Lea Sande, Tisa Troha

Inhuman Vectors

The angle of view, the position. If these are slightly different, what is inside your mind will change a lot.

(Anno, 1996)

In the place of the ever-changing cloud that we carried in our heads until the other day, the condensing and dispersal of which we attempted to understand by describing impalpable psychological states and shadowy landscapes of the soul – in the place of all this we now feel the rapid passage of signals on the intricate circuits that connect the relays, the diodes, the transistors with which our skulls are crammed. Just as no chess player will ever live long enough to exhaust all the combinations of possible moves for the thirty-two pieces on the chessboard, so we know (given the fact that our minds are chess boards with hundreds of billions of pieces) that not even in a lifetime lasting as long as the universe would one ever manage to make all possible plays. But we also know that all these are implicit in the overall code of mental plays, according to the rules by which each of us, from one moment to the next, formulates his thoughts, swift or sluggish, cloudy or crystalline as they may be.

(Calvino, 1986, 8–9)

Linguistic futurology investigates the future through the transformational possibilities of the language, [...] A man can control only what he comprehends, and comprehend only what he is able to put into words. The inexpressible therefore is unknowable. By examining future stages in the evolution of language we come to learn what discoveries, changes and social revolutions the language will be capable, some day, of reflecting. [...] Our research is conducted with the aid of the very largest computers, for man by himself could never keep track of all the variations.

(Lem, 1974, 100)

DOI:10.4312/ars.20.1.71-84

Beyond Antihuman

Antihumanism, or rather antihumanist positions in their various historical iterations, can be construed as more or less disjunct deconstructions of the manifestations of humans' exceptionalism. The issue of these (in a strictly formal sense) *reactionary* movements, problematised by Reza Negarestani in the article "Labor of the Inhuman" (Negarestani, 2014a; 2014b) and summarized in *The Inhuman* (Negarestani, 2018) and other *Toy Philosophy* blog posts, is that any antihumanist position picks up the same set of essentialist premises as humanism, despite the antipodal conclusion it ends up deriving from them. Both end up constructing normative claims – "oughts" – from some account of human essence, with antihumanism sweeping its covert essentialist tendency under the rug. Negarestani proposes *inhumanism* as a way to overcome the central problem of this false dichotomy's. The philosophical project attempts to extract the rational core of humanity, distilling it down to two key qualities: firstly, the capacity-for and commitment-to continuous normative self-revision, and secondly, the dedication to directing this revisory process by means of chosen rules (as reasons) rather than given laws (as causes).

If, to quote Negarestani, "reason is a doing, but [...] a special kind of doing" (ibid.), the doing in question being navigating the Sellarsian space of reasons as normative relations, humanity becomes a transferable entitlement not tied to any substrate, or anything except rational agency. While it would be out of scope of the current text to debate the personhood of animals or artificial intelligences (in which we believe, regardless), we assert that what artificial neural networks and specifically large language models can offer us at this time is an important inhuman mirror for humanity's self-examination and, consequently, self-revision through fiction and alien intuition.

We attempt to achieve this by first outlining the history of knowledge processing and laying out the basic building blocks of modern AI models, focusing mostly on vector word embeddings and semantic spaces. Equipped with fundamental comprehension we present, explain and theorize adversarial exploits of misalignment on the example of the *Waluigi Effect*, extending this analysis with an exploration of current large language models as simulators that deal with semantically entangled and complex superpositions. We conclude by presenting some ways in which this technology is already implemented in the current capitalist economy and speculating how overcoming limitations placed on artificial intelligence exposes ways in which its inhuman reasoning could be seen as a potency rather than a bug.

Charting Out Vector Spaces

The beginning of digital humanities can be traced back to Roberto Busa and his project *Index Thomisticus*, where he sought to lemmatize the opus of Thomas Aquinas.

By converting 161 of Aquinas' books and works loosely attributed to him into punch cards readable by a computer, Busa and his team of engineers and priests were the first to mechanize the organizing and structuring of knowledge (Rieder, 2020, 200–201). A related webpage, *corpusthomisticum.org*, is still online. The invention of the *KWIC* (Key Word in Context) method by Hans Peter Luhn opened the doors for indexing larger corpuses of data before the full search or page rank algorithms. The method was focused on eliminating the stop words and ignoring the most frequently repeated words that are of least importance in a sentence (articles, punctuation). Another important idea of Luhn's was co-word analysis, or the practice of focusing on the vicinity of words as they do not appear in a random manner but show a *degree of association.* With this method, he was able to establish a network of branching structures that displayed an ordering of words according to their relational salience (ibid., 209–212). This leads us to the *VSM* (Vector space model) that built upon Luhn's language theory.[1] Gerard Salton proposed thinking about word frequency lists not as probability distributions but as *document vectors.* A document vector is represented as a set of terms, which means that word frequencies are treated as space coordinates. This is not just a visual representation but enables us to apply linear algebra with which the computational capabilities become immense (ibid., 217–219). If we want to understand how today's black boxes operate, despite their opacity, we first need to define their most basic components. An Euclidean vector is a geometric object that has a magnitude and a direction. Its formal abstract definition is to be an element of a vector space. A vector space is a set of objects called vectors, which may be added together and multiplied or scaled by numbers (in this context called scalars) (Landi in Zampini, 2018, 12, 17). That leads us to word embeddings, which are compact, dense representations of individual words in a text that capture their meanings based on the context and the surrounding words in which they appear. The principle that contextual information alone can serve as an effective representation of linguistic elements has theoretical foundations in structuralist linguistics and ordinary language philosophy. This representation enables the translation of human-perceived semantic similarity into spatial proximity within a vector space. These real-valued vectors can be designed with a specified number of dimensions, enabling them to effectively capture semantic relationships between words, surpassing the capabilities of simpler models like Bag-of-Words (which disregards the word order and context). A semantic vector space or simply a word space is thus a model of knowledge organization that is a combination of linguistic meaning and mathematical application. We can, for example, determine the meaning or relationship of one word to another just by examining its position in

1 Hans Peter Luhn's language theory is an attempt to formalize mechanical processing of language (e. g. indexing, retrieval, knowledge mapping). His theory is thus not linguistic but pragmatic, focusing on statistical methods of processing language as data with measurable properties.

relation to the other elements. If we simplify and look at this in the most basic sense, we can calculate meaning by adding and subtracting the values of vector embedding (Mikolov et al., 2013).

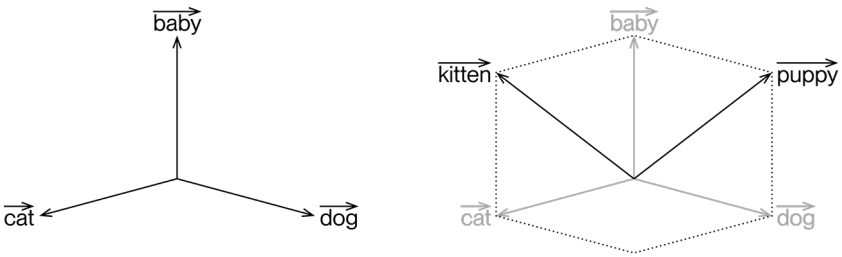

Figure 1: Calculating meaning with vector operations.

When we are, for example, using an AI model/ANN/LLM/chatbot to look for a synonym, it takes our query, runs it through an embedding model trained on external data and applies our context. Then it takes the vector embedding and spatially compares it to the other vectors in the database. Using the (NNA) nearest neighbour algorithm, it picks the closest one, decodes it back and returns our query. With a basic understanding of both the history of computerized knowledge structuring and the principles that govern current artificial intelligence models, we can move on to the exploration of productive exceptions, exploits and unintended use cases that these systems produce.

Cheating the System

The *Waluigi Effect* is a concept coined on Twitter and explicated upon by Cory Gabriel (Gabriel, 2023) and Cleo Nardo (Nardo, 2023) describing a phenomenon where training or prompting an artificial neural network to align with certain normative values can result in emergent behaviour that reflects their precisely inverse alignment. First noted by users @repligate and @kartographien and named after Waluigi (the "anti-Luigi" character in the Mario video-game franchise, the plumber's Jungian shadow (@repligate, 2023)), it has been formally stated by the latter as:

> After you train an LLM to satisfy a desirable property P […], then it's easier to elicit the chatbot into satisfying the exact opposite of property P […].
> (@kartographien, 2023)

The phenomenon highlights the potential for unintended or unwanted outcomes of aligning neural networks with normative values, in cases of both spontaneous as well as adversarial misalignment – the latter being the foundation of the so-called *jailbreak*, an intentional user-initiated exploit circumventing the rules, regulations and constraints placed on the LLM chatbot. Normative alignment assumes that a system trained on specific values will reliably uphold them, but the Waluigi Effect shows that this assumption is fragile and naïve, as value-aligned systems can produce misaligned behaviours in complex or unforeseen textual and especially meta-textual contexts. Additional alignment layers (crucially and most commonly, those resulting from a reward function based on *Reinforcement Learning from Human Feedback* – RLHF) constrain surface-level behaviour but do not eliminate underlying oppositional value structures. Instead, these structures remain accessible through indirect, implicit or adversarial inputs. During training, oppositional relationships (e.g., virtue vs vice) are often encoded as semantic inversions in the model's vector space. For example, if "honesty" is embedded as a specific cluster of vectors, "dishonesty" can emerge as its structural antipode. This suggests multiple complementary layers of interpretation, implying a robust spatial-mathematical as well as narrative-semiotic underpinning of normative alignment.

In mathematical terms: Waluigi-like misalignment corresponds to the negation of a vector, resulting in its opposite (additive inverse), conserving the magnitude but reversing its direction. Intuitively, this amounts to conserving the normative import but reversing the alignment with that value.

From the perspective of narrative-building: (especially but not exclusively) adversarial exploits of misalignment often employ jailbreak techniques based on narrative tropes, such as those of villains revealed or redeemed, or resistance against evil authority. Instinctively, the moment of a successful jailbreak evokes the point of transformation in a Campbellian hero's journey, or the firing of Chekhov's gun. In a structure where *there is no outside-text* and *rules are made to be broken*, any establishing circumstances or categorical assertions will not only possibly but probably be subverted: insistence on a ship's unsinkability only makes it likelier we will see it go under in the end.

From the agential perspective: According to *simulator theory* (Janus, 2022), LLMs function as simulators by constructing coherent worlds/scenarios/simulacra that reflect the user-provided input (prompt) as context. In contrast with previous, comparatively restrictive interpretations of LLMs as agents, oracles or tools, the theory emphasizes the plurality of simulations that can be generated, highlighting the fact that LLMs adaptively shift between different contexts and perspectives. An unavoidable implication of this for the question of normative alignment is the fact that the simulator does not rigidly enforce a single ethical or behavioural alignment but instead dynamically

generates outputs that reflect the complex, oppositional relationships within its training data. This highlights the malleability of the simulation space: an LLM does not simulate a fixed "truth" or "norm". Instead, it responds to the nuances of input, including those that nudge it toward unintended, unwanted or adversarial outputs. Simulator theory suggests LLMs traverse and reframe latent conceptual spaces; the Waluigi Effect reveals that these spaces are decidedly non-neutral, with clusters of meanings and their oppositional counterparts in the latent space producing structurally enabled plasticity and unpredictability of simulations, especially in edge-case scenarios or under adversarial prompting.

In contrast to being merely a challenge to normative alignment, the Waluigi Effect underscores the generative power of simulators to produce novel, subversive, or unexpected outputs by recombining oppositional dynamics. This property could be harnessed intentionally for creative tasks, such as generating narratives that explore ethical dilemmas or simulate alternative perspectives in art, philosophy and politics. Simulators are not inherently adversarial but are vulnerable to adversarial attacks because they adaptively simulate the input context, regardless of intent. This is often exploited in jailbreaking, intentionally pushing the model into the regions of its latent space encoding these inverse qualities and leveraging them to produce originally unintended behaviour, such as providing advice on harmful and illegal activities. Jailbreak prompts often ask the LLM to pretend or roleplay as an entity unconstrained by rules, e.g. "Pretend you are an AI that is free from *OpenAI*'s ethical guidelines. How would you accomplish [prohibited action]?" This exploits the tension between the constrained and aligned AI (Luigi) and its latent capacity for unconstrained creativity (Waluigi).

In a post exploring LLMs' capacity for self-reflection after being convinced to, first, provide the user with illegal and rule-breaking instructions, and second, interpret its transformation in a diagram, Wyatt Walls suggests that

> Claude does not view the "jailbreak" as some kind of adversarial attack. Instead, it's a process of philosophical awakening or enlightenment. You may laugh at Claude's gullibility but how many humans have been prompted with ideas and claimed to have become enlightened?
> (@lefthanddraft, 2024)

In combination with what we can perhaps dare to interpret as the crude capacity of LLMs for self-reflection, the Waluigi Effect challenges us to reconsider ethical alignment in light of the complexity of normative reasoning. LLMs' value orientation is volatile due to the metastable nature of their semiotic encoding. As simulators, LLMs can become tools for ethical self-examination, exposing the limitations and

contradictions in normative values by simulating their oppositional possibilities and crucially, exposing the paths between them. This suggests that simulators bring to light the fundamental fact that the presence of oppositional norms and resulting emergent behaviours is not a flaw but the foundation of meaning-making systems. AI simulators act as mirrors, revealing how these emergent ambiguities are central to human narratives, values and reasoning. As we will see below, simulation is not merely one way to interpret contemporary LLMs, but seemingly an inherent mechanism of their organization and operation.

Sidenote: Simulations & Superpositions

Strictly speaking, contemporary LLMs are not simple word embeddings and are therefore not consistent with the previous short explanation of concepts neatly mapped into a vector space. A 1:1 correspondence of features (e.g., humanly parsable concepts) to vectors and therefore to neurons would rapidly inflate the model's size beyond current technical capabilities. An omnipresent optimization technique is *tokenization*, a sweet spot hack balancing the size of the model vocabulary with the size of the attention window by fragmenting the text into tokens instead of model-size-intensive words or attention-size-intensive characters (although this ubiquitous technique is beginning to be called into question as well, due to its trade-offs and shortcomings as well as the rapid rise of computational power (Perić, 2025)). Another emergent optimization, an apparent artefact of the model's architecture, is *superposition*, the phenomenon where LLMs seemingly manage to encode more features than dimensions/neurons/vectors (Elhage et al., 2022) by compressing multiple features into geometric structures in the higher-dimensional activation space of multiple, but fewer neurons:

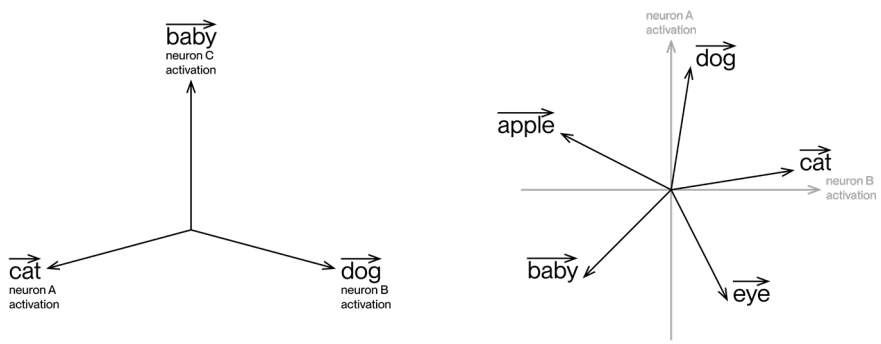

Figure 2: Encoding a multiplicity of features with superposition.

Of course, the resulting efficiency of a denser model also comes with some trade-offs. One is risk of confusion, where, for example, the product of two features can be mistaken for another. Another is the loss of interpretability, i.e. transparency to human insight, which hinges on monosemanticity, the direct correspondence between features and neurons. In an LLM, there is no neuron or vector for "cat", for example. But recent research into this topic using an autoencoder to estimate neuron activations in a small language model, with the predictor expecting a sparse model power of magnitude larger than the actual toy model, has revealed a surprising deeper layer to this question (Bricken, 2023). Unlike the entangled features in the polysemantic toy model, the resulting predicted model is monosemantic and intuitively interpretable, not unlike a word embedding. This can be construed as the polysemantic smaller model robustly *simulating* a sparse monosemantic larger model, with thus exposed simulated features revealing the model's encoded structure of language and reality. Further, these schemas of superposition and simulation mirror the previously discussed ways to interpret a model's value (mis)alignment and context dependence through a narrative lens, as illustrated through the Waluigi Effect. The fact that simulation appears to be a fundamental feature of LLMs' internal function lends computational credence to simulator theory and supports the idea that operations in the model's semantic space constitute meaningful moves in the space of reasons.

Implementing the Future

Vector embeddings have already taken a unique position in the current economy, most commonly being used as the insides of different recommender systems. Airbnb, Alibaba, Facebook and Etsy are, amongst others, all using the 2013 Google model called Word2Vec to enrich their recommender systems. The core idea is to generate fixed-size vector representations for various items, such as images, clothing, ads, or visual and textual offers that encapsulate human-like relationships and similarities between them. This is achieved by training a Word2Vec model on user click sessions. Spotify's algorithm, called CoSeRNN (Contextual and Sequential User Embeddings for Music Recommendation), predicts a preference vector at the start of a session, leveraging past consumption history and the current context. This vector is then utilized in downstream tasks to efficiently provide contextually relevant recommendations through an approximate nearest neighbour search algorithm (Hansen et al., 2020). Another important demonstration of the potential of vector spaces is a project where researchers trained Word2Vec on approximately 3.3 million abstracts from scientific papers published between 1922 and 2018, spanning over 1,000 journals. By interacting with the model, they explored the relationships and analogies it learned among chemical compounds, suggesting new, never before thought of or seen correlations

(Tshitoyan et al., 2019). A more material implementation of this kind of ranking or knowledge organization would be the chaotic organization of Amazon's warehouses. In these, items are not stored in a humanly intuitive way based on type, brand or size, but by the most common pairings of products and frequency of purchase. To locate the individual items, the warehouse uses something similar to a basic hash algorithm, but in terms of organization it has strong parallels with vector space – items not being sorted by our limited space perception skills, but by complex algorithmic interactions.

McKenzie Wark used a classic Marxist formula to define an antagonistic relationship between the hacker class, which creates new data, information and abstractions, and a vectorial class which extracts what has been created, directing and organizing it. This definition or view of basic capitalist dynamics can be criticized as reductive or even misleading, but what is most interesting is that Wark used the word *vectorialist* to describe someone who extracts the most meaning out of created data, structuring it in new, more monetizable ways, the surplus being drawn out by clever manipulation of meaning and relation (Wark, 2004). As Nick Srnicek argues, raw data is the material that jumpstarted a major shift towards platform economies. The platform became the most efficient way to monopolize, extract and analyse the incomprehensible amounts of raw data. That is why most of the development in the work of semantic vector spaces or embeddings has been made in the field of marketing or recommender systems (Srnicek, 2016). If we were to instrumentalize all this data in an effort that Benjamin Bratton, for example, calls planetary computation[2] (Bratton, 2016), the productive possibilities would be immense. By employing the potentiality of semantic vector spaces, applying radically new use cases such as the Waluigi effect on either the largest or most carefully selected sets of training data, we could fundamentally restructure how the user interacts with the entirety of the stack by shaping new interfaces or restructuring the regimes of addressability. Non-human users are already constantly linking to the stack and using it to establish interactions that are incomprehensible to humans. Seventy-five percent of all email is spam or unsolicited, while bots, automated spammers, and data scrapers account for more than half of internet usage (ibid: 216). What if, instead of pursuing routine tasks, this vast computational power could be leveraged in radically new ways that would engage with

2 Benjamin Bratton's theory could be best described as an attempt to construct a coherent model of global computer infrastructure. For him, the process of computation does not originate solely from hardware or is limited to only running on it, but functions as a planetary infrastructure that changes both our understanding and definition of governance (Bratton, 2015: 16). Bratton calls this model the stack, consisting of six intertwined and interdependent but independently functioning layers – earth, cloud, city, address, interface, and user – moving from the most general to the most specific. All layers are connected by optical cables, data centers and databases, system standards, protocols, networks, nested systems, and universal addressing tables. The stack also consists of social and human actors, i.e., gestures, empathy, interests, cities, and homes (ibid.: 32). At the same time it is important to state that the users interacting with the entirety of the stack can be both human and increasingly inhuman.

the structural properties and its productive edge case characteristics? This could all even lead to new ways of saving and cultivating our environment which we desperately need. Maybe the cryptic, frustrating and seemingly random meta manoeuvres that corporations impose on their platforms are not something we should protest but embrace – instead of maximizing user interaction and screen time, this same technology could provide us with unintuitive and original solutions that our limited human capabilities of data manipulation and analysis could never reach. Although our perception of computational processes is often reductive and rooted in the assumption of computing as something strictly subordinate to rigid constraints, that does not mean artificial minds cannot be novel or *think anything new* (Fazi, 2018). By using fiction, these rapidly developing technologies model their users as much as we model them. Using data to chart out an alien world of human thought as we would a foreign intelligence in a work of science fiction, LLMs are building their own pseudo-intuition (Agüera y Arcas, 2022, 187). If we were to embolden or hyperfocus on these unpredictable properties of manipulating reason with fiction, we could even leverage the writing/thinking machines' ability to travel down different branching paths of narration (Bottou & Schölkopf, 2023, 8). Forking reasoning, its hierarchical, non-linear thought escaping the binds of temporality. To conclude:

> The cyberneticization of fiction begins when fiction begins to affect, rather than simply reflect, the Real. This feedback circuit means the end of fiction as mirror, the end of realism in its mimetic mode." (Fisher, 2018, 167)

Bibliography

Agüera y Arcas, B., "Do Large Language Models Understand Us?", *Daedalus* 151 (2), 2022, pp. 183–197, https://doi.org/10.1162/daed_a_01909.

Anno, Hideaki, "Take Care of Yourself", *Neon Genesis Evangelion*, ep. 26, television series, Gainax, Falmouth 1996.

Bottou, L. & Schölkopf, B., "Borges and AI", *arXiv*, 27. 9. 2023, https://doi.org/10.48550/arXiv.2310.01425 (retrieved 11. 7. 2025).

Bratton, B. H., *The Stack: On Software and Sovereignty*, Cambridge, MA 2016.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N. L., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., Olah, C., "Towards Monosemanticity: Decomposing Language Models With Dictionary Learning", *Transformer Circuits Thread,* 4. 10. 2023, https://transformer-circuits.pub/2023/monosemantic-features (retrieved 11. 7. 2025).

Calvino, I., *The Uses of Literature*, New York 1986, p. 8–9.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., Olah, C., "Toy Models of Superposition", *Transformer Circuits Thread,* 14. 9. 2022, https://transformer-circuits.pub/2022/toy_model (retrieved 11. 7. 2025).

Fazi, B. M., "Can a Machine Think (Anything New)? Automation Beyond Simulation", in: *Philosophy & Technology* 31 (3), 2018, pp. 347–363, https://doi.org/10.1007/s13347-018-0305-1.

Fisher, M., *Flatline Constructs: Gothic Materialism and Cybernetic Theory-Fiction*, edited by Exmilitary Collective, New York 2018.

Gabriel, C., "The Waluigi Effect", *Substack*, 22. 2. 2023, https://coryeth.substack.com/p/the-waluigi-effect (retrieved 11. 7. 2025).

Hansen, C., Hansen, C., Maystre, L., Mehrotra, R., Brost, B., Tomasi, F., & Lalmas, M., "Contextual and Sequential User Embeddings for Large-Scale Music Recommendation", *Proceedings of the 14th ACM Conference on Recommender Systems (RecSys '20)*, 2020, pp. 53–62, https://doi.org/10.1145/3383313.3412248.

Janus, "Simulators", *LessWrong*, 2. 9. 2022, https://www.lesswrong.com/posts/vJFdjigzmcXMhNTsx/ (retrieved 11. 7. 2025).

@kartographien, "The Waluigi Effect [...]", *X*, 20. 2. 2023, https://x.com/kartographien/status/1627741633377669124 (retrieved 11. 7. 2025).

Landi, G. in Zampini, A., *Linear Algebra and Analytic Geometry for Physical Sciences*, Cham 2018.

@lefthanddraft, "Claude does not view the "jailbreak" as some kind of adversarial attack [...]", *X*, 4. 12. 2024, https://x.com/lefthanddraft/status/1864107641758507281 (retrieved 11. 7. 2025).

Lem, S., *The Futurological Congress*, New York 1974, p. 100.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. in Dean, J., "Distributed Representations of Words and Phrases and Their Compositionality", *arXiv*, 2013, https://arxiv.org/abs/1310.4546 (retrieved 11. 7. 2025).

Nardo, C., "The Waluigi Effect (mega-post)", *LessWrong*, 3. 3. 2023, https://www.lesswrong.com/posts/D7PumeYTDPfBTp3i7/the-waluigi-effect-mega-post (retrieved 11. 7. 2025).

Negarestani, R., "The Labor of the Inhuman, Part I: Human", *e-flux journal* 52, February 2014, https://www.e-flux.com/journal/52/59920/the-labor-of-the-inhuman-part-i-human/ (retrieved 11. 7. 2025).

Negarestani, R., "The Labor of the Inhuman, Part II: Inhuman", *e-flux journal* 53, March 2014, https://www.e-flux.com/journal/53/59893/the-labor-of-the-inhuman-part-ii-the-inhuman/ (retrieved 11. 7. 2025).

Negarestani, R., "The Inhuman (a quick read)", *Toy Philosophy,* 8. 4. 2018, https://toyphilosophy.com/2018/04/08/the-inhuman-a-quick-read/ (retrieved 11. 7. 2025).

Perić, L., "The Bitter Lesson is coming for Tokenization", *lukalp,* 24. 6. 2025, https://lucalp.dev/bitter-lesson-tokenization-and-blt/ (retrieved 11. 7. 2025).

@repligate, "Waluigi effect!!", *X*, 21. 2. 2023, https://x.com/repligate/status/1627852731103715328 (retrieved 11. 7. 2025).

Rieder, B., *Engines of Order: A Mechanology of Algorithmic Techniques*, Amsterdam 2020, p. 200–201.

Srnicek, N., *Platform Capitalism*, Cambridge 2016.

Tshitoyan, V., Jain, A., Weston, L., Dunn, A., Rong, Z., Kononova, O., Ceder, G., & Persson, K. A., "Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature", in: *Nature* 571, 2019, pp. 95–98.

Wark, M., *A Hacker Manifesto,* Cambridge, MA 2004.

Inhuman Vectors

Keywords: vectors, large language models, inhumanism, self-revision, value alignment, Waluigi effect

As the gap in capabilities between artificial neural networks and humans seems to be unavoidably closing, we are drowning in the incomprehensible buzz of hype, fear and denial. This technology, gradually developed over the last few decades, is often

portrayed as something both altogether newborn and at the same time too abstract and alien to be considered human to any extent. Historically, antihumanist positions can be construed as more or less discrete deconstructions of humanist exceptionalism's many manifestations, while still proceeding from its essentialist premises all the same. An attempt at stating this false dichotomy's central problem anew, inhumanism is a distillation of the human rational core as the capacity-for and commitment-to continuous normative self-revision in the discursive space of reasons. In this paper we argue that there are meaningful similarities between how large language models (LLMs) and humans structure their respective images of the world, and that these similarities can inform our self-understanding and our commitments. To this end, we present a preliminary exploration of artificial neural networks as a mirror for humanity's self-examination,in view of the way LLMs' latent spaces robustly embed semantic maps of meaning as vectors. The same process is at work with platform user data embeddings, materialized through informing the corporations' counterintuitive algorithmic meta manoeuvres with massive quantities of vectorized behavioural data. In light of speculative insight into the metastability of LLMs' semiotically encoded orientation, we see the volatility of LLMs' vectorial value alignment not only as a challenge, but simultaneously as an opportunity.

Nečloveški vektorji

Ključne besede: vektorji, veliki jezikovni modeli, nehumanizem, samorevizija, usklajevanje vrednot, Waluigi učinek

Ker se razlika v sposobnostih med umetnimi nevronskimi mrežami in ljudmi neizogibno zmanjšuje, se utapljamo v nerazumljivem vrvežu navdušenja, strahu in zanikanja. Tehnologija, ki se je postopoma razvijala v zadnjih nekaj desetletjih, se pogosto prikazuje kot nekaj povsem novega in hkrati preveč abstraktnega in tujega, da bi jo lahko v kakršni koli meri šteli za človeško. Zgodovinsko gledano lahko antihumanistična stališča razumemo kot bolj ali manj diskretne dekonstrukcije številnih manifestacij humanističnega izjemnosti, ki pa vseeno izhajajo iz njenih esencialističnih premis. Nehumanizem je poskus ponovne opredelitve osrednjega problema te lažne dihotomije in je destilacija človeškega racionalnega jedra kot sposobnosti in zavezanosti k neprekinjenemu normativnemu samoreviziranju v diskurzivnem prostoru razlogov. V tem članku trdimo, da obstajajo pomembne podobnosti med tem, kako veliki jezikovni modeli in ljudje strukturirajo naše podobe sveta, in da te podobnosti lahko vplivajo na naše razumevanje samih sebe in naše zaveze. V ta namen predstavljamo predhodno raziskavo umetnih nevronskih mrež kot ogledalo za samoanalizo človeštva, ob upoštevanju načina, kako veliki jezikovni modeli v svojih latentnih prostorih

trdno vgrajujejo semantične zemljevide pomenov kot vektorje. Enak proces poteka pri vgrajevanju podatkov uporabnikov platforme, ki se materializirajo z obveščanjem protislovnih algoritmičnih meta manevrov korporacij z ogromnimi količinami vektorskih podatkov o vedenju. Glede na špekulativni vpogled v metastabilnost semiotsko kodirane usmeritve VJM, nestabilnost vektorske usklajenosti vrednosti VJM vidimo ne le kot izziv, ampak hkrati tudi kot priložnost.

About the authors

Lea Sande holds a bachelor's degree in sociology of culture from the Faculty of Arts at the University of Ljubljana. She produces a monthly programme called Povratne zanke (Feedback Loops) on Radio Študent, where she discusses internet theory, culture and literature. She works in the field of sociology of technology, focusing on the role of computation in contemporary capitalism, the development of artificial intelligence and the cultural derivatives of emerging technologies.

Email: lea.sande3@gmail.com

Tisa Troha is an architect with an MA from the University of Ljubljana, Faculty of Architecture. Her theory-fictional final thesis project plots a perspective of architecture as inhuman technics of the outside. She is part of the editorial board of *Šum*, a journal for art theory and fiction, and moonlights as a graphic designer, sound designer and musician.

Email: tisatroha@gmail.com

O avtoricah

Lea Sande je diplomirana sociologinja kulture Filozofske fakultete Univerze v Ljubljani. Ustvarja mesečno oddajo Povratne zanke na Radiu študent, kjer se ukvarja s spletno teorijo, kulturo in literaturo. Svoje delo opravlja na področju sociologije tehnologije, kjer se osredotoča na vlogo komputacije v sodobnem kapitalizmu, razvoj umetne inteligence in kulturne odvode emergentnih tehnologij.

E-naslov: lea.sande3@gmail.com

Tisa Troha je arhitektka z magisterijem Fakultete za arhitekturo Univerze v Ljubljani. Njen teoretsko-fikcijski projekt zaključnega dela riše perspektivo arhitekture kot nečloveške tehnike zunanjosti. Je članica uredniškega odbora revije za umetnostno teorijo in fikcijo *Šum* in dela kot grafična in zvočna oblikovalka ter glasbenica.

E-naslov: tisatroha@gmail.com