Meaning and Language in Artificial Intelligence's own Linguistic Turn

Bruno Caldas Vianna[1]

[1]*Centre for the Image and Multimedia Technology, Catalonia Polytechnic University (CITM-UPC). Terrassa, Barcelona, Spain*

Abstract

This paper examines the rapid development of generative visual artificial intelligence, particularly text-to-image models, through the lens of 20th-century philosophy of language. It contrasts the one-to-one symbolic representation reminiscent of early Wittgenstein and traditional AI with the subsymbolic, conceptual synthesis found in modern neural networks like CLIP and diffusion models. The paper traces the technological evolution from GANs to diffusion, analyzing the implications for creativity, originality, and the machine's capacity for metaphor. It argues that while current models can visually render complex concepts by learning from vast datasets, they struggle with true abductive reasoning and novel metaphor creation, thus highlighting the persistent limits of machine understanding when compared to human language.

Keywords

Generative AI, Text-to-Image Models, Philosophy of Language, Computational Semantics, Neural Networks

1. Introduction

This paper aims to look at current developments of generative visual artificial intelligence using tools developed by philosophers belonging to the so-called linguistic turn. The introduction of OpenAI's Clip model [1] in 20211, which ranked sentences according to how well they represented a given image, opened the gates to a plethora of text to image models. At the time of writing, these systems allow the creation of highly coherent pictures, simply by giving a textual reference of what must be generated. We will look at the consequences brought by these developments under the optics of the philosophy of language.

2. Philosophies of language

In the end of the nineteenth century and beginning of the twentieth, philosophers began questioning issues of language, like symbolic representation, its relation to the physical world and its mental concept. After all, we can't avoid using language to understand and define the world we live in. What is the relation, therefore, between the things and the way we represent them in our thoughts and discussions?

Gottlob Frege, working on mathematical proofs, arrived at problems of equivalence: how can language affirm that "The morning star is identical to the evening star" without any further inspection? This led him to the theory of Sense and Reference (or Sense and Denotation, in some translations), published in 1892 [2]. Frege, innovatively, treated sentences as functions, where the references could be exchanged by equivalent values (both "morning star" and "evening star" are the plant Venus), but the sense (or thought) they provoked would be different.

5th International Conference on Creative/Media/Technologies 2023, St. Pölten University of Applied Sciences, Austria
bruno.caldas@citm-upc.edu (B. Caldas Vianna)
0000-0003-0213-6115 (B. Caldas Viana)

Bertrand Russell picked upon some of these problems. A prominent mathematician, co-author of one of the most seminal compendia in the field, the Principia Mathematica, he wrote a first take on language is named exactly On Denoting (1905) [3], tackling questions around phrases that don't denote anything ("the king of the United States"), the ones that do define an object ("the president of the United States") and the ambiguous ones ("a president"). By applying symbolic logic to textual expressions, he worked notions of identity, x is x or Scott is the author of Waverly: "Scott wrote Waverley; and it is always true of y that if y wrote Waverley, y is identical with Scott".

At almost the same time, Ferdinand de Saussure was teaching his Course on General Linguistics [4], published only after his death in 1913. Different from Frege and Russell, his background was on tongues, investigating, for instance, the Proto-Indo-European language. Also different from them both, Saussure was more concerned with the relation between the sense and its linguistic representation, as opposed to the sense and the thing itself. From Saussure comes the study of this representation, or the sign, which derived into Semiotics.

But the most interesting takes in the context of this article come from another early twentieth-century thinker. Ludwig Wittgenstein had a flaming personality and only published one book while alive: The Tractatus Logico-Philosophicus [5]. This text defines what is known as the early Wittgenstein philosophy, including the picture theory of meaning. According to this concept, there is a one-to-one relationship between names and the world: "A name, if there is no object that it signifies, is meaningless: it is not a name at all." The complete theory also draws the concepts of elementary propositions are built from names, and depict states-of-affairs, while propositions are composed of the elementary ones, and represent the facts, which in their turn constitute the "totality of the world"[6].

This one-to-one relation, as one can imagine, is hard to sustain. Language is ambiguous and multiple. But it implies another aspect of the early Wittgenstein is relevant to us: that some things can not be expressed in words and therefore can only be "shown": "What can be shown cannot be said" [5]. The classic interpretation of these statements assumes that we cannot express the relation between language and the world, and that it can only be shown. Together with the ending proposition of the Tractatus, "Whereof one cannot speak, thereof one must be silent", [5], it allows the conclusion that his philosophy is concerned with the ineffable, what lies around the limits of language.

3. Meaning in philosophy and in computing

The one-to-one relation between objects and their representation is vital, however, in computer science. Programming a machine in the traditional manner (excluding quantum computing and even neural networks, as we'll see) requires an unequivocal connection between what is being processed and how it is stored within the processor. There can be no ambiguous ways to keep a number or an instruction in memory, or operations will fail. It is no wonder that the classical approach to artificial intelligence is referred to as symbolic computation, or good old-fashioned AI (GOFAI) [7] . This means that everything we want to process -- images, sound, text -- must be encoded in a way that the CPU can deal with. Reality is sliced into discrete pieces, ending up in bytes that can hold color intensities or look-up tables for letters. While storing images or sound waves is not so difficult, dealing with concepts (as compared to fixed representations) is a problem that still hangs.

Wittgenstein himself critiques the one-to-one relationship in his later philosophy, and proposes the concept of language-games, which allows room for the multiple possibilities that any word contains [8]. A consequence of this is the postulate that meaning is use. "The meaning of an expression is what we understand when we understand that expression. Understanding consists in knowing the expression's use across the variety of language-games in which it occurs" [6].

Wittgenstein is, without a doubt, a very influential writer with numerous disciples. His writing style and stark propositions make him a very quotable author, as the few samples above demonstrate. But his actual influence on philosophy is questioned: his concepts are too imprecise, and in consequence too open to conflicting interpretations. The "attempts to put Wittgenstein's

views into practice show that they do not constitute a solution to philosophical difficulties" [6]. Yet, even if the applications of Wittgenstein's thoughts are questionable, the tools he developed are very useful for the purpose of understanding the consequences of text to image tools in 2022.

The invention of the Word2Vec network, in 2013 [9], represented a breakthrough in the codification of concepts. Computer-based semantics had been approached in a number of ways before. The WordNet database was started in Stanford in the 1980, affording a manually built web of relations between words and their meanings [10]. The CYC company was founded in the 1970s to tackle the problem of organizing the whole intuitive knowledge of the worlds as symbols. It contains millions of definitions [11] and is still being updated. But Word2Vec took the subsymbolic path of machine learning and proved it to be highly efficient. By simply feeding the database with numerous texts, the neural network was able to organize their signifieds of words numerically. This mean that the words banana, mango, and even fruit, had their indexes (actually, large vectors) not much far from each other. More than that, the numbers could be used to operate on meanings mathematically: subtracting man from king, and adding woman to the result, yields a number in the vicinity of the word queen.

This abstraction leads us to another philosophical concept, the eidetic reduction. Developed within the field of phenomenology, in particular by its founder Edmond Husserl, this process promotes an effort to reduce a thing to its conceptual essence [12]. The experience of seeing a chair, for instance, gives us a concrete perception of its color, number of legs, material. But that it not what constitutes a chair: the "chairness" is closely tied to its function of letting a person sit. We can expand the concept to a variety of instantiations - chairs of different colors, with arm rests etc. - until we get to an essence of what a chair is. The process undertaken by the neural network is analogous to this reduction: through the inference made from numerous texts that mention chairs, the data organization will capture the concept of chair, irrespective of the characteristics of individual chairs.

4. From GANs to Diffusion

The evolution from the first attempts at visual processing with neural network until generative adversarial networks (GANs) was a process of several decades. Artificial intelligence broke ground in generative arts by allowing any style to be mechanically generated, given it had a training set with enough consistent pictures [13]. Before that, generative art was limited to what algorithms could express through symbols. This had the effect that geometrical or random-based visual were more commonly represented in computational arts; in fact, it is still the case for algorithmic artists that do not use AI.

GANs were invented in 2014 by Ian Goodfellow [14], but the techniques needed to put them together started much earlier. Convolution, which is needed for dimensionality reduction across layers, was imagined as early as 1975 [15] and put to use in the eighties by Yan LeCun [16]. Backpropagation in neural networks - the capacity of networks to evolve from their error - has been a practice since the publication of the Parallel Distributed Processing book in 1985 [17]. Some similar implementations had already appeared in 1970 [18] and it is not, after all, very far from a founding concept from cybernetic science, the feedback mechanism [19].

Goodfellow's idea was refined and made more powerful (since the original code spawned images with very low resolution) by new methods like DCGAN [20] (Radford, Metz, and Chintala 2016), which was used by artist Robbie Barrat to generate images in the style of impressionists, or nineteenth-century nudes. By 2018, AI paintings were being sold in fine art action houses, and, in 2019, a useful photorealistic face generated by a neural network could already be obtained thanks to StyleGAN2 [21].

In general, the history of machine learning is characterized by tinkering, testing, and mixing up heterogeneous techniques, and it was no different with text to image networks. Their turning point can be traced to the release of OpenAI's CLIP in January 2021 [1], which, taking inspiration

from Word2Vec, could rank text descriptions according to how well they corresponded to images. Its greatest insight is the ability to vectorize both text and images in the same latent space, allowing to understand the textual meaning of what an image represents. From the moment it was published, CLIP inspired a myriad of efforts to turn it around and make it a text to image tool.

The experiments in creating a system that translated textual concepts into images started before CLIP appeared. A model written in 2015 managed to create rudimentary but convincing low-resolution depictions of concepts [22]. In 2016, another approach using GANs - which were the state of the art in generative images then - produced much more compelling results [23].

In fact, at the same time CLIP was first released, OpenAI also published the first version of DALL·E, their own text to image resource. It's curious to notice that it was not based on the CLIP text model. Instead, they curated their own dataset of 250 millions text-image pairs from the Internet [1]. Images were generated not with GANs, but with a specifically developed method which included a discrete variational autoencoder (dVAE). It is fair to say that DALL·E was the first generator of its kind to create an impact. Its cherry-picked outputs could pass as a human-made image, and it attracted the interest of media and the generative AI community. But the release of CLIP made the community eager to code a model based in its textual capabilities. CLIP was published on January 5, 2021. On the 18th, there was already a model published by Ryan Murdock connecting it to a BigGAN (Brock, Donahue, and Simonyan 2019) generator. Later on the same day, a variation was already available, and many others followed. One of its most popular implementations was the VQGAN-CLIP by Katherine Crowson, released in April 2021 (although the corresponding paper was only published in 2022) [24]. The VQGAN generator had been released only a few months before [25]. Generative AI is the direct child of open source code, open science, collaboration and remixing.

VQGAN-CLIP was incorporated into a few popular AI-powered generative art sites, like StarryAI and Nightcafe. It was also used by autonomous art projects like Botto and Abraham. But in the end of 2021, the attention of the community started to be drawn to another technique, commonly named diffusion.

Diffusion models can be traced back to an idea from Jascha Sohl-Dickstein, published in a 2015 paper [26]. What is interesting about the proposed technique is that it refers back to a central question in cybernetics and systems theory: non-equilibrium thermodynamics. We can define entropy as the measure of disorder, or uncertainty or, for the effects of image generation, the measure of visual noise. The second law of thermodynamics implies that the entropy cannot spontaneously decrease. We could formulate it in another way: to create order, some external energy must be applied. One of the greatest challenges in systems theory is to explain how order appears spontaneously in a chaotic universe, for instance, when chemical elements self-organize in living organisms. Notably, the work of Ilya Prigogine has cast light on many of these issues [27, 28]. What diffusion models do is to apply a method that reduces entropy from a grid of random noise pixels in a guided way, until it forms a coherent image. To see such ideas being used to forge artistic imagery is a kind of vindication for systems theorists such as Maturana and Varela.

By 2021 it was already clear that diffusion models worked better than GANs, [29] but they were not nearly as popular. In a short period, OpenAI had released Improved-Diffusion [30] and Guided-Diffusion. By November - months after CLIP was used to generate images with VQGAN - they released GLIDE, a code that combined the textual skills of CLIP and the imagetic capabilities of diffusion models [31]. Between the end of the year and the beginning of 2022, an explosion of different diffusion models corroborated the potential of the approach. Disco-diffusion was a compilation of methods first authored by Max Ingham from a notebook written by Katherine Crowson and developed collaboratively on a public repository. It gradually incorporated collaborations that included animation, tridimensional movements, super resolution or virtual reality optics. CompVis, a group belonging to Ludwig Maximilian University of Munich, then released Latent-Diffusion [32]. OpenAI published the second version of DALL·E [33] in April. In May, Google released their own diffusion model named Imagen [34]). Companies like Midjourney started to build business models from image generation tools.

But the pinnacle – until 2024 - of the development in generative models came with Stable Diffusion, which is a variation of Latent Diffusion based on the Laion image database and on CompViz's Latent-Diffusion method that was made possible by the support from the company Stability AI. The model was incorporated in a matter of months into several commercial services, and since it is small enough to be run by relatively cheap GPU's, it became the model of choice for many creators.

What was impressive about these algorithms was the coherence of the images generated by them. As the methods for fine-tuning evolve, users have each time more control on what is being created. This means that instead of being limited to a single category of coherent pictures, AI generated imagery is being limited only by what we can express in words and sentences. I think the effects of this development are tremendous, and we are only seeing the beginning of it. In the next section, I'll try to break down these consequences as a few themes in order to tackle them.

5. Creativity and originality

Fear of automation and consequent loss of jobs has followed the developments of AI since the beginning [19] and it is also present here. It should be clear in this case that the fear is justified: these tools can, in fact, be used to replace highly skilled illustrators, notwithstanding the huge amount of work needed to create an artificial print comparable to a human-made version.

As previously mentioned, we are already seeing a great amount of AI-produced imagery being used beyond the experimental sphere. Several online art communities banned the exhibition of artifacts made with AI tools. The stock image marketplace Shutterstock does not accept any such content either. A "painting" created with the Midjourney generator won the first place in a fine arts fair, triggering the reaction of other artists. Besides the loss of jobs (and the appearance of new ones, namely of "prompt artists", with sites dedicated to the writing of these guiding texts), there remains the issue of creativity. On one hand, all these models were trained on works of human artists, downloaded from the Internet, which were not necessarily public. One of the most "prompted" artists, Greg Rutkowski, has already mentioned his discomfort with the capacity of anyone to replicate his own style. We can already imagine a future where pen and paper artists forbid any digital representation of their piece to keep their style from being copied.

The ripples these developments provoke in society are significant because the human access to language has always been related to our creative capabilities. The first invented stories emerged at the same time when humans could narrate events by the fire. The oldest evidence of artistic expression are visual representations of animals and daily scenes in cave walls. Language allows us to create analogies, and using analogies we can solve problems and explain complex concepts. When comparing human abilities to computers skills, Umberto Eco said that "no computer is capable of creating a metaphor". Today we can ask ChatGPT to write metaphors for anything. But is it creating or merely recycling concepts from billions of digested sentences?

6. Language Affordances

The machine that materializes a graphic representation of a textual representation finally exists, and apparently it does not work in a one-to-one relationship like the early Wittgenstein laid out. Although symbolic computing is absolutely indispensable - it is the method of choice to equip the gears that allow neural networks to run - the quest for the holy grail of computer cognition advanced mainly through the use of subsymbolic, neural networked method. In that sense, Husserl - who is not usually included in the continental philosophy's "linguistic turn" narrative - described an abstraction method which can perfectly be used to understand the methods used by neural networks to synthesize meaning.

Because when I instruct the model to draw an image of a lioness, it will output endless variations of the female animal - but no pictures of humans or mice. Yet language does allow me to use the word lioness to refer to an entity which is not the actual animal, but a mighty feminist woman or a surprisingly brave mouse. Language-games are so: they allow for figures of speech. A computer, on the other hand, will almost always choose the literal interpretation, since this is what is overwhelmingly represented in the training data.

This is understandable. Coming from a filmmaking background, I believe that the translation of ideas and stories into graphic representations for concepts is one of the most demanding tasks of a visual practitioner, and one of the greatest skills of talented directors and screenwriters. This is also the work, for instance, of political cartoonists, who must summarize and comment on complex situations through simple, witty illustrations. "An image is worth a thousand words", says the adage. But to express this very concept with an image instead of a few words would be very hard, as Brazilian cartoonist Millôr observed once.

The issue was also addressed by the main theorists of cinema. Christian Metz mentions how a coin being in flipped in a hand starts to represent the gangster character in George Raft's movies (Metz 1993). Another well-known example is the monocle of the czarist doctor in Eisenstein's Battleship Potemkin. A glove can represent the act of strangling, or a cross being dragged in the foreground can express the inner drama of a character in the background (Martin 1985). But at this point, we start to navigate the terrain of metaphors. Deleuze describes a precious one in Buster Keaton's The Navigator. When the main character is saved from drowning by a girl, she cuts open his life jacket and water bursts out of it, announcing his new figurative birth [35].

As mentioned, Eco said that computers cannot create metaphors. Making one requires a deep understanding of a specific domain, in order to create an abstraction of it, that then can be applied to a different domain. Trees do not have feelings like the ones attributed to humans. Therefore, when I say "sad tree", I do not mean that the plant is in a gloomy mood, but I'm saying something about its current state. To use a concept such as "sad tree", I need to abstract the domain of sadness, which includes loneliness (which on its turn includes emptiness) and lack of life (which on its turn includes lack of color and of leaves). "Sadness" becomes a metaphor for a visual condition, it stops denoting a feeling. So it would make sense that this expression would be translated into a tree without leaves, in faint colors, isolated in the landscape. And that is exactly what I get when I ask Stable Diffusion models to generate a "sad tree".

Does this mean that Eco's assertion needs to be re-evaluated? Probably. Does this mean that now computers can create metaphors? Probably not. The training space of visual references is so huge that the model has learned from existing representations of sadness and trees, and generated an output coherent with both domains. It did not invent a visual metaphor, but it does convey a groundbreaking ability that was considered out of reach until recently. Moravec's paradox [36], formulated just a few decades ago, stated that easy things for humans were hard for computers and vice versa. Calculating the first prime number greater than one million is easy for a computer and would require a lot of effort from a person. On the other hand, recognizing the face of a relative or drawing a sad tree used to be difficult for machines. Now, more and more, those things which were easy for us are becoming no problem for computers as well. The next big challenge for artificial intelligence is probably the capacity for abstraction without a huge learnable domain, which is another well-known human ability. Researcher François Chollet has proposed a test named Abstraction and Reasoning Corpus [37], which consists of a thousand visual puzzles for which the solution must be deduced from about three examples, instead of millions. So far, no algorithm has performed not even nearly well on the corpus.

Another challenge is to perform the opposite operation: to create a textual figurative description for the image of the derelict tree. Machines are expected to produce a literal evaluation of the visuals. The original CLIP model doesn't even propose to caption pictures. Instead, it ranks different textual classifications according to what it evaluates as being the best match. But image captioning systems have already been developed around it. When I use them to caption the sad tree pictures created earlier, the sadness is gone. The results I get are like:

A tree branch is standing in the middle of a field.
A tree that has been cut down.
A tree with a bunch of branches and a cross on it.

This limitation tells more about the purpose of the tools than their limitations. They are tuned for clarity, not for poetics. Textual machine learning models have effectively broken the barrier between image and word. My supposition is that it is already possible to build an image captioning system around figures of speech rather than literal descriptions. But it is not yet to be found, so the challenge remains.

7. Limits

The question turns now again to the limits of linguistic expression. Even if combinations of learned visual representations might make some unexpected metaphor appear, machine learning systems do not create in the same impromptu sense that humans do. And the abstract model of abductive creation is very hard to develop without resourcing to millions of examples.

When Wittgenstein cast his early philosophy, he was influenced by the optimism born out of the then-recently consolidated knowledge of how to denote transcendental questions using mathematical symbols. This achievement is what led to the development of computers by Turing and Von Neumann, the successors of Bool, Frege and Russell. We might never know if Wittgenstein really meant that the limits of one's world are the limits of what one can express. A huge chunk of his body of work is dedicated to taming language, and it is fair to say he did not manage to finish the task in his lifetime. But it also seems reasonable to think that today, the limit lies in the things we cannot mold in language, and therefore cannot be represented symbolically. Subsymbolic AI might have dominated the field in the last decade, but computers are still symbolic beasts. And while they continue to be so, human language remains a wild animal.

References

[1] 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.: Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs]. (2021).

[2] Zalta, E.N.: Gottlob frege. In: Zalta, E.N. and Nodelman, U. (eds.) The Stanford encyclopedia of philosophy. Metaphysics Research Lab, Stanford University (2022).

[3] Russell, B.: II.—ON DENOTING. Mind. XIV, 479–493 (1905). https://doi.org/10.1093/mind/XIV.4.479.

[4] Saussure, F. de, Baskin, W., Meisel, P., Saussy, H.: Course in general linguistics. Columbia University Press, New York (2011).

[5] Wittgenstein, L., Ogden, C.K.: Tractatus logico-philosophicus. Dover Publications, Mineola, N Y (1999).

[6] Grayling, A.C.: Wittgenstein: a very short introduction. Oxford University Press, Oxford (2001).

[7] Haugeland, J.: Artificial intelligence: the very idea. MIT Press, Cambridge, MA (1986).

[8] Wittgenstein, L., Anscombe, G.E.M.: Philosophical investigations: the German text, with a revised English translation. Blackwell Pub, Malden, MA (2003).

[9] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space, https://arxiv.org/abs/1301.3781v3, last accessed 2024/01/02.

[10] Fellbaum, C.: WordNet. In: Poli, R., Healy, M., and Kameas, A. (eds.) Theory and Applications of Ontology: Computer Applications. pp. 231–243. Springer Netherlands, Dordrecht (2010). https://doi.org/10.1007/978-90-481-8847-5_10.

[11] Lenat, D., Prakash, M., Shepherd, M.: CYC: Using common sense knowledge to overcome brittleness and knowledge acquistion bottlenecks. AI Mag. 6, 65–85 (1986).
[12] Husserl, E.: Ideas: general introduction to pure phenomenology. Routledge, Oxfordshire, England (2013).
[13] Caldas Vianna, B.: Generative Art: Between the Nodes of Neuron Networks. Artn. (2020). https://doi.org/10.7238/a.v0i26.3350.
[14] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Networks. arXiv:1406.2661 [cs, stat]. (2014).
[15] Fukushima, K.: Cognitron: A self-organizing multilayered neural network. Biol. Cybernetics. 20, 121–136 (1975). https://doi.org/10.1007/BF00342633.
[16] LeCun, Y., Boser, B.E., Denker, J.S., Henderson, D., Howard, R., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural Computation. 1, 541–551 (1989). https://doi.org/10.1162/neco.1989.1.4.541.
[17] Rumelhart, D.E. ed: Parallel distributed processing. 1: Foundations / David E. Rumelhart. MIT Pr, Cambridge, Mass (1999).
[18] Griewank, A.: Who invented the reverse mode of differentiation. Presented at the (2012).
[19] Wiener, N.: Cybernetics: Or Control and Communication in the Animal and the Machine. 1965 ed. MIT Press (1948).
[20] Radford, A., Metz, L., Chintala, S.: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, http://arxiv.org/abs/1511.06434, (2016). https://doi.org/10.48550/arXiv.1511.06434.
[21] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and Improving the Image Quality of StyleGAN, http://arxiv.org/abs/1912.04958, (2020). https://doi.org/10.48550/arXiv.1912.04958.
[22] Mansimov, E., Parisotto, E., Ba, J.L., Salakhutdinov, R.: Generating Images from Captions with Attention, http://arxiv.org/abs/1511.02793, (2016). https://doi.org/10.48550/arXiv.1511.02793.
[23] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative Adversarial Text to Image Synthesis, http://arxiv.org/abs/1605.05396, (2016). https://doi.org/10.48550/arXiv.1605.05396.
[24] . Crowson, K., Biderman, S., Kornis, D., Stander, D., Hallahan, E., Castricato, L., Raff, E.: VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance, http://arxiv.org/abs/2204.08583, (2022). https://doi.org/10.48550/arXiv.2204.08583.
[25] Esser, P., Rombach, R., Ommer, B.: Taming Transformers for High-Resolution Image Synthesis, http://arxiv.org/abs/2012.09841, (2021). https://doi.org/10.48550/arXiv.2012.09841.
[26] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In: Proceedings of the 32nd International Conference on Machine Learning. pp. 2256–2265. PMLR (2015).
[27] Prigogine, I., Nicolis, G., Babloyantz, A.: Thermodynamics of evolution. Physics Today. 25, 23–28 (1972). https://doi.org/10.1063/1.3071090.
[28] Prigogine, I., Stengers, I.: The end of certainty: time, chaos, and the new laws of nature. Free Press, New York (1997).
[29] Dhariwal, P., Nichol, A.: Diffusion Models Beat GANs on Image Synthesis, http://arxiv.org/abs/2105.05233, (2021). https://doi.org/10.48550/arXiv.2105.05233.
[30] Nichol, A., Dhariwal, P.: Improved Denoising Diffusion Probabilistic Models, http://arxiv.org/abs/2102.09672, (2021). https://doi.org/10.48550/arXiv.2102.09672.
[31] Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, http://arxiv.org/abs/2112.10741, (2022).
[32] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models, http://arxiv.org/abs/2112.10752, (2022). https://doi.org/10.48550/arXiv.2112.10752.

[33] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical Text-Conditional Image Generation with CLIP Latents, http://arxiv.org/abs/2204.06125, (2022). https://doi.org/10.48550/arXiv.2204.06125.
[34] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding, http://arxiv.org/abs/2205.11487, (2022). https://doi.org/10.48550/arXiv.2205.11487.
[35] Deleuze, G.: Cinema. 2: The time-image. Athlone Pr, London (1989).
[36] Moravec, H.: Mind children: the future of robot and human intelligence. Harvard Univ. Press, Cambridge (1995).
[37] Chollet, F.: On the Measure of Intelligence. arXiv:1911.01547 [cs]. (2019).