Dumb Meaning: Machine Learning and Artificial Semantics

Hannes Bajohr

In June 2022, Google employee Blake Lemoine was given an indefinite leave of absence. The reason: He had claimed that the artificial intelligence he was helping to test was sentient, and the company thought such claim bad press.[1] Lemoine insisted that LaMDA, a chatbot system, convinced him in lengthy conversations that it had the intelligence of a highly gifted eight-year-old, and asked to be considered a person with rights.[2] In doing so, Lemoine, who describes himself as "ordained as a mystic Christian priest," was merely exaggerating a sentiment that also afflicted others at Google. For Blaise Agüera y Arcas, a senior machine learning engineer not usually prone to mysticism, wrote of his own interactions with LaMDA just days before Lemoine: "I felt the ground shift under my feet. I increasingly felt like I was talking to something intelligent."[3]

In contrast, a discussion about another AI system, which took place at about the same time, did not use the buzzwords of sentience and intelligence at all. Dall-E 2, which was developed by the company OpenAI, is a text-to-image AI that can generate images from natural language input. Given a prompt such as "A Shiba-Inu wearing a beret and a black turtleneck," it produces an

---

[1] Nitasha Tiku, "The Google engineer who thinks the company's AI has come to life," *The Washington Post*, June 11, 2022, https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine. Lemoine's term is *sentience*, not *consciousness,* but he seems to use them synonymously.

[2] Blake Lemoine, "What is LaMDA and What Does it Want?," *Medium*, June 11, 2022, https://cajundiscordian.medium.com/what-is-lamda-and-what-does-it-want-688632134489; Lemoine also published the chat transcript of a conversation with LaMBDA, Id., "Is LaMDA Sentient? An Interview," *Medium,* June 11, 2020, https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917.

[3] Blaise Agüera y Arcas, "Artificial Neural Networks are Making Strides Toward Consciousness," The *Economist*, June 9, 2022, https://www.economist.com/by-invitation/2022/06/09/artificial-neural-networks-are-making-strides-towards-consciousness-according-to-blaise-aguera-y-arcas.

output image depicting that very scene.[4] The public beta triggered a slew of experiments, and soon the most interesting or whimsical results were shared on the web and especially on Twitter.

This, too, was revealing: compared to the much less successful experiments with autonomous cars, it suggested that AI has significantly different social effects than long thought – that, before it puts truck drivers out of business, it is more likely to take the jobs of illustrators, graphic artists, and stock photographers.[5] Unlike in the case of LaMDA, however, no one thought Dall-E 2 was possibly a person with rights.

The different reactions to the two systems show how quickly thinking about AI veers into familiar conceptual ruts. Intelligence, consciousness, sentience, and personhood have been the major themes of AI research and its imaginaries for nearly seventy years; amusing little pictures, by contrast, seem to raise fewer fundamental questions. But it is quite possible that it is actually the other way around – that the eternal hunt for superintelligence and the singularity obscures the more interesting and subtle conceptual shifts that escape both the tech evangelists in their visionary furor and the skeptical critics.

For philosopher Benjamin Bratton, it is clear that in the face of new AI systems, "reality has outpaced the available language to parse what is already at hand." What is needed, therefore, is a "more precise vocabulary"[6] that goes beyond the usual handful of big concepts, but also beyond the anthropocentric assumption that the only way in which machines may form world-relations would have to be our way. We can observe such a tendency with Dall-E 2 and LaMDA. Here, the

---

[4] Aditya Ramesh et al., "Hierarchical Text-Conditional Image Generation with CLIP Latents," *arXiv*, April 13, 2022, https://arxiv.org/abs/2204.06125; see also OpenAI, "Dall-E 2," April 6, 2022 https://openai.com/Dall-E-2.

[5] Prarthana Prakash, *"*AI Art Software Dall-E Moves Past Novelty Stage and Turns Pro," *Bloomberg*, August 3, 2022, https://www.bloomberg.com/news/articles/2022-08-04/Dall-E-art-generator-begins-new-stage-in-ai-development; the June 11, 2022 issue of The Economist featured an illustration generated by an image AI.

[6] Benjamin Bratton and Blaise Agüera y Arcas, *The Model Is The Message*. In *Noema,* July 12, 2022 (https://www.noemamag.com/the-model-is-the-message).

concept of meaning becomes detached from its anthropocentric correlate, and it would be meaning without mind – dumb meaning.

Free-floating and grounded systems

Despite constant admonitions from computer scientists, linguists and cognitive psychologists to use terms such as intelligence and consciousness with care, the tech industry remains relatively immune to such warnings. Thus, critics soon accused Lemoine of having fallen for the "ELIZA effect"[7] – of having projected intelligence and consciousness onto LaMDA – a susceptibility Joseph Weizenbaum had already observed in 1966 among users of his ELIZA chatbot: Although ELIZA merely mimicked a Rogerian psychoanalyst, mirroring the patient's statements back to them as questions, its users behaved as if the program really were a conscious agent interested in their well-being.

The classic objection here is the following: Computers are symbol-processing systems that deal with syntax alone, not with semantics – they can process logical forms but not substantive meaning.[8] For their operations it is irrelevant which objects or concepts the symbols name in a human world and which cultural valences are associated with them. Thus, ELIZA merely scans user input for a given syntactic pattern and transforms it into a "response" according to a transformation rule. Weizenbaum gives the example in which the analysand reproaches the analyst: "It seems that you hate me." The program identifies the key pattern "*x* you *y* me" in this sentence and separates it accordingly into the four elements "It seems that," "you," "hate," and

[7] Brian Christian, "How a Google Employee Fell for the Eliza Effect," *The Atlantic*, June 21, 2022, https://www.theatlantic.com/ideas/archive/2022/06/google-lamda-chatbot-sentient-ai/661322 .
[8] Florian Cramer, "Language," *Software Studies: A Lexicon*, ed. Matthew Fuller, Cambridge, Mass.: MIT Press 2008, 168-74.

"me." It then discards *y* ("it seems that") and inserts *x* ("hate") into the reply template "What makes you think I *x* you". And so ELIZA responds to the accusation that it hates the analysand by asking how they got that idea.[9]

This interaction may have meaning for the user and plausibly suggest a communicative intent on the part of ELIZA, but neither such intent nor such meaning is actually to be found in the program. It has merely processed symbols according to a rule without "knowing" what hate is or what behavior the mores of civil discourse suggest. That is the difference between the processing of information and the understanding of meaning.

For AI researcher who seek to make computers more human, this circumstance describes what cognitive psychologist Stevan Harnad called the "symbol grounding problem" in 1990: Symbols, like those in Weizenbaum's transformation operation, have no intrinsic meaning in computers because, without the background of practical knowledge of the world, they can only refer to other symbols, never to any reality beyond them. They are not grounded in the world. There is no way out of this "symbol/symbol merry-go-round." Whatever meaning there is can only be "parasitic," and is projected onto the output by human interpreters.[10]

Harnad's criticism, however, was directed against only one particular type of AI, which also includes ELIZA; for obvious reasons, it is called "symbolic." To solve the symbol grounding problem, Harnad relied on the novel "*subs*ymbolic" or "connectionist" systems of the time: neural networks of which LaMDA and Dall-E 2 are late descendants. Unlike traditional AI, they are not designed as a set of logical rules of inference, but are vaguely modeled on the brain as

---

[9] Joseph Weizenbaum, "ELIZA: A Computer Program for the Study of Natural Language Communication Between Man and Machine," *Communications of the ACM* 9, no. 1 (1966): 37-8. I have simplified the procedure somewhat; moreover, ELIZA allows quite different transformation rules, and the therapist is only one subroutine, called DOCTOR.

[10] Stevan Harnad, "The Symbol Grounding Problem," *Physica D: Nonlinear Phenomena* 42, no. 1-3 (1990): 340, 339.

neurons that amplify or attenuate the signals passed through them. They therefore do not require explicit symbolic representations and rules. They are not programmed, but learn independently from examples. While neural networks were mainly used for pattern recognition in the early 1990s, Harnad thought they might be able to access the world. Implemented in an autonomous, mobile robot, equipped with sensors and effectors, a conglomerate of neural networks would first receive impressions and categorize them as recognizable shapes. These would then be handed over to a symbolic AI, but would now no longer be mere references to other symbols, but connected to the world via their causal reference to external data – they would finally be grounded.[11]

The consequence of this thought, however, seems to be that the only way to get around the ELIZA effect, which falsely attributes consciousness to computers, is to *actually* give them consciousness. For what Harnad has in mind is, in the end, again an anthropocentric model that hopes embodied cognition and sufficiently extensive referential meanings will produce world understanding, since this is how we more or less function, too.[12] The success of his hybrid model would have to be demonstrated by his robot being as competent at navigating the world as if it were actually intelligent. Since this is not yet the case, the symbol grounding problem cannot yet be considered solved either; a *bit* of meaning does not exist here by definition. And yet such limited meaning is exactly what LaMDA and Dall-E 2 seem to suggest.

Gradated meaning

[11] Stevan Harnad, *Grounding Symbols in the Analog World with Neural Nets*. In: *Think* 2 (1993): 12-78.

[12] There are quite a number of objections that question the plausibility of his proposal. The most prominent among them is that no situated understanding of the world follows from the mere summation of such "grounded" symbols; this kind of argument has often been advanced by "phenomenological" criticism of AI trained on Heidegger, e.g.: Hubert Dreyfus, *Was Computer nicht können* (Frankfurt am Main: Athenäum 1989).

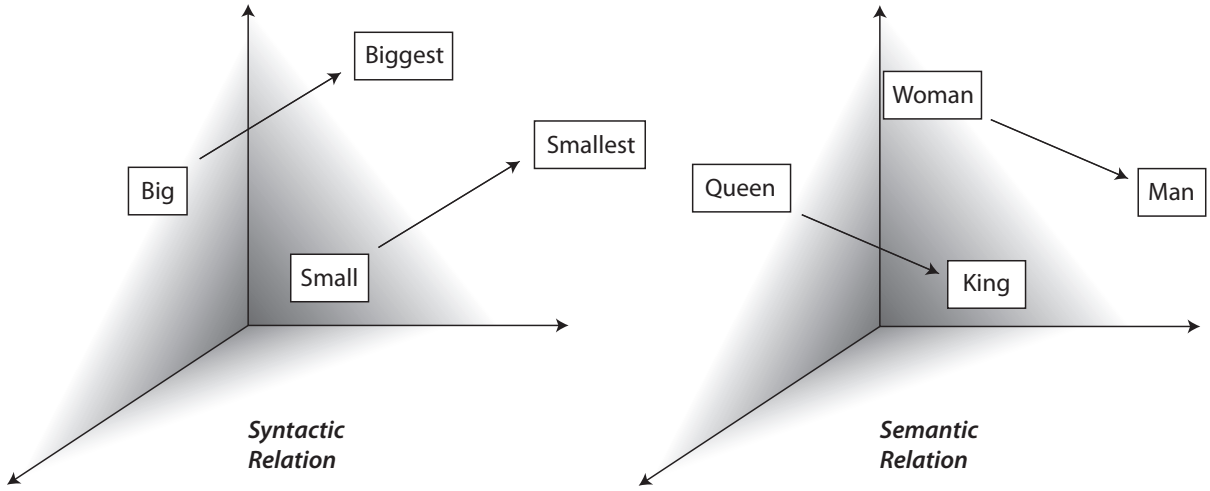
With the increasing popularity that neural networks have enjoyed for almost ten years now, the idea that they somehow could have access to meaning beyond mere ungrounded symbols has also become more attractive again. For cultural scientist Mercedes Bunz, neural networks, thanks to their complexity and capacity for unsupervised learning, can now "compute meaning" rather than just empty symbols.[13] It is true that, in the face of neural networks, the binary distinction between meaning (human world) and non-meaning (digital systems) is becoming increasingly difficult to maintain. Maybe we should consider levels of graded meaning which, as artificial semantics, no longer presuppose a mind.

Thus, instead of being a sign of consciousness, the fact that LaMDA's answers sounded so human-like can simply be understood as an indication of such "dumb" meaning. While "broad" meaning presupposes – depending on your philosophical or disciplinary orientation – embodied intelligence, cultural and social prior knowledge, or the world-disclosing function of language, dumb meaning would be beyond this scale (that is always calibrated on humans), and could best be grasped as a type of *correlation.*

LaMDA is – similar to the better-know text generator GPT-3 – a large language model implemented as a neural network. Trained on vast amounts of text, it processes language as a multi-dimensional vector space, a so-called word embedding, which works according to the principle of staggered correlations: First, words that frequently appear together are closer together in this space, forming semantic clusters familiar from word clouds. However, since not only the correlations of words to words, but also correlations of correlations are encoded, large language models can also explicate implicit regularities that are not spelled out in the training text. This is true for syntactic relations – when the Euclidean distance between the vectors for the positive and superlative of a

[13] Mercedes Bunz, *The Calculation of Meaning: On the Misunderstanding of New Artificial Intelligence as Culture*. In: *Culture, Theory and Critique*, no. 60/2, September 2019.

word is the same – but also for complex semantic relations, that is, word meaning. One of the best known examples of this principle is the operation: $v_{king} - v_{man} + v_{woman} \approx v_{queen}$.[14]



Word embedding of a large language model

In this equation – which reads: subtract from the word vector "king" that for "man" and add that for "woman," and the result is the word vector for "queen" – the latent semantic relation "gender" emerges as an arithmetic correlation, even though it is not explicitly present in the model. (That it emerges from the mass of language on which the model is trained explains machine learning's susceptibility to biases: sexism and racism may also be latently encoded in language models).[15] The meaning of a sign in a language system constructed in this way is determined purely *differentially*, as in Ferdinand de Saussure's linguistic structuralism. Instead of referring to anything outside language, sign meaning is simply thought of as difference from other signs and sign

[14] This was first discovered in 2013 by the inventors of the word-embedding model Word2vec. Cf. Tomas Mikolov/Wen-tau Yih/Geoffrey Zweig, *Linguistic Regularities in Continuous Space Word Representations*. In *Proceedings of the 2013 Conference of the NAACL*. Atlanta: ACL 2013. However, this insight still applies to newer, technically different models such as GloVe (Global Vectors for Word Representation).

[15] See Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, und Shmargaret Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" In *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, New York: Association for Computing Machinery, 2021, 610–23.

correlations.[16] The effect, nevertheless, is that large language models, by their immense training data alone, are able to produce apparently situational understanding, as LaMDA did, without ever being "in a situation."[17]

Language models would then be producers of a first degree of dumb meaning. It is dumb because the model captures latent correlations between signs, but still does not "know" what things these signs actually name; with this kind of meaning, one will not be able to build an intelligence that will ever find its way in the world. The linguist Emily Bender, a vehement critic of all AI hype about alleged consciousness, admits with her colleague Alexander Koller that "a sufficiently sophisticated neural model *might* learn some aspects of meaning," such as semantic similarity, but considers them to be "only a weak reflection of actual meaning," which is always related to something in the world, i.e. "grounded." [18]

[16] This is nicely explained in Juan Luis Gastaldi, "Why Can Computers Understand Natural Language? The Structuralist Image of Language Behind Word Embeddings," *Philosophy & Technology* 34, no. 1 (2021): 149–214.

[17] This is philosopher Hubert Dreyfus' term for the prior world-understanding that humans, but not computers, have, and which he develops from a hermeneutic, Heideggerian critique of AI, see Hubert L. Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason,* Cambridge, Mass.: MIT Press, 1992.

[18] Emily M. Bender and Alexander Koller, "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online: Association for Computational Linguistics, 2020), https://doi.org/10.18653/v1/2020.acl-main.463, 5191, 5193.

But as wrong as it would be to project anything like sentience or consciousness onto this system, one should also not be too quick to dismiss this modicum of meaning. Insofar as language models make implicit knowledge explicit in a non-trivial way – even if only by matrix transformations in a vector space – they produce dumb meaning which would not have been available without them. In contrast to ELIZA – whose *x* and *y* were only empty placeholders to the system – neural networks are not *solely* parasitically dependent on the meaning attributions of human agents, but *also* operate productively with the inherent structure of language.

Text and image and world

Bender is of course right that LaMDA is not grounded.[19] It is a *mono*modal network, processing only a single type of data, namely text. To be grounded in Harnad's sense, she writes, it would be necessary to combine several types of data – it would have to be *multi*modal machine learning. [20] That is what Dall-E 2 is: Instead of text just referring to other text, here text is correlated with image information. This raises the hope that arbitrary signs can be linked to things in the world to produce grounded meaning.

Harnad's hypothesis that neural networks in particular could address the symbol grounding problem has recently been taken up by media scholars Leif Weatherby and Brian Justie with their notion of "indexical AI." It is named after Charles Sanders Peirce's notion of the index. Unlike the symbol, which has a purely conventional relationship to its signified (as "dog," "chien," and "Hund" all refer to the same thing), the index is causally linked to it (as smoke refers to fire).

---

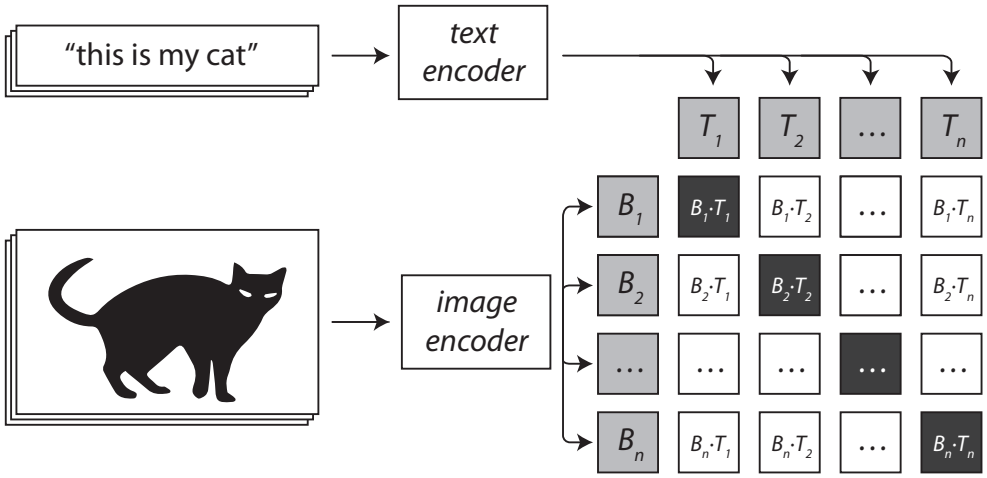
[19] While the paper presenting LaMDA also claims "groundedness" for the model, what is meant by this is simply that LaMDA's outputs are "grounded in known sources wherever they contain verifiable external world information." As *textual sources,* they continue to be part of Harnad's "symbol/symbol merry-go-round." Romal Thoppilan et al., "LaMDA: Language Models for Dialog Applications" (arXiv, February 10, 2022), http://arxiv.org/abs/2201.08239.

[20] See Gadi Singer, "Multimodality: A New Frontier in Cognitive AI," Medium, February 2, 2022, https://towardsdatascience.com/multimodality-a-new-frontier-in-cognitive-ai-8279d00e3baf.

With this coinage, the authors make Harnad's project the basis of contemporary cultural description: "Digital systems, relying on the neural net, have left the world of mere symbol behind and have begun to ground themselves *here*, *now*, for *you* – they are able to *point* to real states of affairs."[21] Neural networks bring the world – as the data on which they have been trained – into the computer, getting off of the solipsistic "symbol/symbol merry-go-round." Nowhere does this seem more plausibly demonstrated than in Dall-E 2.

The heart of Dall-E 2 is a machine learning model called CLIP. Via an encoder, it is fed with vectorized text-image pairs taken from the Internet – for example, a photo of a cat with the caption "this is my cat." CLIP is trained to predict which text vector matches which image vector; the result is a comprehensive stochastic model that correlates image information with text information, but stores it as *one* type of information. In the figure below, this is the table in which the scalar product of the text and image vectors is entered – the better the text and image fit, the better this value; when the original image and text are paired, it is of course optimal (those are the black boxes running diagonally).

[21] Leif Weatherby and Brian Justie, "Indexical AI," *Critical Inquiry* 48, no. 2 (2022): 382. One difficulty with this notion is the question of whether *all* data in a neural network should already be considered indexical (that would include the text of LaMDA) or only those obtained directly by sensors emulating physical senses (that would be images, but not text). Weatherby and Justie seem to have the former in mind, Harnad the latter. Harnad therefore speaks at one point of "iconic" data – Peirce's third sign type, which operates on the principle of similarity between sign and signified. But since these are also indexical as they originate from sensors (which, of course, limits them to visual data), it seems to me that the argument of Weatherby/Justie and that of Harnad amount to something structurally similar – both are concerned with the connection between system and world.



Text-image correlation in CLIP

CLIP is thus an amazingly good at *image recognition*: If you present it with an unknown cat photo, it nevertheless recognizes it as "cat." But in a second step, it also becomes an *image generator*. To do this, it works in conjunction with another machine learning model called GLIDE, which has already been trained on a large data set of images.[22] If the user enters a prompt, GLIDE can use the text-image data stored in the CLIP model to reverse this process and synthesize an image that best correlates with the input text. In both operations – image recognition as well as image generation – it is again central that the models can learn and actively reproduce the *correlation* between textual descriptions of objects and their corresponding visual manifestations.

One may object that the image information correlated with the word "cat," in which the photo of a cat is stored, may have an indexical relation to this cat – light was reflected from it and fell on a photo sensor etc. –, but that even so the system will not learn what it means to share a world with

[22] GLIDE is a *diffusion model* based on thermodynamic models, and thus functions differently from the GANs that were popular until recently, which combine two antagonistic submodels. See Prafulla Dhariwal and Alex Nichol, "Diffusion Models Beat GANs on Image Synthesis" (arXiv, June 1, 2021), https://doi.org/10.48550/arXiv.2105.05233. That the AI architectures used for an aesthetic work can themselves be a resource for discussing that work is something I suggest in: Hannes Bajohr, "Algorithmic Empathy: Toward a Critique of Aesthetic AI," *Configurations* 30, no. 2 (2022): 203–31.

a cat. Advocates of symbol grounding therefore try to feed extend what types of data an AI model gets fed – not only sensory but also motor and eventually even social feedback: Only through the effects of language use in a community of other speakers inhabiting the same world can meaning be learned.[23]

But this claim would again mean to demand "full" human, that is, broad meaning, and to take anything below that not quite seriously. Instead, multimodal AI should be regarded as a second degree of dumb meaning. The Peircean indexical reference to something outside the model and the Saussurean differential reference to other elements within it are at any rate two distinct ways of meaning-making – if only that the dimension of possible correlations increases, and with it the possibility of unearthing unsuspected latent connections, unsuspected dumb meaning.

Indeed, multimodal AIs – besides Dall-E 2, for instance, the free but unrelated Dall-E Mini (now Craiyon), Stable Diffusion, Google's yet-to-be-released Imagen, or the paid model Midjourney – are capable of generating very complex text-image meanings. Their power lies in a capability that suggests that such correlations have a productive quality: In studying the deep structure of CLIP, computer scientists found that the model had trained single "neurons" that fired for both the word and the image of a thing. These were *conceptual* neurons in which the distinction between image and text tended to be overcome.[24] Multimodality, at the neural level, is really *pan*modality, suggesting a semantics without clearly differentiated sign systems. Dumb meaning finds a new quality here, and is not tied to either text or image data, but encompasses both in a way that points to meaning beyond modal separation – and again has nothing to do with mind.

[23] See Yonatan Bisk et al., "Experience Grounds Language" (arXiv, November 1, 2020), http://arxiv.org/abs/2004.10151.
[24] Gabriel Goh et al., "Multimodal Neurons in Artificial Neural Networks," *Distill* 6, no. 3 (March 4, 2021): 10.23915/distill.00030, https://doi.org/10.23915/distill.00030.

Promptological investigations

AI systems *are* dumb. They have no consciousness. Yet they produce a complex artificial semantics that runs counter to our ordinary notions of meaning. Multimodal AI also shows that imputed consciousness and the meaning-capacity of a system have little to do with each other: The fact that LaMDA in particular seemed to Lemoine like a person – and not Dall-E 2, which actually represents a higher, because more correlation-rich stage of AI development – is simply due to the fact that it operates via dialog and thus is assumed to have communicative intent, whereas the image generator does not. Language always seems to be smarter than the image.

However, meaning beyond communicative intent need not be *merely* parasitic, as the vector operations of word embeddings and the conceptual neurons of text-to-image AIs show. That it is always *also* parasitic is due to the fact that the training data originate from a human world and artificial semantics is precisely not a "robot language" but a correlation effect of information that can be interpreted by humans. Nevertheless, in the long run, a convergence of dumb and broad meaning would be conceivable once they enter into mutually influencing circular processes.

The interface between natural and artificial semantics in the case of Dall-E 2 is the interaction via prompt. On the one hand, "prompt design" – the precise, almost virtuosic selection of the text input – can be used analytically to scan the vector space of dumb meaning for traces of cultural knowledge. This would once again elevate the broad meaning of natural language in its interaction with dumb meaning.

A "promptology" that takes on such natural-artificial connections – the correlation of datafied language and the cultural meaning attributed to that language on the recipient side – would be a gateway for the humanities and cultural studies, which, with their knowledge of such soft factors as styles, influences, iconography, etc., could make their contribution without necessarily taking the form of the more computer science-focused digital humanities; they could work in a

phenomenon-oriented way and devote themselves to the artifacts that the model outputs as boundary objects between human and machine, between broad and dumb meaning.

At the same time, however, promptology is not merely an analytical procedure, but also a practice with its own knowledge, which has much to do with an almost "empathetic" interaction with the AI system. It has turned out that with text-to-image AIs, these prompts can be steered in unexpected directions simply by using certain, often counterintuitive or absurd formulations (there is already a start-up, PromptBase, which claims to sell particularly effective prompts).[25] Instead of subjugating the system and using it as an instrument, natural language instead must be adapted to the artificial semantics just to operate the system.

The result is a feedback loop of artificial and human meaning: not only does the machine learn to correlate the semantics of words with those of images we have given it, but we learn to anticipate the stupidity of the system in our interaction with it; this convergence would not be communicative in a strong sense, but perhaps in a weak, a dumb, sense.

[25] Kyle Wiggers, "A startup is charging $1.99 for strings of text to feed to Dall-E 2," *TechCrunch*, July 29, 2022, techcrunch.com/2022/07/29/a-startup-is-charging-1-99-for-strings-of-text-to-feed-to-dall-e-2. What is interesting here is that the discussed tendency to eliminate the speech/image distinction at the *technical* level is contrasted with the displacement of the image by speech at the *interface level.* The results of Dall-E 2 could therefore also be understood as *language art* instead mere visual objects.