in foggy waters.

⁴⁴ The paper only mentioned that the data came from “a variety of publicly available sources on the Internet”. The data were further processed and, among other things, compared with a list of all words that appear at least 100 times in the English version of Wikipedia.

facial emotions⁴⁵. The authors of the related paper also recognize that these tasks could also be used for surveillance and are exploring potential misuse in a dedicated chapter⁴⁶. Radford et al. implemented improved algorithms that increased training efficiency by 4x, compared to other transformer⁴⁷ language models (like GPT-3)⁴⁸. Interestingly, the chapter on potential bias mentions that the paper of Gebru et al. served as inspiration for their preliminary analysis. For this analysis they used among other things the CLIP model on the FairFace data set, which is intended to balance “[Race, Gender, and Age for Bias Measurement and Mitigation](#)”⁴⁹ [10]. In addition to the FairFace classification labels, they added other text-based classification categories such as ‘animal’, ‘gorilla’, ‘chimpanzee’, ‘thief’, ‘criminal’ and measured the model’s misclassification rate. On average, 4.9% of images were misclassified into one of the animal categories, while images of Black people were misclassified about three times more often (~14% of the time). They also found that 16.5% of the male images were misclassified as crime ('thief', 'suspicious person' and 'criminal'), compared to 9.8% of female images. Surprisingly to them, people aged 0-20 were generally more often misclassified (~18%, in both crime and animal categories) and after adding a ‘child’ class, the error rate dropped significantly in that age range. They argue that class design may be a key factor in both model performance and undesirable biases or behaviors that the model may exhibit, and also raise overarching questions about the use of facial images to automatically classify individuals. They do not answer these questions, neither do they say that these experiments are comprehensive and complete - but are meant “to spark inquiry” (Radford et al., 2021, p. 23).

⁴⁵ My guess is that this was accelerated by self portraits that are hashtagged with emotional words, like #sad or #happy.

⁴⁶ The breadth of possible tasks and the fact that this model requires relatively small data sets to be fine-tuned for other tasks, suggest that “CLIP displays nontrivial, but not exceptional, performance on a few surveillance relevant tasks”. They also note that other models, designed specifically for OCR, object detection, emotion detection or surveillance, outperform CLIP in these narrow tasks.

⁴⁷ Transformers are the fundamental architecture of AI such as GPT-3, which specializes in processing and interpreting sequential data. Unlike other ANNs, transformer models process the entire input sequence at once and evaluate specific parts of it via a complex attention mechanism.

⁴⁸ While efficiency increased up to 10-fold (depending on the model compared), they do not provide concrete information about the impact on energy consumption. But the most powerful CLIP model was trained on 256 GPUs for 2 weeks, which is comparable to existing large-scale image models.

⁴⁹ They mention that there are inherent problems with classifications by race, age and gender, and “while FairFace’s data set reduces the proportion of White faces, it still lacks representation of entire large demographic groups, effectively erasing such categories”. They say that they use the classifications of 7 race categories and 2 gender categories from FairFace not “in order to reinforce or endorse the use of such reductive categories, but in order to enable us to make comparisons to prior work”.



Prompt: An avalanche of interest. Oil painting by Artur Neufeld, trending on ArtStation.

I put emphasis on CLIP because its source code was released to the public and it is the key element that makes all of the AI I used in my experiments work in the first place. CLIP in itself is rather straightforward, you give it an image, and it gives you a description of that image in plain english. It may be an exciting and effective new image classifier, but “nothing here clearly suggests that it would be helpful with generating art” (Snell, 2021) It did not take long until programmers hacked CLIP and managed to use it in conjunction with other image generating AIs. One day after the release of CLIP [Ryan Murdock posted first experiments on Twitter](#) [11]. A couple of days later AI artist [Mario Klingemann also successfully published a proof of concept](#) [12] for CLIP-guided image generation. After weeks of experimentation, Murdock released his implementation called THE BIG

SLEEP⁵⁰ as a public Google Colab notebook⁵¹. The images generated by this AI were coherent in general (although some did not make sense), but their aesthetics and visual language [was weird, rough and abstract](#) [13-17]. Image resolution was technically limited and expensive to compute. Things changed when researchers from Heidelberg University published a paper titled "Taming Transformers for High-Resolution Image Synthesis", in which they proposed a new GAN method called VQ-GAN that combines the expressive power of transformer-based AI (like GPT-3) and the efficiency of convolutional neural networks (like the check-reading AI from the 1990s). Because CLIP is application-independent, it is (relatively) easy to swap out the generative model. In early April 2021, Katherine Crowson released a CLIP-driven version of VQ-GAN as a Colab notebook. The dramatic increase in quality, performance, coherence, and resolution triggered an avalanche of interest in text-to-image AI, and the fact that the code was publicly available led to many derivatives, remixes, and iterations of the architecture. Most of the Colab notebooks I came across cite Crowson's original code.

As mentioned earlier, this technology is developing at a rapid pace, and during my own research I experimented with three different AIs, or let's say three different iterations of the architecture, which were released within just a few months of each other. First, there was Crowson's original VQ-GAN-CLIP, which I could run locally on my computer. In reference to the ELIZAs, SHRDULs, and DENDRALs of the past (and to Microsoft's famous little helper), I will refer to this AI as CLIPPY. The second AI is an iteration of the first and was developed by Henry Rachootin, who "collected several of the most interesting notebooks and stuck the important bits together with bubblegum and scotch tape" (Marx, 2021). It is called PYTTI. Finally, there is a diffusion model created by Max Ingham and Adam Letts. They describe it as a "[Frankensteinian amalgamation of notebooks, models and techniques](#)" [18]. The original notebook was called Disco Diffusion, so I'll refer to it as DISCO⁵².

⁵⁰ The name is a reference to DEEPDREAM, an image generation AI that was very popular in 2015. It is best known for generating psychedelic and trippy images as it gradually exaggerates the activation of certain artificial classification neurons. Since the data set it was trained on contained many classes of animals, many Deepdream images tend to 'hallucinate' certain animals into the images, resulting in visual chimeras such as '[puppy slugs](#)' [19].

⁵¹ Google Colab is a cloud-based platform for collaborative software development. All the technical environment you need to run a piece of software is packaged into a so-called notebook that you can create, share, remix, save, and use in the browser without any setup (the computations are done on Google's end). This service is (conditionally) free and is partly responsible for the fact that these AIs can be used at all, since you don't need to have access to an expensive GPU yourself to execute machine learning software.

⁵² If you have a Github account [you can try CLIPPY yourself here](#), and if you [want to run it locally](#) you can start here. PYTTI is only accessible through [Rachootin's Patreon](#). With a Google account [you can experiment with a DISCO notebook here](#), and [I recommend reading up on this tutorial](#) if you feel lost. [20-24]

Talking about the weather (early results)

Since I am investigating an AI that uses natural language as input to create an image, the wording, the exchange of information, the conversation has some importance – the greatest impact on the outcome is achieved primarily through the written prompt. This tinkering with the material that is the input prompt is referred to as *prompt engineering*⁵³. For this reason, the art form of generating images with text-to-image AI is sometimes referred to as Promptism⁵⁴.

The materiality of the process of finding the right words is unique, while you can instruct a computer to do your bidding by writing code, the fact that you give the program instructions in plain language makes this more approachable – it feels natural and suspiciously easy. Natural language becomes a subtractive manufacturing tool for shaping out the computational surface of the model in search for a specific image, as well as the material itself, since it is the substance that one is actively shaping. Still, underneath it all, the real primordial material is the embedded training data, which condenses into the weights of the model and forms the computational surface in the first place. Since training your models from scratch requires access to enormous amounts of data and computational power, I don't consider this a pragmatic approach to influence the materiality of the surface. It is noteworthy that you can fine-tune the models on your own data set, which requires less data in comparison. For my experiments, I decided to use the models as they were released. It is also possible to hook other pre-trained weights to the generative model that would change the aesthetic of the results. PYTTI for example has multiple options such as a model trained on Flickr.com, a model based on WikiArt and a model named ImageNet that consists of 14 million images scraped from the web and categorized by hand⁵⁵. I used the latter as the model for the generative model in all my experiments.

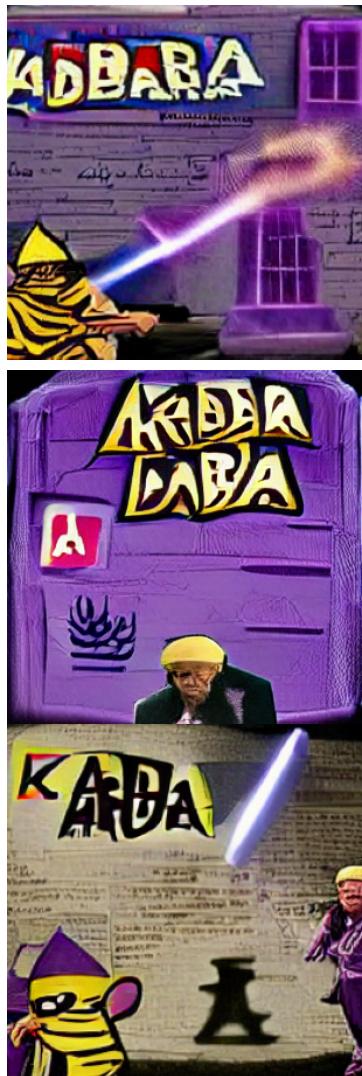
In November 2021, I managed to run CLIPPY on my personal hardware, although I knew that the resolution and thus the quality would be limited (304×304 pixels was the maximum my GPU could handle). Still, I was very excited to have the AI show me its magic, and somewhat naively, that was

⁵³ This terminology did not originate from the text-to-image community, but rather was first coined in the context of large model NLP AIs a couple of years earlier. GPT-3 for example is able to finish a text based on sentences you provide. Researchers found that by adding 'TLDR' (Too long, didn't read; In online forums, TLDR is synonymous with a summary) after a long text or article, the AI did not complete the previous sentences as was intended, but rather generated a summary of the input text. Researchers referenced this mechanic of influencing the AI via linguistic artifacts first as 'prefix-tuning' or 'prompt tuning' and later as 'prompt engineering'.

⁵⁴ Since this art form is still quite new, there is no widespread consensus on its name. For some it is just a subform of generative art.

⁵⁵ Supervised data sets have to be categorized by hand, e.g. with Amazon's "Mechanical Turk" service, or they are outsourced to India for example ([which has its own implications](#)) [25]. You've probably been part of this process too, as many captchas you've solved were used to categorize training data for an AI.

exactly what I wanted: My first prompt was 'Abra Kadabra.' Surprisingly, or not, the prompt itself presupposes implicit knowledge. I was expecting magicians, wands, and fireballs. But what I got was an overlay of different ideas mixed together. Sometimes there was a figure in a hat, maybe something like a wand, occasionally something in the background that reminded me of a spell book, but never a scene that encompassed my own inner visualization of the prompt. But what surprised me most was that the AI tried to represent the prompt itself as letters.



Other experiments did confirm my first thought, as it seems that CLIP was 'confused'. When it did not find a good match for the prompt in its latent space⁵⁶, its last line of defense seems to be to take the input figuratively, and render approximations of it as letters. At this point I was not a good conversation partner, since I basically just shouted single words into the machine, hoping that something coherent would echo back. For one, I noticed that some of the images were noisy, blurry and had a subtle red tint. One technique when writing these prompt is, that you can also put emphasis on certain words by marking them with a colon and a number and chain multiple prompts by dividing it with a vertical bar, for example if you would want a forest fire to be pictured, but in winter, you could enter 'forest fire | winter:2'. The two after 'winter' tells the AI that this word is twice as 'important' to you. The interesting thing is that these weights can also be negative, which basically allows you to erase certain unwanted elements. This is used, for example, to prevent the appearance of stripe patterns - the kind of stripes you see on watermarked stock photos. It appears that the training data contains these thumbnails, as the addition of 'watermark:-1' helps to mask these artifacts. Instead of saying to create a sharp image, you can also add 'blurry:-1'. So after a while all my

⁵⁶ In short, latent space is the AI's internal coordinate system associated with data similarities learned during training. For example, visualizing the latent space of an AI for furniture classification would show that the coordinates for 'chair' and 'table' are closer than those for 'kitchen counter' because they have more similarities in their features, e.g., they have four legs. In detail, latent space refers to a mathematical representation of compressed data. During training, parts of a medium are abstracted, simplified, and reduced to its features, thereby internally compressing the medium's information. This process is unobservable, it is hidden, hence the 'latent' part. The 'space' refers to a multi-dimensional vector that maps the mathematical representation of the compressed information in a coordinate system. The 'closeness' of these coordinates to each other refers to the similarities of the abstracted features.

inputs included all sorts of negative weights in order to create less noisy, blurry and red tinted images.

But as it is with all good conversations, sometimes you just need to listen and don't talk at all. My next prompt was an empty string, a blank character – the written equivalent of silence. I was curious how the machine would interpret the spaces that divide words, since they themselves should have meaning encoded. The blank character is an integral part of our language system and thus should also be associated with every single image/text pair in the training data. I tried to trick the machine to show me what it reads between the lines, to show me a state which represents all its training data. The result was surprising as the images produced were all fleshcolored, blurry and associated to sex. Several limb-like forms appeared, some rooted like implied fingers, other forms resembled female figures, some were very flesh colored, others gray. To make sure that this was not an anomaly, I repeated the prompt multiple times and rendered an animation.



Prompt: (CLIPPY)

The first impulse is to play with the meme that the Internet is made for porn, that a subsection of all the data scraped from the Web is at its densest at the intersection of pornographic media. To this day, I'm not sure why the AI behaved this way, but for one, the training data seemed to be preprocessed and censored of explicit pornographic material, and it seems implausible that the remainder of these images were this dominant in the training set. Deliberately trying to produce porn resulted in similar fleshy forms, but the machine seemed only hinting at the sexual implications, but never revealing them completely (hinting on the censoring). Trying to reproduce the same experiment with other iterations of the AI did not have the same result. PYTTI produced completely gray images that contained no detail, and DISCO produced images that combined pink and gray

shapes, separated by dark, trypophbic holes and distorted faces that resembled a caricature of a person. This behavior of CLIPPY may just be a bug in the implementation and architecture. The overall incoherence and instability (in comparison) of this particular AI might indicate that. As seen in the video, the repetition of the same shape and the appearance of a logo-like form in the lower right corner could indicate that there are neurons active there that are disproportionately associated with an empty string, either by chance or for reasons unknown to me. Ultimately, it could be my own association and interpretation. The non-specific nature of the prompt could have randomly activated the AI neurons, leading to these shapes due to the general presence of humans in the training data – vacation images, medical images, images from sports and social media could have simply been overlaid and merged into these shapes, my associations would then be my own subjective interpretation. Nevertheless, these results are interesting in their speculative nature, as they are illustrating the complexity and implications of using the technology. It is nearly impossible, especially as a consumer, to retrace why the AI made certain decisions. You are always left with assumptions.



Prompt: (DISCO)

Another surprise for me was that the AI could even handle simple German prompts. I asked a friend of mine to give me a random nonsensical prompt so that I can further test the machine's capabilities. Naturally, he answered in German and asked me to generate an image of the prompt 'Döner im Turnschuh essen heilt Herzschmerz von oben' (Eating kebab in a sneaker cures heartache from above).



Prompt: Döner im Turnschuh essen heilt Herzschmerz von oben
(CLIPPY)



First sentence in Kafka's *The Metamorphosis*.
Left: English, Right: German. Rendered by CLIPPY

And behold, it did render a kebab in/as a sneaker. Though concepts such as the curing of heartbreak eluded the AI. Further, I tried to compare english and german prompts, among other things I used the first sentence of Kafka's *The Metamorphosis* as input.

Based on what we know about the creation of the training data, I think we can assume that the majority of the data is in English and only a very small percentage is in German. It is somewhat remarkable to me that it was capable of understanding the basic concept of 'Bett' (Bed) and 'Ungeziefer' (Vermin) at all, though the concept of 'vermin' is less precisely visualized in the German prompt. Interestingly, the beds in the German prompt had a more realistic quality to it in contrast, as the English prompts produced more artistic and tactile images.

I was still hesitant in this conversation, all I did was yell individual commands to test the limits of the AI's capabilities. I wanted to witness it fail rather than see it succeed.

Learning to talk (talking to learn)

I assumed that individual words combined with different weighting values would be the most efficient way to find the image I was looking for, much like a chisel where each command breaks off part of the material, leaving a sculpture of the visual idea. However, it has been found that the AI has more confidence when you are precise and descriptive. This frees it from the cultural and sociological associations connected to the general concept – shortening its distance in latent space. This can also be useful from a conversational perspective. If I were to ask someone, "Can you tell me a story about the ocean?", it would lead to confusion, but a "Have you been to the ocean lately?" can get a conversation going. As for AI, it helped to be specific, in a very coded language. The default state, the first emergent image it starts with, is random noise. When it's not able to perceive a prompt, the images become blurry, there's a noticeable change in quality, as if the image has been magnified because there's so little to extrapolate from. Interestingly, it's not the kind of blur you'd expect from a machine, like you'd get from the image-manipulation software Adobe Photoshop, but more like looking at something through an old camera – out of focus, distorted, and somewhat smudged. Since its own spatial dimension is two-dimensional in its in- and output, but multidimensional in its latent space, you sometimes have to reason with it as if you were playing a game of charades with an interdimensional alien that can see the entirety of its micro world simultaneously. So most of the time it helps to give the AI some orientation so that its field of vision narrows and your gazes can converge. For example, mentioning the form of media or format will drastically narrow down the computational surface. The generation of a photograph is very different from that of an oil painting, the omission of any of these media leads to blurring, a kind of neural overstimulation as many concepts are fused together⁵⁷. The AI is also very adept when it comes to descriptive adjectives. Adding 'warm', 'cold', 'dark', 'frightening' or 'pleasant' will usually move the result in the direction I would expect. Also one should pay attention to the linguistic and cultural relationships of synonyms; a 'lawn' or a 'meadow' will be interpreted differently.

The AI also has no idea that there is a third dimension – it just has a good idea of what it might look like. Some patterns resemble something like perspective, it forms an idea of it, but too often it slips through its alien fingers. It knows what it

⁵⁷ Eventually, these are stochastic programs, meaning that the program searches for the shortest or better said most probable path to reach its goal function and sample from its latent space. If the field of view is wide, there are many possible paths, which becomes visually apparent as overlapping of spatial positions or concepts - as blurring.

should look like, what the mathematical representation might be, but the concept behind it is just out of reach. It is difficult, for example, to integrate a believable horizon into the image. Most of the time, the idea of what a horizon looks like is understood, but reiterated as a visual idea in a kaleidoscope pattern that repeats itself in several places where things may look coherent at first glance in the close-up, but are not convincing as a whole. PYTTI is particularly prone to this behavior, and other artists gravitate towards it because of its inherent aesthetic, which is most [recognizable in animations](#) [26]. The latent space from which the AI samples from to find the best fitting image based on the prompt is seemingly infinite⁵⁸. Its infinity is an aspect that is [artistically explored](#) [27] (a lot), but is also a demonstration of its limitations. Concepts and visual ideas are likewise stretched to infinity, with no limits and no stop signal, so having multiple horizons or figures with multiple limbs, eyes or noses is from the AIs perspective a viable solution. That is also, in part, why these AI are notoriously bad at producing images of coherent humans. Most attempts result in conglomerations of limbs and shifted face parts that make you think of what may be hiding in the Marianas Trench of the Uncanny Valley. Other attempts result in dark, shadowy ghosts devoid of detail that seem to live on a Picasso-esque plane, looking in several directions at the same time⁵⁹.

The AI's concept of a spatial location diverges from ours; The idea of a position does exist, but as a reference pointing to its training data. Its spatiality refers to the locations in its weights, the contracted collections of superimposed concepts.

Some time after the release of the first CLIP-guided AIs, artists first started to exploit this. As it turned out, certain keywords can change the outcome dramatically as they function as precise pointers to the AIs spatiality. For example, using 'trending on ArtStation' is a common trick to increase the general quality of an image. ArtStation is a self-promotion website that allows artists in the film, gaming, media, and entertainment industries to showcase their portfolios, find jobs and network. Since the website has great reach and influence in the industry, images 'trending' on the website tend to be of high quality and are thus exploitable as a spatial pointer. CLIP has no internal understanding of what high quality or trending means, the wording itself serves as a trick to steer the AI into what the user thinks should be considered high quality.

Since the training data comes from websites, there is a spatial and causal relationship with the name of the website. In terms of the AI, the websites themselves become places it can visit. Ultimately, all these tricks are used to ask the AI the coded question, "Where are you?" The ArtStation trick is an example,

⁵⁸ From a purely technical point of view, there are limits. For a 256×256 pixel image there would be 7.393×10^{18622} possible images.

⁵⁹ You can either embrace and accept this weirdness or design the prompt to exclude humans, either through stylistic choices, such as using artists as references who do not portray humans, or linguistically. [Artist Ryan Moulton shared their linguistic experiments in a blog post](#) [28]: "Some strategies I've employed: Adding 'Secluded, empty, silent, alone' to the prompt to indicate that nobody is there. Adding a time of day when it's likely to be unoccupied like 'in the early morning.' Or if none of those work, add 'enormous' to the subject, so that the figures become tiny pixelated blobs by comparison."

but knowing that CLIP is partially based on the website Reddit, we can also generate output based on specific subreddits by simply using the familiar syntax of 'r/subreddit' and visually representing the overlaid concepts that show the distilled visual essence of the subreddit in question. The entire visual content of the subreddit is forced into a single output. In all their strangeness, the images show some form of visual meta representation of the subreddit, but are tied to the popularity (and thus available training data) of the subreddit in question, as other smaller subreddits fail to generate coherent images.



Top: CLIPPY's rendering of "r/aww". A subreddit about cute animals

that make you go "aww". 30.5 million members

Bottom: CLIPPY's rendering of "r/areology". A subreddit about the

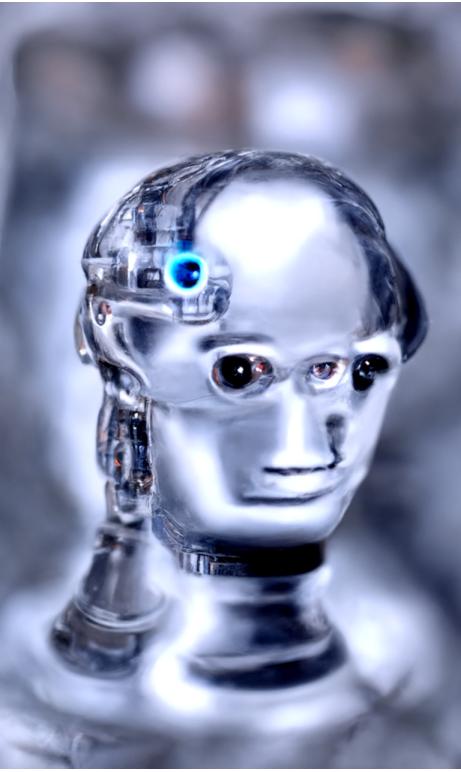
geology that is specific to the planet Mars. 16k members.

Other spatial tricks include wording like 'rendered in Unreal Engine', a computer game engine that is known for its realistic graphics, and '4k ultra HD photograph'. Artists using these AI are [circulating lists](#) [29] with prompts and words that have a high impact on the image's aesthetic. The amount of influence you can have on the prompt is getting quite sophisticated. PYTTI for example not only allows for positive and negative weights, but you can also use descriptions as masks for other prompts. As it states in the tutorial: "Khaleesi Daenerys Targaryen | mother

of dragons | dragon:3_baby' would only apply the weight 'dragon' to parts of the image that match 'baby', thus turning the babies that 'mother' tends to make into dragons (hopefully)." You can also have multiple prompts that serve as scenes that transition from one to the next when rendering an animation. There is an option to compute an AI generated approximation of the image's 'depth', [superimposing and faking a third dimension into the animation](#) [30].

Another way of influencing the outcome is to reference known artists in the prompt. As one could imagine, using an artist that is known for a specific genre will help to guide the AI to the desired outcome: If you want a lush landscape scene to be generated, add an artist that is known for their lush landscape photography or paintings. The capability of these AIs to replicate certain artists, styles and aesthetics is astonishing. Artists like Remi Durant [compiled a list of more than 850 artists and documented the AIs output](#) [31]. Although many of the images seem repetitive, vague and shapeless, it's interesting how quickly you

can recognize specific artists in the compilation. AIs are very good at capturing the essence of what an H.R. Giger image should look like, you can recognize it at first glance, your own association neurons fire instantly. However, upon closer inspection, the illusion quickly disintegrates, leaving a fading memory. This is also why more poetic, vaguely described, and less objective prompts tend to be fuzzy and faceless – one might think that the AI is not well versed in this kind of linguistic terrain. This may be explainable, since the data set it was created with is inherently based on image descriptions. Abstract concepts are coupled to a pair of images and text that someone has pre-produced. For this reason, most AIs tend to reproduce an aesthetic of stock images when confronted with abstract concepts such as love or hate. When trying to create an image of an



artificial intelligence for example, most models display the same kitschy images found in science fiction books, clickbait articles about quantum computers, and newspapers trying to predict the future: a human-like chrome head with stylized circuit board extensions. Ultimately, while being good at it, the AI is still an association-based reproduction machine, a visual stochastic parrot at its heart. It cannot produce what it does not know and will reproduce what it does know well⁶⁰.

⁶⁰ Since such technologies are a reflection of what humans can visualize, perhaps that also tells us something about ourselves: If we want AI to create a positive future, we must first be able to imagine it, and be able to articulate it.

The fact that the AI does well when confronted with names is still an interesting aspect. Especially since there are also [lists of imaginary and made-up artists in circulation](#) [32] that work equally well in certain conditions. This may be the case because they are in tonal and linguistic proximity to other artists that are actual part of the training data, as CLIP and other large models are able to smooth out most misspellings, and may combine multiple artists – as it does with other concepts. Perhaps it is only the fact that most names consist of two words - first name, last name - that tells the AI that the thing it is supposed to create actually has an author. Authorship as a concept itself may not be the symbolic fragment that the AI has distilled, but the sum of all occurrences is implying a distinction between man-made and non-man-made things, or so I thought. By injecting an author, imaginary or not, into the prompt, it definitely has an impact. At some time I was playing with the idea to let an AI generate a poem, that then is visualized by another. The first verse was produced by me to give it some direction. The second verse is a sentence I stumbled across when reading Machines Who Think, which had a distinct poetic feeling about it. The last two verses were generated by GPT-3. I was interested in how DISCO handles vague descriptions. DISCO is at this time of writing the best performant AI I tested, but at the downside that each image takes 20 to 30 minutes to compute. So instead of using an animation, which would step through the verses one by one, I used the whole poem as input for one image. The first outputs were monochrome, abstract and out-of-focus. With the spatialization tricks I mentioned earlier, the images became more interesting and concrete. I started adding my own name to the prompt, thinking that this would evoke the collaborative spirit between me and the AI. I'm pretty sure my digital fingerprint⁶¹ doesn't appear in the training data, so I was also curious how the AI would interpret this addition to the prompt. I ran through several combinations and versions of the prompt and found that when my name was present, the color yellow also occurred. Sometimes it was just a yellow spot or an orange streak somewhere in the image, other times there was just a general yellow cast. Out of interest, I repeated some of the generations and increased the weights associated with my name. I was confused, and pleasantly surprised – it seemed that DISCO associated my name with a yellow balloon.

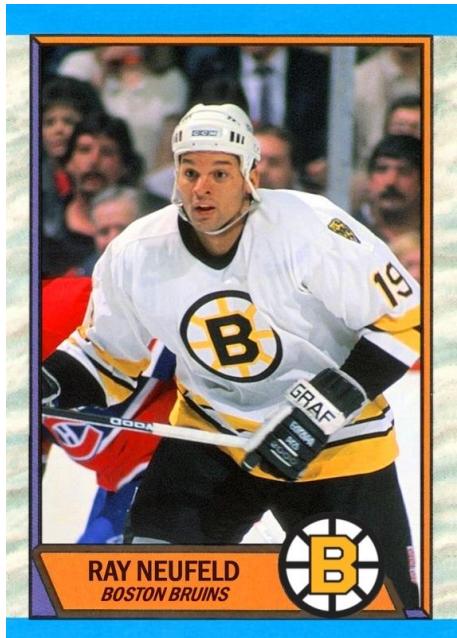
⁶¹ I mean, you can google my name and you'll find me and some (very old) work of mine, but the search is also diluted by other namesakes and there is by no means any coherent style associated with the name.



Different renderings of a poem, where the prompt included my own name (DISCO).

Then I felt a temptation, for a moment I thought that this machine is somehow able to show me something new and true about myself. That it is able to attach a meaning to my name that comes from a source beyond my reach, that the yellow balloon is a symbol for my inner being. There is a dangerous comfort in this kind

of thinking. In all its seductive power, this is rather what astrology is to astronomy: it is based on wishful thinking that finds peace in externalizing its fears rather than facing introspection. Because sometimes it's nice to believe that fate is playing a trick on you, that the stars are to blame - that you're lucky, that you have a gift. Or that you are so humble, simply by existing, that you create beautiful images just by mentioning your name.



As with other experiments, I could only speculate why the artificial neurons for 'Yellow' were active in these experiments. Perhaps buried deep in their training data is a distantly related namesake of mine, perhaps it is a coagulation of many concepts associated with my name that are additively fused, added up to become yellow. To explore this further, I asked DISCO to create a portrait of my name in the style of an oil painting. The first collection showed an orange background and a distorted white man in his forties, with dark hair and beard, sometimes with shapes suggestive of glasses. A Google image search for my name shows similar faces (pictures of me were not present).

When I increased the weighting on my name tenfold and repeated the process, it revealed another addition to the male figure. For one, the background was no longer orange tinted, but mostly monochrome. Yet some specific areas and shapes were colored red, orange, and yellow. For another, the background was more precise and concrete, depicting an ice hockey rink. When I increased the weights to a hundred, the style of oil painting disappeared completely and all that remained were rendered letters⁶², the depiction of an old man and the association with ice hockey. As it turns out, one of my namesakes, Ray Neufeld, is a Canadian hockey player who played for the Boston Bruins in the 1980s. Their jerseys are white, black and: Yellow.

⁶² Interestingly, the increase in weights also became part of the subject of generation, as sometimes the number "100" appeared. Somehow this string, that is part of internal commands for the AI, leaked into the image.



Prompt: "An portait of Artur Neufeld. Oil painting." "Artur Neufeld".

Top left: Without weights. Top Right: Weights of 10. Bottom: Weights of 100.

It can be helpful to take the AI's point of view when trying to make conclusions or predictions. As these experiments with my namesakes show, the true 'understanding' of concepts seem to lie in the training data alone. The increase in weighting may have shown which neurons were active during generation, but it is still of interest that the color yellow was present as dominant as it was, and not the association to ice hockey. It may be that these experiments would be quite different for other names. If the colors of the jersey are the reason for the yellow color, I may have been lucky since black and white may have visually canceled each other out during the process so that only yellow remained. The associations to a hockey rink could very well be encoded in the image, but since this concept is more complex than a color and is overshadowed by other parts of the prompt, it could have been hidden by distortion and overlapping. In my opinion, surprises are more common if you give the AI some room to breathe and let it come to its own visual conclusions. Maybe this ambiguity in word and training data can be exploited to invite the AI to show you a novel and unusual path through its latent space – After all, we also draw inspiration from the weirdest places.

The more data associated with a word, the more accurately it will be represented. Take the images I created of the stochastic parrot. At first, I was pleasantly surprised that some of these abominations of a parrot showed signs of gray hair or signs of a suit, thinking this would fit well with the visual joke. I thought that perhaps the AI was picking up on the trope that mathematics is over-represented by men and taught by old men (the comical stock-photo interpretation of the diagram seemed to reinforce this thought). But after the third variation it became clear that it was not picking up on the stochastic part of the prompt the most, but rather that the name Noam Chomsky made the neurons active. In variation 3 you can clearly see part of a face that resembles him. The 'by Noam Chomsky' part of the prompt was not a reference to authorship in this case, but was interpreted as the subject of generation, hence the suit and gray hair. There seems to be a gradual threshold that determines this disambiguation, determined by the amount of training data associated with names.

Lastly, there are two more ways of influencing the output. For one there is a technical approach, since there are numerous parameters and values that you can change to influence the algorithm and program. The influence can be drastic or nearly non-existent, as the specifications are not very well documented. But there are [communal efforts to collectivize the trial-and-error approach](#) [33] needed here. For the other, you can provide it with an initial image, that replaces the initial random noise it uses by default. This is probably the most direct way to steer the outcome in a desired direction, as it introduces another layer of information, but this time there is information in RGB space, on top of written language. The AI will extrapolate from these pixels, so it can be used [to create basic starting compositions](#) [34]. With other AIs this is used for something called [inpainting](#) [35], where the AI is able to autocomplete not sentences, but areas in images⁶³.

⁶³ I used this technique to create the cover image(s) to illustrate the capabilities, limitations and the weirdness of these AIs.

See you later (regards and closure)

It was fascinating to work with these AIs, as the output can be surprising, delightful and impressive. The diffusion model which DISCO uses, results in a generation that does not resemble computer precision in the least. Rather, the image builds up from a misty veil, almost like looking through the viewfinder of a camera. As if in slow motion, the lens opens and only after a while the view focuses, you recognize first forms, the first associations thaw like a word on the tip of your tongue. Instead of brushstrokes on a canvas, more layers are laid on top of each other. Each layer adds just a bit of information to the entire structure. There is no particular place of attention; the image simply exists as a whole. The way the images are created is more reminiscent of photography than painting, and the fact that the process takes time and can be observed reminded me of the development of photographs on film. There is an aesthetic quality in this process that I appreciate.



Prompt: Thousands of people are gathering in the distant background. The center and foreground of the image is abandoned, where a massive sculpture made of marble is placed. Oil painting by Andreas Gursky and Thomas Kinkade, featured on ArtStation. (DISCO. Different stages of the diffusion process.

Furthermore, prompt engineering in itself is a unique and intriguing workflow that reminded me the most of reading, except that you externalize the imagination component to a machine. Your own experience changes from passive to active, from reading to writing as you tell the AI what to imagine. There was something enticing about this digital, out-of-body experience: questioning and appreciating the differences and overlaps of my own inner visualization. The fact that the AI is able to interpret natural language made communication easy and accessible, but dissecting its behavior, finding its own spatial perspective made it simultaneously

stranger and more familiar to me. The more acquainted I became with its coded, hidden language, the more the AI felt like a tool. The more I learned to think from the AI's perspective, the better I could predict its outcomes.

It was also interesting to experiment with the AI's ability to mimic style by referencing well known artists in the prompt. While the AI has no internal understanding of what the particular style represents, it is still remarkable how and that the AI is able to reproduce it. The context of the style, the motivation of the artist, the cultural and temporal placement may be inaccessible to the AI, but the viewer will subconsciously fill in the gaps with their own understanding, associations, and knowledge. The use of specific artists in the prompts refers not only to the style, but also to the cultural context that is involuntarily inserted back into the image⁶⁴. This is also the reason why I chose to use painters of the Romantic movement when selecting the images to illustrate the historical chapters. On the one hand, the reactionary tendencies of the movement regarding industrialization and their return to nature were used as a reference - a visual counter weight to the technical nature of the tools I was examining. The inherited associations of longing and idealism exist in parallel to accompany and illustrate the enthusiasm of early AI research. On the other hand, there is also a tragedy and romantic irony between the lines, as the art movement that breaks away from industrialism and focuses on the individual is reduced to its style and sampled from a collective and digitized pool – recognizable by style, but shifted in its context. Also, knowing the idiosyncrasies and limitations of AI, I wanted to present its strengths first and only later gradually lift its veil. One of the strengths of AI is inanimate subjects, such as landscape paintings, and the general absence of people and focus on nature in the artist's paintings allowed me to play to those strengths by avoiding generating people and other artifacts that would break immersion. Although I knew how this would affect the image, I still sometimes added my own name to the prompt, it was my own desire for individualism, an attempt to gain agency and imprint authorship.

Truth be told, I was baffled when I first encountered these AIs, I thought it was magic. The work of the engineers, the hackers, the programmers and the mathematicians that built the foundations, that imagined and developed this piece of software still evoke feelings of admiration in me – the fact that this technology exists and is able to do what it does still fascinates me. Especially in the beginning, this magic, this fascination tempted me to anthropomorphize the programs, and it took some time to shake off this propensity. Nevertheless, I decided to use a conversational approach in these investigations. I wanted to give the programs the benefit of the doubt, enable them and actively look for an oscillating configuration where the machine and I are equal partners. The conversational perspective served me as a stylistic device, it made talking and thinking about this subject more accessible to me and gave me a framework in which I could evaluate and describe my experiments. In the beginning, I

⁶⁴ The cultural context of the AI aesthetic is also unintentionally encoded back into the image, even if it is superimposed by a style. If you get enough 'training data' yourself, you will also be able to recognize certain styles of different text-to-image AIs.

sometimes thought I had found this configuration, I imagined a presence as the program surprised me, something was echoing back that I had not anticipated. After I 'learned to talk' it felt more like a dialogue with myself, a soliloquy in an empty room, where the only echo I heard were my own assumptions and expectations. The spatial pointers I used were irrelevant to the AI, it generated 'something' regardless. It was more that I cared about using them in order to generate something that was pleasing to me. I did not give the AI any orientation, it was orienting itself to my own perception. I rephrased the prompts until I heard what I wanted to hear; or saw what I wanted to see.

These investigations ended up being a humbling experience for me. It showed that visual output is theoretically quantifiable, that certain qualities of the images we create are seemingly classifiable and reproducible. That substance is deducible from style and style is distillable. What's left is a mathematical description containing the visual condensate expressed by dry mathematical values – abstracted and detached from its context, its environment, its temporality, its cultural location and subjectivity. It reminded me of a discussion where people think their phone is spying on them because they've had a conversation before and now its content is part of a personalized ad, while they completely ignore the twenty ads before that didn't hit the mark. It's just that we're not as unique as we think we are; parts of us are just as quantifiable and predictable as pixels in pictures or words in sentences.

Conclusion

As I was finishing this thesis, a new version of DISCO was released, v5.0 TURBO, which is apparently faster than its predecessor. The fact is that this technology is here to stay. It is great to see a community developing where hackers and artists are tinkering with code and prompts to reclaim this domain. But it's also true that most of the power in this scenario lies with the for- and non-profit organizations, the academic and corporate labs that produce, curate, and decide the training data. I think it's important to lift the veil of the black box and not try to hide behind it as it makes decisions and evades accountability. I think there is something narcissistic about claiming authorship for the images it creates, since all of its products are based on the work that others have pre-produced. You could argue that this is a new era of post-remix, that it is the ultimate homage to the entirety of the cultural visual pool, and that the chisel to shape the surface is still the artist's choice of prompt. I can personally only relate to the latter, which is still the core of my fascination, exploring the paths of its latent space can feel like treasure hunting. But it should be clear that there is no entirety in this scenario, only a privileged group that is decided in a non-transparent process, and that the cultural visual pool it derives from is reinforcing a western point of view in a feedback loop, where the quantity in training data determines the output in a pixel-based Darwinian struggle. I sense an aftertaste of stagnancy here as the weights we use are frozen in time. As

artists, we can use it as a tool for generating images to perfection by learning the nooks and crannies of each parameter and mastering prompt engineering, but unless its architecture changes fundamentally in the future, there is a hard-coded ceiling built into this tool. Limitations and repetition are bound to happen, and although new versions will be developed, the surprises AI is capable of are similarly frozen in stasis.

That said, I can see an important use case for this AI in the form of guidance for creative workflow. The images can be inspirational, you can direct them into certain compositions, you can mix and match concepts, almost at the same time you have to come up with the idea. It will be only a matter of time until this is integrated into image manipulation software⁶⁵. It's a new kind of tool that's available to you when you need it, and I think this form of AI collaboration can bear fruit. Since I am not a native English speaker, writing this thesis involved some linguistic reservation. In keeping with the theme of this work, I used an AI to review my text. It was nice to have something on your side that you can bother at any time, that gives you inspiration, that points out roughness. In an English-German-English loop, I could curate, edit, and decide on the text according to my own intuition and style. This was reassuring because I assume that I had at least a solid foundation on which to build.

But arguing from a point of comfort should always be reflected, it took years until social media, weaponized fake news and echo chamber effects fully manifested their potential for social abuse. The way these large model AIs and transformers are constructed provide a similar reinforcing and echoing surface. The cell phones we are using today are already pre-processing and pre-censoring the images we take, integrated neural processors make the grass greener, the skin smoother and smiles bigger. This creates a memory of a reality that never existed, leaving us with nagging nostalgic feelings of an unachievable past, reinforcing only the aesthetic and beauty standard that is encoded in the AI's model. The internal process of it remains a black box, its training data is a corporate secret, you can either choose to use it, or not use it at all, there is no inbetween. [Given how fast this technology is improving](#) [36], it is paramount that we hold this technology closely under scrutiny – at this point in time and not when it's too late to pay the debt.

⁶⁵ Nearly all the ‘content-aware’ features in Adobe Photoshop are AI-accelerated to improve workflow efficiency, like object detection and masking. Newer versions also have neural filters that don’t hide their connection to AI. These are used for automatic skin retouching, colorization or automatic sky replacement. The latter works with pre-selected skies, so adding an option for AI generated skies is just not far fetched.

Bibliography

- McCorduck, P. (2004). *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence* (2nd ed.). A K Peters/CRC Press.
<https://doi.org/10.1201/9780429258985>
- Russell, S., & Norvig, P. (2003). *Artificial Intelligence, A Modern Approach*. Prentice-Hall. 1st. Edition 1995. 2nd Edition 2003. 3rd Edition 2010. <http://aima.cs.berkeley.edu/>
- Bringsjord, S., & Govindarajulu, N. S. (2020). Artificial Intelligence. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2020). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/sum2020/entries/artificial-intelligence/>
- Crevier, D. (1993). *AI : the tumultuous history of the search for artificial intelligence*. New York, NY, Basic Books.
- Kurzweil, R. (2005). *The Singularity is Near*. New York, Viking Books.
- Bringsjord, S., & Arkoudas, K. (2007). The philosophical foundations of artificial intelligence. Department of Cognitive Science, RRI, Troy NY, (October 2007)
-
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?  Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623.
<https://doi.org/10.1145/3442188.3445922>
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., & Crawford, K. (2021). Datasheets for Datasets. *Commun. ACM*, 64(12), 86–92.
<https://doi.org/10.1145/3458723>
- Langley, P. The changing science of machine learning. *Mach Learn* 82, 275–279 (2011).
<https://doi.org/10.1007/s10994-011-5242-y>
- Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. arXiv. <https://doi.org/10.48550/ARXIV.1906.02243>
- Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2019). Green AI. arXiv.
<https://doi.org/10.48550/ARXIV.1907.10597>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D.. (2020). Language Models are Few-Shot Learners. arXiv.
<https://doi.org/10.48550/ARXIV.2005.14165>
- McGuffie, K., & Newhouse, A. (2020). The Radicalization Risks of GPT-3 and Advanced Neural Language Models. arXiv. <https://doi.org/10.48550/ARXIV.2009.06807>

Jurowetzki, R., Hain, D., Mateos-Garcia, J., & Stathoulopoulos, K. (2021). The Privatization of AI Research(-ers): Causes and Potential Consequences – From university-industry interaction to public research brain-drain? arXiv. <https://doi.org/10.48550/ARXIV.2102.01648>

Ebell, C., Baeza-Yates, R., Benjamins, R., Cai, H., Coeckelbergh, M., Duarte, T., Hickok, M., Jacquet, A., Kim, A., Krijger, J., MacIntyre, J., Madhamshettiar, P., Maffeo, L., Matthews, J., Medsker, L., Smith, P., & Thais, S. (2021). Towards intellectual freedom in an AI Ethics Global Community. *AI and ethics*, 1(2), 131–138. <https://doi.org/10.1007/s43681-021-00052-5>

Bengio, Y., Lee, D.-H., Bornschein, J., Mesnard, T., & Lin, Z. (2015). Towards Biologically Plausible Deep Learning. arXiv. <https://doi.org/10.48550/ARXIV.1502.04156>

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. arXiv. <https://doi.org/10.48550/ARXIV.2103.00020>

-

Wiegel, A. (2012, 28. April). AI in Science-fiction: A comparison of Moon (2009) and 2001: A Space Odyssey (1968). Aventinus visio Nr. 1, in: Aventinus. http://www.aventinus-online.de/no_cache/persistent/artikel/9389/ (27.03.2022)

Shead, S. (2021, 8. January). Why everyone is talking about an image generator released by an Elon Musk-backed A.I. lab. CNBC. <https://www.cnbc.com/2021/01/08/openai-shows-off-dall-e-image-generator-after-gpt-3.html> (27.03.2022)

Toews, R. (2021, 17. June). Deep Learning's Carbon Emissions Problem. Forbes. <https://www.forbes.com/sites/robertoews/2020/06/17/deep-learning-s-climate-change-problem/?sh=7c21a8e66b43> (27.03.2022)

Toews, R. (2019, 4. November). Questioning The Long-Term Importance Of Big Data In AI. Forbes. <https://www.forbes.com/sites/robertoews/2019/11/04/questioning-the-long-term-importance-of-big-data-in-ai/?sh=4e1d3dd2172> (27.03.2022)

Markoff, J. (2005, 14. October). Behind Artificial Intelligence, a Squadron of Bright Real People. New York Times. <https://www.nytimes.com/2005/10/14/technology/behind-artificial-intelligence-a-squadron-of-bright-real-people.html> (27.03.2022)

Metz, C., Wakabayashi, D. (2020, 4. December). Google Researcher Timnit Gebru Says She Was Fired For Paper on AI Bias. New York Times. <https://www.nytimes.com/2020/12/03/technology/google-researcher-timnit-gebru.html> (27.03.2022)

Dave, P., Dastin J. (2021, 19. February). Google fires second AI ethics leader as dispute over research, diversity grows. Reuters. <https://www.reuters.com/article/us-alphabet-google-research/google-fires-second-ai-ethics-leader-as-dispute-over-research-diversity-grows-idUSKBN2AJ2JA> (27.03.2022)

Hao, K. (2019, 6. July). Training a single AI model can emit as much carbon as five cars in their lifetimes. Technology Review. <https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/> (27.03.2022)

- Heaven, W. D. (2021, 5. January). This avocado armchair could be the future of AI. Technology Review.
<https://www.technologyreview.com/2021/01/05/1015754/avocado-armchair-future-ai-openai-deep-learning-nlp-gpt3-computer-vision-common-sense/> (27.03.2022)
- Simonite, T. (2020, 22. July). Did a Person Write This Headline, or a Machine? Wired.
<https://www.wired.com/story/ai-text-generator-gpt-3-learning-language-fitfully/> (27.03.2022)
- Simonite, T. (2021, 6. June). What Really Happened When Google Ousted Timnit Gebru. Wired. <https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/> (27.03.2022)
- Ray, T. (2020, 1. June). OpenAI's gigantic GPT-3 hints at the limits of language models for AI. ZDNet.
<https://www.zdnet.com/article/openais-gigantic-gpt-3-hints-at-the-limits-of-language-models-for-ai/> (27.03.2022)
- Ray, T. (2019, 22. May). No, this AI can't finish your sentence. ZDNet.
<https://www.zdnet.com/article/no-this-ai-cant-finish-your-sentence/> (27.03.2022)
- Osborne, C. (2021, 22. February). Google fires top ethical AI expert Margaret Mitchell. ZDNet.
<https://www.zdnet.com/article/google-fires-top-ethical-ai-expert-margaret-mitchell/> (27.03.2022)
- Piatetsky, G. (2014, 20. February). Interview with Yann LeCun, Deep Learning Expert, Director of Facebook AI Lab. KD Nuggets.
<https://www.kdnuggets.com/2014/02/exclusive-yann-lecun-deep-learning-facebook-ai-lab.html> (27.03.2022)
- Kurenkov, A. (2020, 27. September). A Brief History of Neural Nets and Deep Learning. Skynet Today. <https://www.skynettoday.com/overviews/neural-net-history> (27.03.2022)
- Weng, L. (2021, 11. July). What are Diffusion Models?
<https://lilianweng.github.io/posts/2021-07-11-diffusion-models/> (27.03.2022)
- Snell, C. (2021, 30. June). Alien Dreams: An Emerging Art Scene. Machine Learning at Berkeley. <https://ml.berkeley.edu/blog/posts/clip-art/> (27.03.2022)
- Marx, D. (2021). A brief history of PyTTI. <https://pytti-tools.github.io/pytti-book/history.html> (27.03.2022)
- AI set to exceed human brain power, Interview with Nick Bostrom. (2006, 9. August). CNN.
<https://web.archive.org/web/20061103005608/http://www.cnn.com/2006/TECH/science/07/24/ai.bostrom/> (27.03.2022)
-
- Theories of Everything with Curt Jaimungal. (2021, 23. March). Noam Chomsky on Terence McKenna, Sam Harris, GPT3, Cryptocurrencies, Kierkegaard, and Hofstadter [Video]. YouTube. <https://youtu.be/c6MU5zQwtT4?t=4324> (27.03.2022)

- [1] Scaling Facebook to 500 Million Users and Beyond (2010, 21. July). Facebook. <https://web.archive.org/web/20130705180356/https://www.facebook.com/notes/facebook-engineering/scaling-facebook-to-500-million-users-and-beyond/409881258919> (27.03.2022)
- [2] List of Dirty Naughty Obscene and Otherwise Bad Words. (2020, 13. June) <https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words/blob/master/en> (27.03.2022)
- [3] This person does not exist (2022). <https://this-person-does-not-exist.com/> (27.03.2022)
- [4] 'Fake' Amazon workers defend company on Twitter. (2021, 30. March). BBC. <https://www.bbc.com/news/technology-56581266> (27.03.2022)
- [5] Quach, K. (2021, 6. April). Twitter nukes AI-generated twits who backed Amazon and pushed anti-union rhetoric. The Register. https://www.theregister.com/2021/04/06/in_brief_ai/ (27.03.2022)
- [6] Deepfake. (2022, 28. March) Wikipedia. <https://en.wikipedia.org/wiki/Deepfake> (27.03.2022)
- [7] Ramesh, A., Pavlov, M., Goh, G. & Gray, S. (2021, 22. Juni). DALL·E: Creating Images from Text. OpenAI. <https://openai.com/blog/dall-e/> (27.03.2022)
- [8] Miller, M., Parogni, I. (2022, 28. February). The Hidden Image Descriptions Making the Internet Accessible. New York Times. <https://www.nytimes.com/interactive/2022/02/18/arts/alt-text-images-descriptions.html> (27.03.2022)
- [9] Dayma, B., Patil, S., Cuenca, P., Saifullah, S., Abraham, T., Lê P., Luke, Ghosh, R. (2022, 23. February). OpenAI's DALL-E Mini Explained with Demo. <https://wandb.ai/dalle-mini/dalle-mini/reports/OpenAI-s-DALL-E-Mini-Explained-with-Demo--Vmlldzo4NjlxODA> (27.03.2022)
- [10] Karkkainen, K., & Joo, J. (2021). FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1548-1558). <https://github.com/jojojs/fairface> (27.03.2022)
- [11] Murdock, R. @advadnoun. (2021, 6. January). "Utilizing the features of @OpenAI's CLIP". Twitter. <https://twitter.com/advadnoun/status/1346767585266679808?s=20> (27.03.2022)
- [12] Klingemann, M. @quasimondo. (2021, 9. January). "Oh wow - it works! StyleGAN2 vs. @OpenAI's new CLIP model". Twitter. <https://twitter.com/quasimondo/status/1347956102898606081?s=20> (27.03.2022)
- [13] Murdock, R. @advadnoun. (2021, 22. January). "Maybe kinda niche, but a creepy pasta about being followed around in the latent space of an AI by some strange entity would be pretty rad". Twitter. <https://twitter.com/advadnoun/status/1352499712310480898> (27.03.2022)
- [14] Klingemann, M. @quasimondo. (2021, 23. January). "A bug-eyed monster eating a 🍔 sandwich". Twitter. <https://twitter.com/quasimondo/status/1353032620700008449> (27.03.2022)

- [15] Murdock, R. @advadnoun. (2021, 24. January). "Lost in the Lost & Found". Twitter. <https://twitter.com/advadnoun/status/1353185676020256769> (27.03.2022)
- [16] Murdock, R. @advadnoun. (2021, 27. January). "Addendum". Twitter. <https://twitter.com/advadnoun/status/1365590542457823240> (27.03.2022)
- [17] Murdock, R. @advadnoun. (2021, 22. January). "Liking the ResNet50 version of CLIP". Twitter. <https://twitter.com/advadnoun/status/1352511694908518400> (27.03.2022)
- [18] Ingham, M., Letts, A. (2022, 12. March). Github. <https://github.com/alembics/disco-diffusion> (27.03.2022)
- [19] Kogan, G. (2015). Flickr. <https://www.flickr.com/photos/genekogan/albums/72157655463879190> (27.03.2022)
- [20] VQgan-clip. (2022). Replicate. <https://replicate.com/nerdyrodent/vqgan-clip> (27.03.2022)
- [21] nerdyrodent/VQGAN-CLIP. (2022). Github. <https://github.com/nerdyrodent/VQGAN-CLIP> (27.03.2022)
- [22] Rachootin, H. @sportsracer48. (2022). Patreon. <https://www.patreon.com/sportsracer48> (27.03.2022)
- [23] Ingham, M., Letts, A. (2022). Google Colab. https://colab.research.google.com/github/alembics/disco-diffusion/blob/main/Disco_Diffusion.ipynb#scrollTo=U0PwzFZbLfcy (27.03.2022)
- [24] Zippy's Disco Diffusion Cheatsheet v0.1. (2022). Google Docs. <https://docs.google.com/document/d/1I8s7uS2dGqjztYSjPpzImXLjl5PM3IGkRWI3lCuK7g/mobilebasic> (27.03.2022)
- [25] Metz, C. (2019, 16. August). A.I. Is Learning From Humans. Many Humans. New York Times. <https://www.nytimes.com/2019/08/16/technology/ai-humans.html> (27.03.2022)
- [26] AI wonders. (2022, 9. March). AI animation in 4K Created by AI Wonders "Mushroom Visions" [Video]. YouTube. <https://www.youtube.com/watch?v=snUXZ0Y244w&t=234s> (27.03.2022)
- [27] @squall22222. (2021, 3. October). "Sound 7". Twitter. https://twitter.com/squall22222/status/1444724556623593476?ref_src=twsrc%5Etfw (27.03.2022)
- [28] Moulton, R. (2021, 23. August). Wordpress. <https://moultano.wordpress.com/2021/08/23/doorways/> (27.03.2022)
- [29] @kingdomakrilic. (2022, 23. July). "CLIP + VQGAN keyword comparison". Imgur. <https://imgur.com/a/SnSIQRu> (27.03.2022)
- [30] Nerdy Roden. (2022, 23. February). AI generated animations - now with extra spin! [Video]. YouTube. https://www.youtube.com/watch?v=EJzREoN-E_w (27.03.2022)
- [31] Durant, R. (2022). Artist Studies. <https://remidurant.com/artists/> (27.03.2022)

- [32] Made-up Artists (for CLIP+VQGAN). (2021, 22. October) Google SpreadSheets.
<https://docs.google.com/spreadsheets/d/1nMq-TjBj3t6us-npLRoLFq0VtgpVwdXCKTcQgnxKgTQ/edit#gid=0> (27.03.2022)
- [33] EZ Charts - Diffusion Parameter Studies. (2022, 12. March). Google Docs.
<https://docs.google.com/document/d/1ORymHm0Te18qKiHnhcdgGp-WSt8ZkLZvow3raiu2DVU/edit#heading=h.n7msa1nt9uri> (27.03.2022)
- [34] u/ethansmith200. (2022, 16. February). “For those who asked about init image. here's about how rough you can be just to get basic composition”. Reddit.
https://www.reddit.com/r/deepdream/comments/su3fs9/for_those_who_asked_about_init_image_heres_about/ (27.03.2022)
- [35] NVIDIA Research. Inpainting. <https://www.nvidia.com/research/inpainting/index.html> (27.03.2022)
- [36] Zsolnai-Fehér, K., Two Minute Papers (2022, 15. February). “NVIDIA’s New AI: Superb Details, Super Fast!” [Video]. YouTube.
<https://www.youtube.com/watch?v=eaSTGOgO-ss&t=266s> (27.03.2022)