Shaping Vectors: Discipline and Control in Word Embeddings

Pierre Depaz

Shaping Vectors: Discipline and Control in Word Embeddings

Pierre Depaz <pierre.depaz@nyu.edu>

Depaz, P. (à paraître). Shaping Vectors: Discipline and Control in Word Embeddings. *A Peer-Reviewed Journal About*, *13*(1). https://aprja.net/

Abstract

This article investigates how the word embeddings at the heart of large language models are shaped into acceptable meanings. We show how such shaping follows two educational logics. The use of benchmarks to discover the capabilities of large language models exhibit similar features to Foucault's disciplining school enclosures, while the process of reinforcement learning is framed as a modulation made explicit in Deleuze's control societies. The consequences of this shaping into acceptable meaning is argued to result in semantic subspaces. These semantic subspaces are presented as the restricted lexical possibilities of human-machine dialogic interaction, and their consequences are discussed.

Keywords: Word embeddings, semantic subspaces, discipline, control, benchmarking, reinforcement learning.

Biography

Pierre Depaz is currently a Lecturer of Interactive Media at NYU Berlin. His research focuses on understanding how software operates procedural translation of non-computational entities, and how it affects humans' perceptions and affordances with the world. ORCID: https://orcid.org/0009-0009-1489-247X

Note

This article has benefited greatly from thorough discussions with and copy edits by Sara Messelaar Hammerschmdit.

Introduction

When following the direction from *man* towards *programmer* in a space composed of word vectors, computational linguists Bolukbasi et al. encountered a problem -- the resulting value when starting from *woman* was *homemaker* (Bolukbasi et. al., 2016). In order to correct this mistake (*programmer* should be to *woman* as *programmer* is to *man*), they developed algorithms to "de-bias" word embeddings—the vector representation of text—and thus provide a different configuration of words that would be considered less sexist.

Word embeddings are ways to organize words in space such that their proximity or distance to other words holds semantic information. However, an unwanted proximity or distance might be interpreted as bias by researchers and users alike (Noble, 2018; Bender et. al., 2021; Steyerl, 2023), and can be understood as a sense-making problem, in which a given semantic output does not correspond to the expectation. And yet, as Bolukbasi and their colleagues show, it is possible to reconfigure semantic fields such that they make more acceptable sense. This article investigates how word embeddings, as used in large language models (LLMs), are the result of *shaping processes*, and how these shaping processes are akin to educational processes.

We define shaping processes as the different steps in the development of a technical artefact, in order to modify both its function and user perceptions. This article focuses on two specific processes, benchmarking and reinforcement learning, to highlight the overall tendency in which such shaping processes inscribe themselves. As such, the central question we address is: *under which logic do shaping processes take place? How are technical processes implementing such logics in order to discover meaning-making capabilities in LLMs? And who determines the kind of sense that is being made by a large language model?* We hypothesize that these processes can be productively analyzed through the dual lens of *discipline* and *control*, as put forth, respectively, by Michel Foucault (Foucault, 1993) and Gilles Deleuze (Deleuze, 1992), particularly in their discussion of education; through this, we show that shaping logics, when it comes to generative cognitive technologies, influence the development and assessment of meaning-making abilities both in the machine and the human.

We begin by exploring how meaning can be encoded digitally by making the relationship between syntax and semantics in computer environments explicit. By comparing binary encoding and vector encoding, we highlight the complexities of the latter, particularly when assessing meaningfulness. We then

trace how those vectors are being shaped — that, is being rendered operationally meaningful — within LLMs. Specifically, we pay attention to two particular steps in the creation process of an LLM: benchmarking and reinforcement learning. We highlight how these techniques, a combination of discipline and control, contribute to normalization and standardization of meaning, but also from its modulation and adaptation, and result in semantic *subspaces*.

Discussing Alan Turing's proposal of machine intelligence as an educational problem, we conclude by turning to theories of co-construction of intelligence (Bachimont, 2004; Stiegler, 2010) to sketch out, through examples of linguistic normalization, hallucinations, and prompting, how such word embeddings can operate logics of control themselves.

1. From a bit to a vector

The question of discursive communication in technical systems is inseparable from the question of encoding. Whether as frequency-modulated hertzian waves, pixel arrays, or smoke clouds, different encodings enable different discourses (Postman, 1985). This section focuses on the shift from one encoding to the other and its semantic implications, looking at both the bit and the vector as a means to represent information in digital environments and highlighting how sense-making shifts from one to the other.

1.1 External reference in the bit

Before the electrification of computers, the use of binary distinction greatly facilitated automation, from the programming of textile patterns in jacquard looms to the processing of punch cards in census exercises (Ceruzzi, 2003). In the context of mechanical work, the binary sign's only significant property is that it has two mutually exclusive states; from these states, it becomes possible to encode representation (in the form of binary digits) and action (in the form of Boolean logic). Binary is entirely decontextualized, and it does not matter whether the binary sign is represented as a pair of 0/1, red/blue, low/high, cold/hot, as long as it is a disjointed pair[1].

While enabling flexible representation, this lack of context requires additional cognitive apparatuses, such as references and conventions against which a particular configuration of binary can be checked. Like all codes, there is a need for a cipher to access the meaning encoded in the binary representation (Kittler,

[1]In practice, the representation of binary digits as a pair of 0 and 1 is the most convenient.

2008). From 01001010 as input, the convention of 4-delimited base 2 encoding allows us to retrieve decimal numbers, here the number 74. Once such number has been decoded, we can further decode it into a letter, following here the reference table of the American Standard Code for Information Interchange (ASCII), in which case the number 74 will be interpreted as the upper case letter *J*. An equivalent for actions encoded in binary are *truth tables*, establishing the results of particular combinations of Boolean logic operations.

This decontextualized binary sign was contemporary with another decontextualization: that of the message. Claude Shannon's theory of communication famously proposed that meaning was irrelevant when calculating the means of communication and that one should, therefore, focus on maximally faithful recreation of the input signal, avoiding any kind of noise interference (understood as the corruption of the initial value of the transmitting medium) (Shannon, 1948). Encoding information through specific signs, whether Morse code or binary code, lent itself particularly well to this paradigm of information transmission. However, such a system holds a second assumption: it assumes the meaningfulness of the source. Indeed, in order to decode a message under Shannon's theory at all, one must presuppose there is sensical message to decode.

While binary encoding might be first seen as a decontextualized sign, as a technical object, it also exists in a network of relations, involving at least reference documents, transmission media and human agents that are all necessary for it be productively operationalized. Such productivity is achieved specifically by setting aside meaning to focus on syntax.

1.2 Internal reference in the vector

From the 1950s until the 2010s, the binary digit remained the dominant form of encoding information in digital systems. Throughout the 1970s, though, another form appeared, known as Vector Space Models (VSM). Originally proposed by Gerald Salton, this technique for information retrieval relied on the key insight, proposed by linguist John Firth in 1957 that "[we] shall know words by the company they keep" (Firth 12), hence departing from an essentialist view of language, towards a pragmatic one, in which the context of a given word should be part of its encoding (Salton et. al., 1975). Such encoding became particularly popular in broader digital information system after Yoshua Bengio and his team combined it with neural network algorithms at the dawn of the twentieth century (Cardon, 2018).

A vector is a mathematical entity that consists of a series of numbers grouped together to represent another entity. Often, vectors are associated with spatial operations: the entities they represent can be either a point or a direction. In computer science, vectors are used to represent entities known as features, measurable properties of an object (for instance, a human can be said to have features such as age, height, skin pigmentation, credit score, and political leaning). Today, such representations are at the core of contemporary machine learning models, allowing a new kind of *translation* between the world and the computer (Rieder, 2020).

In machine learning, a vector represents the current values of a given object, such that a human would have a value of 0 for the property “melting point”, while water would have a value of non-0 for the property “melting point”. Conversely, water would have a value of 0 for the property “gender”, while a human would have a non-0 value for that same property. However, this implies that each feature in this space is related to all the other dimensions of the space: a human could *potentially* have a non-0 value for the property “melting point”. Vectors are thus always containing the potential features of the whole space in which they exist and are more or less relatively tightly defined in terms of each other.

If binary enabled a *syntactic* exchange (everything can be encoded as a series of 0s and 1s), vectors enable a *semantic* exchange (everything can be described in terms of everything else). Combining vectors entails a more malleable manipulation of meaning throughout lexical fields. As a vector goes from *Berlin* to *Germany*, it represents the concept *capital city* (Guo et. al., 2023).

Because features exist in relation to one another, and meaning is constructed through the local similarity of vectors, semantic space both flexibly stores meaning (each number in a vector can subtly change without affecting overall meaning) and systematically retrieves it (all vectors exist in the same dimensions).

1.3 Expected meaning, unexpected meaning

The nature of meaning differs depending on encoding – but this is by not exclusive to digital inscription systems. For instance, Jack Goody’s work on lists and Bruno Latour’s on perspective, both suggest epistemological consequences inherent in the choice of one particular syntactic system over another (Goody, 1986; Latour, 2013). While binary encoding allows a translation between physical

phenomena and concepts, between electricity and numbers, and while Boolean logic facilitates the implementation of symbolic processing, vectors open up a new perspective on at least one particular level: the spatial dimension of their semantics.

The breadth of the data encoded, packaged in online corpora such as Common Crawl, is valuable insofar as it is mostly syntactically correct natural language. However, it does not follow that its recombination by way of large language model generation will be sensical because the source of such recombination cannot be attributed to a meaningful agent. The problem with language generation based on vector encoding is, therefore, that meaning is ontologically uncertain because it is statistical (software engineers tried to wrangle uncertainty out of the electrical circuits by forcing the continuous voltage into the discrete binary). Such uncertainty brings the acceptability of meaning into question — which can have either potentially boring or dramatic consequences. While binary encoding limits the acceptability of meaning to faithful signal reconstitution, vector encoding gives it a more complicated dimension.

Reconstituting meaning from binary encoding has always been a clearly defined problem, involving only mathematical reconstitution of the original message. Correctness of meaning, on the other hand, began as a computer-syntactic problem, but shifted with vectors to become a human-semantic problem.

2. Shaping vectors

We now turn our attention to techniques deployed by producers of LLMs to shape word embeddings of LLMs into models capable of meaningful output. After looking at the use of benchmarks for capability discovery, we argue that these processes operate as a form of discipline, as theorized by Michel Foucault. Then, we turn to reinforcement learning as an example of such shaping, but this time through the lens of a form of control, following Gilles Deleuze. We then conclude this section by reframing *training* in terms of *education*, drawing on Alan Turing's seminal paper, "Computing Machinery and Intelligence".

2.1 Benchmarks and the disciplining of vectors

Originally, a digitally encoded message was considered intelligible when it successfully compiled and behaved according to specification. But as

programming became an engineering discipline (Campbell-Kelly, 2003), engineers' focus on metrics, such as efficiency and reliability, ushered in new ways of qualifying the value of a program as a productive object. From the 1970s on, benchmarks emerged as reproducible tests to signal entities' comparative productivity. Through standardized procedures, they measure and rank, for instance, the time taken to sort lists of items, the number of triangles that can be drawn at a given frame rate, or the temperature of a CPU chip when processing a certain set of tasks.

Conventional engineering metrics, such as speed, play only a minor role in determining the quality of today's large language models. While contemporary benchmarks are still centered around the concept of performance, it is no longer measured on discrete machine tasks, but rather on subjective human ones -- focusing on content rather than form.

Engineering benchmarks for LLMs thus take on a different dimension, involving conceptual assessments, rather than technical efficiency. For instance, the General Language Understanding Evaluation (GLUE) (Wang et. al., 2019) benchmark is a test for machines that assesses performance in domains such as lexical semantics, predicate-argument structure, logic, as well as knowledge and common-sense. These tests have a normative power, deciding the extent to which something is correct or not, and are thus part of disciplinary technologies, i.e., technologies that rely on the creation, supervision, and maintenance of norms (Galloway, 2004). Here, benchmarks enable engineers and other users to determine the relative performance of one LLM compared with others.

The recent application of LLMs to other kinds of benchmarking tests, namely standardized tests designed for humans,suggests a parallel between the logic of benchmarking and that of education. In the past years, LLMs have successfully passed the Chartered Financial Analyst exam (I & II), the Bar exam, the SAT, the GRE , the Biology Olympiad Semifinal Exam, the Certified and Advanced Sommelier Exam, and the United States Medical Licensing Exam (Varanasi, 2023). As well as assessing LLMs' capabilities, such tests allow for the adjustment and regulation of cognitive processes, and act as value judgments for the meaningfulness of an output produced by an agent whose capabilities are to be asserted, whether human or machine-simulated. Referring to the educational system of the 20th century, Foucault writes:

> These 'regulated and concerted systems' fuse together the human capacity to manipulate words, things and people, adjusting

> abilities and inculcating behaviour via 'regulated communications' and 'power processes', and in the process structuring how teaching and learning take place. (218-219)

At the heart of the practice of teaching is a defined and regulated relation of surveillance that acts to improve the efficiency of its subject. The power process here is that of the standardized test, as it measures and compares decontextualized performance (Ryan, 1991). This happens through normalization, the shaping of entities in order to make them comparable and rankable, an operation already at play in engineering benchmarks (Heaven, 2023). One key difference, however, is that the discipline that Foucault describes in the school primarily aims at disciplining bodies, particularly in terms of sexuality, whereas the disciplining of vectors happens on the other side of the cartesian distinction which underpins mainstream artificial intelligence research.

Adherence to standardized benchmarks is not the only way that researchers shape acceptable meaning in LLMs. Once a certain kind of technical performance is confirmed, its social performance must also be assessed and eventually modified. To do that, there is a feedback mechanism, involving both negative and positive signals.

2.2 Reinforcement and the spaces of control

Word embeddings underpinning LLMs are malleable: LLMs can propose different semantic outputs based on the different weights and attentions (Guo et. al., 2023). A notorious example of such malleability is that of Microsoft's chatbot, Tay, who remodelled itself to generate more discriminatory and offensive content after just one day interacting with social media users (Glance, 2016). While benchmarks assess generic capabilities and output quantitative information about the performance of an LLM, they only assess acceptability on a factual and syntactical level, and not on a social or moral level. Additionally, as a commercial product, its outputs must comply with particular legal frameworks that specify what can and cannot be said. Beyond standardization, this then requires LLMS to adapt the semantic space they encoded to *ad hoc* requirements.

Such modulation happens through processes known as reinforcement learning, whether with human or AI feedback. Reinforcement learning judges each output against standards to support subsequent optimization. It involves having a trusted authority (such as a human who has been told what to expect from an ideal LLM output) provide feedback to the training model to reinforce certain

semantic features (e.g., preventing any output that is deemed discriminatory, copyright infringement, or harmful to the user) (Kaelbling, 1996).

While benchmarking focused on abstract comparability through normative testing, reinforcement learning involves more subjective normalization of meaning through feedback and iteration in order to align the model with what is considered a legally, morally, and socially acceptable meaning.

Such logic updates a disciplinary approach to undetermined behaviour and enters the realm of control. In his 1992 essay, Gilles Deleuze describes a new kind of era, ushered by a new kind of machines—computers—that would also suggest new mechanisms to govern individuals. This era of the society of control relies on adaptability, modulation, and deformation in order to best match the desired situation. Deleuze writes:

> [...] the different control mechanisms are inseparable variations, forming a system of variable geometry the language of which is *numerical* (which doesn't necessarily mean binary). Enclosures are *molds*, distinct castings, but controls are a *modulation*, like a self-deforming cast that will continuously change from one moment to the other, or like a sieve whose mesh will transmute from point to point. (3)

During reinforcement learning, the word embeddings of a LLM are shaped into a particular meaning through *ad hoc* interfaced actions such as "thumbs-up" or "thumbs-down", which are subsequently backpropagated through the weights of the network, slightly re-arranging embeddings into a semantic space whose landscape better matches the expectations of the judging entity. Furthermore, such a process can be conducted iteratively, blurring the distinction between what is in training and what has been trained—Deleuze identifies a similar change in the human educational process, wherein education is replaced by continuing education and the educated subject can become a uniquely shaped object -- an *objectile* (Savat, 2005). The objectile is the result of a unbounded modulation, rather than the singular structural shaping of a sculpture. Instead of the standard formatting of foucaldian educational institutions, Deleuze suggests the dawn of a new mode of education which involves personalized frames of action for each subject, an individualized, yet clearly controlled subject.

2.3 Educating intelligences

The question of education has been asked since the beginning of contemporary

history of AI research, considering that education was a crucial step in establishing the intelligence of a subject.

In Alan Turing's seminal "Computing Machinery and Intelligence" (1956), he concludes his investigation into whether machines can think by focusing on how to make them do so. Drawing parallels from the development of human cognition, he identifies three components: the initial conditions (genome for humans, model architecture for LLMs), the formal education (schooling for humans, training for LLMs), and epiphenomenal events (interactions for humans, reinforcements for LLMs).

Such formal education for LLMs stresses learning by example (Campolo, 2023) and capability discovery through human-context benchmarks, either in the form of specialized machine learning tests (e.g., GLUE, BLUE, LMSYS) or broader "real-world" knowledge tests (e.g., the SAT, MCAT, or LSAT).

However, the educational process within an institutional setting does not, as Foucault has shown, limit itself to the transfer of knowledge, but involves also the normalizing of bodies and minds. Since LLMs do not have a corporeal incarnation beyond matrices of weights written to files and globally-networked data and compute centers, it is on the "mind" of the LLM that the educational process of benchmarking and reinforcement learning operates.

Critically inspecting the two educational logics at play in the shaping of vectors—benchmarking as discipline, reinforcement learning as control—highlights two concerns. First, the harmonization of acceptability standards through benchmarks determines the narrow kinds of intelligence which can be expected when interacting with models (i.e. scholarly, academic, bookwormish, test-oriented, to the extent that some researchers have started to look into ways to prevent LLMs from cheating on tests (Zhou et. al., 2023)). For instance, as of 2024, LLMs tend to perform relatively poorly on non-verbal reasoning (Potter, 2024). Since the passing of those assessments operate as a sort of test, we can subsequently anticipate *the kind* of intelligence that those models display based on their assessment techniques. Second, the fine-tuning of acceptability through reinforcement learning takes a performing academic model resulting from the passing of benchmarks, and presents to the end-user a refined version with particular values embedded in them. Due to the limited amount of companies being able to deploy such reinforcement learning, these values then have similar consideration across the globe (Awad et. al., 2018). Not only is the factual intelligence standardized, but the values ascribed to those facts is equally

controlled.

3. Shaping users

Deleuze's conception of continuous education as the on-going shaping of intelligences and abilities implies that the structural distinction between what is inside the enclosure and what remains outside is blurry at best. According to the logic of control, the shaping of LLMs does not stop before release to the public. Continuous, user-provided feedback and software updates constantly re-shape their word embeddings (Gao, 2024). This last section investigates the potentially shifting positions of tested and tester once LLMs are deployed to -- and interacting with -- a broader audience.

3.1 Cognitive technologies and semantic spaces

We take the position here that all intelligence is, to a certain extent, artificial, insofar as it embedded in technical artifacts and symbol systems, as suggested by historians and philosophers of technics (Leroi-Gourhan, 2009; Stiegler, 2008 Bachimont, 2004). Technical apparatuses help us think through problems aided by the use of specific cognitive organizational devices, such as lists, tables, or formulas, as shown by Jack Goody on his work on graphical reason. While Goody interprets these techniques as a means of organizing representations of the world, Stiegler conceptualizes these technologies as tertiary retentions in which the memories of things and practices are externalized and reified into technical artefacts. In both cases, the technical written artefact is co-constructive of thought.

Digital technology is no exception. Its flagship artefact, the digital computer, exhibits properties such as modularity, translation, computation, connection, and simulation (Lev Manovich, 2001), cognitive operations that, by reorganizing the formalities of the concepts they manipulate, also change our understanding of these concepts (e.g.. digital technologies allow us, for the first time in the history of humankind, to copy a text without reading it). Electric-symbolic encoding of meaning thus has an influence on how we understand and make sense of the world.

Attending specifically to texts that exist first and foremost within a digital eco-system, such as websites, digital documents (either in plaintext or in formats such as .PDF, .DOCX, .ODT or .MD), or social media messaging, we can follow Alexandra

Saemmer to consider the *computext*, which is a kind of text that includes "both the algorithms operating weights and calculus on the traces left by the users, as well as the traces themselves, organised in databases" (Saemmer, 2020). Programming, considered as a technique providing the background for the dynamic evolution of meaning, already hints at the fact that software code is a writing of writing. Similarly, "computexts frame and guide the writing process; however, the user no longer writes *in* these tools, but literally writes *with* them" (Saemmer, 2020).

We understand technologies, whether physical or cognitive, to be points of integration in a broader environment and means of interaction within such environments (Hayles, 1999). In the case of LLMs, the environment is not just that of academic research, corporate investment, material infrastructure, raw datasets, and mainstream rhetorical discourses whose networked interaction have brought into being this specific technology, but also the (semantic) environment created *within* such kind of technology.

3.2 Subspaces and prompt engineering

The post-processing of lexical fields in computational systems has been thoroughly researched in the context of search engines (Sack, 2017), social media (Saemmer, 2020), and word-processing (Kirschenbaum, 2016). Nonetheless, the way vector-encoded LLMs affect our linguistic and discursive practices is still under developed, and we sketch out here some threads of how they might do so.

As LLMs retrieve information from their word embeddings, they navigate semantic spaces. However, such a retrieval of information is only useful if it is meaningful to us, the users; and in order to be meaningful, it navigates across vectors that are in close proximity to each other, focusing on re-configurable, (hyper-)local coherence to suggest meaningful structuring of content (i.e., guessing the next word that is the closest to the current word based on the path already travelled). The proximity (or distance) of vectors to each other is therefore essential to how the LLM output is perceived as intelligible to us. Meaning is no longer created through symbolic-logical combinations, but by spatial proximity in a specific semantic space. Because proximity of certain tokens involves distance to others, this implied process of exclusion can be described as a *subspace*, one in which some statements are more likely to be output than others.

To illustrate one of the features of such spatial organization of meaning, we can

pay attention to the phenomenon of so-called “hallucinations”, textual or visual propositions that are considered by the user to be inacceptable with respect to the “ground truth” (the concept in machine learning referring to the base of facts from which reasoning should start). This occurs whenever LLMs suggest something that is considered slightly too remote from such truth, or reality, and yet still adjacent to it. The hallucination is an approximation, in the sense that it is only a proximity to the syntactic configuration that would yield a semantic load grounded in reality. User interactions with hallucinating models thus redraws the line between fact and fiction, as text becomes *a version of* itself, moving from mechanical print to quantum spatialization. While the content seems realistic, and its syntax may well be semantically correct and convincing, the trust users have in the output of the system can only be superficial (Förster, 2023).

Second, the restitution of training data and processes contributes to highlighting (or hiding) particular pieces of information. Models trained by corporations that are particularly attuned to a restrictive notion of copyright (e.g., Google, Microsoft, OpenAI) prevent any replication of styles or creations by artists (or their descendants) who might be able to initiate a lawsuit. LLMs are also prevented from, for instance, providing any expression of personal preference. Previous models based on reinforcement learning, like Microsoft’s Tay, have shown that they are not restrained by contextual social cues such as moral and legal standards. No longer treating text as a value-less mass, such examples of socially-embedded models, insofar as they are consumer products, are explicitly refusing to enter certain semantic spaces. Here, the reinforcement learning’s disciplining of embeddings is made clear, with LLMs beginning answers to inappropriate or unacceptable queries with “as a large language model, …”, which explicitly enacts a techno-political framing akin to a political aesthetics, in which what is visible and what is hidden are determined by their political nature (Rancière, 2000).

Things become somewhat murkier when the LLM does not acknowledge this shaping of the semantic space. In the case of the image generation model RuDall-E, developed and trained by Russian software engineers, it is impossible to prompt the model to generating images of a pro-European revolution in Ukraine or any visual references to the on-going war in Ukraine (Dubow, 2023). Here, it is no longer merely forbidden to be express these outputs in a straightforward manner, but rather pre-emptively forcelosed. We can qualify these different shapings, some through reinforcement learning, some through initial training data, as the creation of *subspaces*, specific configurations of word embeddings,

one in which attention is forced towards particular centers of gravity. The emerging practice of “prompt engineering” consists of providing LLMs with an initial semantic configuration through written instructions. This “prompt engineering” can be conceived as explicitly directing the LLM’s attention towards specific subspaces (e.g., providing a prompt like “Drawing on your expertise as a…” or “You are great pedagogue. Explain to me…”). In this case, end-users harness the malleability of subspaces by deploying technologically-adapted language to shape the navigable space of embeddings into a configuration that best meets their needs. However, prompts can also be entered at the system level, either by the technology company itself, a corporate re-brander of a white-label system, or even by individual power users on local machines. These system prompts exert another shaping of the semantic space that occurs in-between the user’s final prompt and the model’s ultimate output,. Such a practice means that institutions or superusers are using access to the model re-orient answers, and whose experience in training it grounds their perceived ability to decide on the semantic subspaces from which a linear answer should be drawn. End uses assume that an LLM draws on all of its training to produce an answer, and yet it only operates on a partly visible subset.

Conclusion

Vector embeddings as a new form of encoding enables new ways of shaping the content of language. Particularly, they add a layer of self-reference to digitally-encoded language (since words and tokens make sense in the context of other words and tokens) and of uncertainty (since the origin of a given output is no longer a given in the process of decoding meaning). In order to reconstruct semantics from syntax generation, two main processes are involved in the *shaping of semantic spaces*.

We have shown how this shaping operates through two logics. The disciplinary logic, in a slide from engineering benchmarks towards educational benchmark, uses external standards to assess the productive performance of the language models. Such a disciplining process takes on modular features through the controlling process of reinforcement learning. By providing feedback and examples to reach a configuration that yields acceptable outputs. The control logic, drawing on the malleability of software, uses fine-tuned continuous adjustments to validate what is acceptable or not at a value-level. Both of these logics are akin to how standardised test in human education establish normalized knowledge practices, and how continuous education ensures a new kind of

framing in computer-powered societies. Ultimately, these processes ultimately narrow down the frame of expressivity and semantic combination of the LLM.

Finally, we sketched out how such combination of discipline and control in shaping word embeddings can affect users, by suggesting that linguistic interaction only takes place in semantic subspaces. Through dialogue, the user probes the spatial configurations of meaning, but the exact topology of these configurations nonetheless remains elusive, and can thus impact what can be said, and – for the first time in the era of computation -- even what can be imagined.

References

Awad, Edmond, et al. “The Moral Machine Experiment.” *Nature*, vol. 563, no. 7729, Nov. 2018, pp. 59–64. *www.nature.com*, https://doi.org/10.1038/s41586-018-0637-6.

Bachimont, Bruno. “Signes Formels et Computation Numérique : Entre Intuition et Formalisme: Critique de La Raison Computationnelle.” *Instrumente in Kunst Und Wissenschaft - Zur Architektonik Kultureller Grenzen Im 17. Jahrhundert*, edited by H Schramm et al., Walter de Gruyter Verlag, 2004.

Bender, Emily M., et al. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜.” *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ACM, 2021, pp. 610–23. DOI.org (Crossref), https://doi.org/10.1145/3442188.3445922.

Bolukbasi, Tolga, et al. Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings. arXiv:1607.06520, arXiv, 21 July 2016. arXiv.org, https://doi.org/10.48550/arXiv.1607.06520.

Bowman, Sam. *Intelligence Testing—Asterisk*. https://asteriskmag.com/issues/04/intelligence-testing. Accessed 2 Mar. 2024.

Campbell-Kelly, Martin. *From Airline Reservations to Sonic the Hedgehog : A History of the Software Industry*. Cambridge, Mass. : MIT Press, 2003. Internet Archive, http://archive.org/details/fromairlinereser00mart_0.

Campolo, Alexander, and Katia Schwerzmann. “From Rules to Examples: Machine Learning’s Type of Authority.” *Big Data & Society*, vol. 10, no. 2, July 2023, p. 20539517231188725. SAGE Journals, https://doi.org/10.1177/20539517231188725.

Cardon, Dominique, et al. “Neurons Spike Back: The Invention of Inductive Machines and the Artificial Intelligence Controversy.” *Réseaux*, vol. 36, no. 211,

2018. https://mazieres.gitlab.io/neurons-spike-back/index.htm.

Ceruzzi, Paul E. *A History of Modern Computing*. 2nd ed., MIT Press, 2003.

Deleuze, Gilles. “Postscript on the Societies of Control.” *October*, vol. 59, 1992, pp. 3–7.

Dubow, Ben. “Why Putin’s Faith in Russia’s ‘Homegrown Midjourney’ Is Misplaced.” The Moscow Times, 27 Dec. 2023, https://www.themoscowtimes.com/2023/12/27/why-putins-faith-in-russias-homegrown-midjourney-is-misplaced-a83577.

Firth, John Rupert. “A Synopsis of Linguistic Theory, 1930-1955.” *Studies in Linguistic Analysis*, Blackwell, 1957, pp.1-32. Cambridge University Press.

Foucault, Michel. *Surveiller et punir*. Gallimard, 1993, https://www.cairn.info/surveiller-et-punir--9782070729685.htm. Cairn.info.

Galloway, Alexander R. *Protocol: How Control Exists after Decentralization*. MIT Press, 2004.

Gao, Yunfan, et al. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997, arXiv, 27 Mar. 2024. arXiv.org, https://doi.org/10.48550/arXiv.2312.10997.

Gewirtz, Paul. “On ‘I Know It When I See It.’” *Yale Law Journal*, Jan. 1996. openyls.law.yale.edu, https://openyls.law.yale.edu/handle/20.500.13051/8935.

Glance, David. “Microsoft’s Racist Chatbot Tay Highlights How Far AI Is from Being Truly Intelligent.” *The Conversation*, 27 Mar. 2016, http://theconversation.com/microsofts-racist-chatbot-tay-highlights-how-far-ai-is-from-being-truly-intelligent-56881.

Goody, Jack. *The Logic of Writing and the Organization of Society*. Cambridge University Press, 1986, https://doi.org/10.1017/CBO9780511621598.

Guo, Zishan, et al. Evaluating Large Language Models: A Comprehensive Survey. arXiv:2310.19736, arXiv, 25 Nov. 2023. arXiv.org, https://doi.org/10.48550/arXiv.2310.19736.

Hayles, N. Katherine. *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. University of Chicago Press, 1999. University of Chicago Press, https://press.uchicago.edu/ucp/books/book/chicago/H/bo3769963.html.

Heaven, Will Douglas. “AI Hype Is Built on High Test Scores. Those Tests Are Flawed.” *MIT Technology Review*, 2023,

https://www.technologyreview.com/2023/08/30/1078670/large-language-models-arent-people-lets-stop-testing-them-like-they-were/.

Kaelbling, L. P., et al. “Reinforcement Learning: A Survey.” Journal of Artificial Intelligence Research, vol. 4, May 1996, pp. 237–85. www.jair.org, https://doi.org/10.1613/jair.301.

Kirschenbaum, Matthew G. Track Changes: A Literary History of Word Processing. Harvard University Press, 2016. www.degruyter.com, https://doi.org/10.4159/9780674969469.

Kittler, Friedrich. “Code (or, How You Can Write Something Differently).” *Software Studies: A Lexicon*, edited by Matthew Fuller, The MIT Press, 2008, p. 0. Silverchair, https://doi.org/10.7551/mitpress/9780262062749.003.0006.

Latour, Bruno. “« Les ‘vues’ de l’esprit ». Une introduction à l’anthropologie des sciences et des techniques.” *Sociologie de la traduction : Textes fondateurs*, edited by Madeleine Akrich and Michel Callon, Presses des Mines, 2013, pp. 33–69. OpenEdition Books, https://doi.org/10.4000/books.pressesmines.1191.

Leroi-Gourhan, André. *Le Geste et la Parole - tome 1: Technique et langage*. Albin Michel, 2009.

Manovich, Lev. *The Language of New Media*. MIT Press, 2001. /z-wcorg/.

Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018.

Potter, Brian. Could ChatGPT Become an Architect? 4 Mar. 2024, https://www.construction-physics.com/p/could-chatgpt-become-an-architect.

Postman, Neil. *Amusing Ourselves to Death: Public Discourse in the Age of Show Business*. 1st edition, Viking Penguin, 1985.

Rancière, Jacques. “1. Du partage du sensible et des rapports qu’il établit entre politique et esthétique.” *Le partage du sensible*, La Fabrique Éditions, 2000, pp. 12–25. Cairn.info, https://www.cairn.info/le-partage-du-sensible--9782913372054-p-12.htm.

Rieder, Bernhard. “From Frequencies to Vectors.” Engines of Order, Amsterdam University Press, 2020, pp. 199–234. JSTOR, https://doi.org/10.2307/j.ctv12sdvf1.9.

Ryan, James. “Observing and Normalizing: Foucault, Discipline, and Inequality in Schooling: BIG BROTHER IS WATCHING YOU.” *The Journal of Educational Thought (JET) / Revue de La Pensée Éducative*, vol. 25, no. 2, 1991, pp. 104–19.

Sack, Warren. “Out of Bounds: Language Limits, Language Planning, and the

Definition of Distance in the New Spaces of Linguistic Capitalism." *Computational Culture*, no. 6, Nov. 2017. computationalculture.net, http://computationalculture.net/out-of-bounds-language-limits-language-planning-and-the-definition-of-distance-in-the-new-spaces-of-linguistic-capitalism/.

Saemmer, Alexandra. "From the architext to the computext. Poetics of the digital text, facing the evolution of devices." Communication langages, vol. 203, no. 1, Apr. 2020, pp. 99–114.

Salton, Gerald., et al. "A Vector Space Model for Automatic Indexing." *Communications of the ACM*, vol. 18, no. 11, Nov. 1975, pp. 613–20. ACM Digital Library, https://doi.org/10.1145/361219.361220.

Savat, David. "Deleuze's Objectile: From Discipline to Modulation." *Deleuze and New Technology*, edited by David Savat and Mark Poster, Edinburgh University Press, 2005, p. 0. Silverchair, https://doi.org/10.3366/edinburgh/9780748633364.003.0004.

Shannon, C. E. "A Mathematical Theory of Communication." *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, Jan. 2001, pp. 3–55. Semantic Scholar, https://doi.org/10.1145/584091.584093.

Steyerl, Hito. "Mean Images." *New Left Review*, no. 140/141, Apr. 2023, pp. 82–97.

Turing, Alan M. "Computing Machinery and Intelligence." *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, edited by Robert Epstein et al., Springer Netherlands, 2009, pp. 23–65. Springer Link, https://doi.org/10.1007/978-1-4020-6710-5_3.

Varanasi, Lakshmi. "GPT-4 Can Ace the Bar, but It Only Has a Decent Chance of Passing the CFA Exams. Here's a List of Difficult Exams the ChatGPT and GPT-4 Have Passed." *Business Insider*, 5 Nov. 2023, https://www.businessinsider.com/list-here-are-the-exams-chatgpt-has-passed-so-far-2023-1.

Wang, Alex, et al. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. arXiv:1804.07461, arXiv, 22 Feb. 2019. arXiv.org, https://doi.org/10.48550/arXiv.1804.07461.