The Fallacy of Autonomous AI

Dominik Schlienger[1]

[1]University of the Arts Helsinki, Department for Music and Technology, Töölönlahdenkatu 16, 00100 Helsinki, Finland

Abstract

In the mainstream media, concerns are voiced about the potency of AI as a threat to humanity. Some of the academic literature that gives credence to that threat, does so in reference to posthumanism, where we find, besides conceptional tools for thinking technology in an indeterministic way, also an appeasement towards *strong AI*. This paper's aim is to demonstrate that a possible threat by AI does not come from the alleged attainability of autonomous, cognising machines. Starting out from the conception of technology as a socio-material arrangement, if becomes clear that AI, like any other technology, is socially *performed*. What is more, supposedly autonomous machines run on algorithmic, linguistic, written *code*, as I show in an analysis of computer language as Derridean writing. As such, they are extensions of *human* cognition. To proclaim machines conscient and autonomous, is hence not just misleading per se, but disguises the human agency that uses the AI-autonomy as a proxy.

Keywords

Code as Derridean writing, Fallacy of AI autonomy, AI ethics, Posthumanism and AI

1. Introduction

In recent headline news, governments and business leaders described AI as an "extinction-level threat" to humans[1, 2], and warned that drastic regulation will be necessary[3]. That media discourse attained credibility through news of a 'founder figure' resigning from a senior position with a leading player in AI due to "growing dangers from developments in the field"[4], compounded by governments turning to 'AI experts' for policy advice[5].

The discourse on AI as a threat to humanity, is reflected in computer science's increased interest in ethics in technology. Along with the growing understanding for the ethical question in technology development, comes a desire to understand the social impact of emerging technologies like AI. The prognostics cover a wide range from cautious analysis[6] to more outlandish takes on the subject[7].

Reflections on that discourse in a particular segment of computer science, seem to be rooted in a literal interpretation of some posthumanist concepts pervading academic writing on AI and ethics. Besides conceptional tools for thinking technology in an indeterministic way[8, 9, 10], Posthumanism provides an argument for the advent of *strong AI* and, what is more, a sense of appeasement towards it, for example in N. Katherine Hayles' works[11, 12, 13, 14].

I aim to show here that the discourse on AI is rooted in two mutually exclusive rationales. Separated by a fine line, we have the question of *how we do AI* on the one side, and *what AI does to us* on the other. I will argue, that the latter question is based on a misconception of technology and that the posthumanist argument for AI as a conscient other, as contended by Hayles, is based on a logical fallacy.

For Hayles, the human tradition is just a necessity to allow the *emergence* of the posthuman through a conceptualisation of intelligence as co-produced with intelligent machines. There are no essential differences or absolute demarcations between bodily existence and computer simulation. This "privileges information over materiality, considers consciousness as an epiphenomenon and imagines the body as a prosthesis for the mind"[11]. Hayles states that "cognitive technologies are now a potent force in our planetary cognitive ecology" and that there are "rapidly escalating complexities created by the interpenetration of cognitive technologies with human systems" The two last tenets, for Hayles "are not debatable*[sic]*"[14]. She further insists that there are technical systems that are autonomous; that there

7th Conference on Technology Ethics (TETHICS 2024), November, 6–7, 2024, Tampere, Finland

*Corresponding author.

dominik.schlienger@uniarts.fi (D. Schlienger)
0009-0002-4340-9116 (D. Schlienger)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

is cognition in technical systems; and limited human decision power in such systems. These notions are not uncommon in posthumanist thought, as recent overviews confirm, some of them acquiescing to the logical consequences of such thinking[15, 16], some exercising it as a conceptual possibility[17].

Although I do not concur with the type of criticism of posthumanism that aims to defend the Enlightenment at best[18], and white male hegemony at worst, I still attempt to draw attention to what I see as a weakness in Hayles' argumentation that underpins the current discourse according to which computational machines could be autonomous to the extent the media portrays. In contrast, I assert that the reliance of computational machines to run on algorithmic code makes the concept of technical autonomy as a *cognising other* logically impossible: Through language, machines are extensions of the cognitive practices that constitute the language they run on[19]. The computer is to the brain what the hammer is to the hand. To understand this *simile* in the way intended, it will be necessary to provide a bit more background on some of the subjects I discuss in more detail in my thesis[19]: Firstly, the relation between tool and body, and secondly, the relation of the body (as a tool) to other bodies.

1.1. The Body — Tool — Machine Gradient

As fellow musicians and colleagues in the performing arts would agree — the notion of the body as an instrument, and reciprocally, tools as an extension of the body, is essential part of (musical) practice. Based on that, I define the difference between a tool and a machine as the degree of motoric engagement of the body in the tool and its absence in the machine.[1]

This gradient of bodily involvement is not one of an increase in technical complexity – the technical complexity of the human body is high, for some tools it is low, for many machines high again. What I suggest is, that the gradient is one of agency. At first this might be misleading, as it insinuates a decrease in human agency towards the machine-end of the gradient. My hypothesis, however, is that we are looking at a *displacement* of agency, rather than a replacement, and an increase in agentic complexity, a multiplication of agents.

1.2. Socio - Material Arrangements

In *On the Mode of Existence of Technical Objects*[21], French philosopher Gilbert Simondon describes technology as something ontologically distinct, motivated by something akin of empathy for technical objects: "It is important that the technical object [. . .] remains an instrument, even a friend, in our relation to the world" as he stated later in an interview[22]. Importantly, Simondon saw autonomy in technology not primarily as a technical feature: "Automatism, and its utilisation in the form of industrial organisation, which one calls automation, possesses an economic or social signification more than a technical one."[21]

Bruno Latour, using Simondon as an important source, goes one step further and disentangles the *technical* from the technical object[2] as something the technical mode of existence "leaves in its wake"[23], opening conceptual approaches to understand the technical as entirely performative. American Anthropologist Lucy Suchman emphasises the performative nature of technologies by describing (seemingly) autonomous machines as *socio-material arrangements*[24]. And software exemplifies the socio-technical nature of machines like no other: Every line of software code, no matter how many times it has been run, has been written by humans, and what is more, not by one human alone, but by a community of programmers, developers, inventors, hardware designers and participating users. However automated a machine is, deployed by human actors, its cognising agency is human — a social practice. In the following section, I will assert, by example of Hayles' argumentation, that autonomous cognition in machines, far from being "not debatable", is a non sequitur.

[1] If we suppose that the gradient that links the body to tools is the same as the one linking tools to machines, Marcel Mauss's "Techniques of the Body" provides the missing link[20].

[2] For a longer critical discussion on that disentanglement, see my thesis, p. 80 and pp. 83-94 [19]

2. Code as Derridean Writing

If AI is a tool like any other technical tool, and as such, not a cognising *other* but an extension of *our* embodied cognition, then the dangers of AI lie with *what we do with AI.* On the contrary, if it can be logically argued that the cognition in AI is not human, then it is possible that AI *does things to us.* As a working hypothesis, I propose that through *code* as computer languages, and algorithms written in such languages, computational machines, effectively, are linguistic devices and hence have grammar, syntax, and semantics just like natural languages. They are a system of human meaning making.

In[12] Hayles argues that computer languages are *not* languages in the sense of natural languages, which consequently makes it possible to argue that the meaning making in computational machines is disconnected from human meaning making. Yet, as I discuss elsewhere in more detail[19], a linguistic analysis shows, that computer languages indeed do constitute languages, and in particular, constitute Derridean *writing*, based on French philosopher Jacques Derrida's work *On Grammatology*[25]. Here a summary of my argumentation:

Derrida adopts from Saussure the binary oppositions that all speech or text has to articulate if it intends to make sense[26]. Herein lies a conceptual analogy to binary code: on the technical (procedural) level, all code works in a similar way, and the complexity of sense-making can be explained through the proliferation of multiple simple structures. The complexity or limitation of meaning lies not in the binary opposition in sense-making, but in the complexity of the available signal-paths.

In Umberto Eco's *Watergate Model*[27] signal paths link the watergate of a conveying system, with a control station downstream, the conveyed system. The code is what connects them, element by element. The elements of the former become the expression of the latter, and the latter becomes the content of the former. For Eco, there are no signs, only sign-functions. This defines the difference between a signal and a sign: "A signal is a pertinent unit of a system that may be an expression system ordered to a content but could also be a physical system without any semiotic purpose" A signal can be entirely free of meaning. But when the signal is recognised as an antecedent of a foreseen consequent it may be viewed as a sign.

Writing, for Derrida, leaves traces. What is written, remains, and can be transported, read later, spoken, activated, compiled, executed, debugged, updated, or left to (code-)rot. One essentially Derridean aspect of this type of universal writing comes from the notion of différance, an intentional misspelling of 'différence', which Derrida uses to indicate firstly, that by writing the word differently (as a grapheme) there is an additional meaning to the conventional understanding of the word when we hear it as a phoneme (they are pronounced exactly the same way). Secondly, he means to indicate the double meaning of the French word 'différer' which translates to differ as well as to defer. Différance in the sense of to differ stands for the binary sense-making encapsulated in Saussure's linguistic sign. Différance stands for defer, however, to indicate that meaning can be deferred through chains of signifiers, and as a grapheme, in particular, also postponed.

Therefore, computer language and written natural language do not conceptually differ. Furthermore, due to Iterability, is is possible for a written text to remain readable after the person who wrote it is gone [18]. It can also be taken out of its original context and, when reiterated, work in a different context. The writing itself remains, it is a trace. That a sentence means something entirely different according to whom to, when, and how I say it does not change the literal text, the code. Further, Derrida insists that "there is nothing outside context", so the signifier's *conditio sine qua non* is context. Hence, computational machines are linguistic systems, congruent with Eco's Watergate-Model. The code they run on is iterable in a Derridean sense, and can be deferred to another place and time.

Yet Hayles concludes that computer languages are *not languages in a Derridean sense* using the same sources, (Derrida and Saussure), based on the claim that iterability could not be applied to code, "where contexts are precisely determined by the level and nature of the code. [...] Code may be rendered unintelligible if transported into a different context—for example, into a different programming language or a different syntactic structure within the same language"

And here lies the fallacy of the argumentation for an autonomously cognising computational machine: The assumption, that in computer languages, in contrast to any other languages, meaning could be hard-coded into the signifier. But this is not the case: The whole machine, including its "dependencies", like operating system, hardware, and firmware constitute the code (the signal-paths of the watergate), and hence are part of the iterable writing, part of the signifier, the trace. Hayles implies that for code to constitute writing in a Derridean sense, the code would need to be understood by any system. But this contradicts Derrida by implying that code, as the signifier, should *mean* the same (rather than simply *be* the same) in every iteration: The signified would need to stay constant while the signifier changed through iteration — neither true for code nor natural language; a text in French written 200 years ago means something slightly different if read today. And its meaning might be opaque but readable for Italians, but unintelligible to most English speaking people.

Only through a muddling of signifier and signified is it possible to argue that the meaning making in computational machines is in any way disconnected from the meaning making of the human agent(s) deploying, building and activating that machine. What is more, the fact that machines, just like language and any other type of code, need performing — relational activating, to make meaning, makes the point even more salient that *we do AI*, and AI is not actually able to *act on us*.

The (directional) process from an antecedent to a consequent, a subject broached in the description of the watergate-model, is instrumental in meaning making. For meaning to be constituted, there needs to be an antecedent and a consequent activated through code which *runs*.

3. Cognition at run-time

Computational machines need to be powered up, they only work at *run-time*. Consequently, if machine did something that could equate to cognition, it could only be while they *run*. So, one could argue that run-time as *motor action* is a precondition for cognition[28]. Reversely, it means that every entity that *runs* could in theory, potentially cognise. On this basis, motor-neural activity is enough for basic cognitive activity. To contest this would be anthropocentric, humanist exceptionalism.

The above might seem to support Hayles' thesis of *nonconscious cognisers*[14], that implies that many nonconscious entities, for example plants, have (limited) volition through cognition (they react to stimulus and adjust their actions accordingly). Yet Hayles' inclusion of technical systems as cognisers happens on the basis of the logical fallacy that computational machines running on code written by humans can be ontologically disconnected from human meaning making.

Nota bene there is no problem in itself if machines have cognitive abilities; we could interact with them, more or less aptly, like we do with other cognitive beings we encounter. The problem is precisely that it is *not* the machine that has the cognitive ability when running on code written by humans. With the claim that machines can act autonomously, we are up against *somebody's* agency, a person or persons whose intentions are not necessarily clear as we are negotiating with a proxy. To recognise human cognition in computational machines is not anthropocentric. But to deny it *is*: Counterintuitively maybe, if we follow Hayles' argumentation that the cognition in plants and in computational technical systems (running on code) are the same in kind, we accidentally anthropomorphise all non-human entities, rendering most of posthumanism's rationales moot. To emphasise the intertwining and the interdependencies of entities in a multimodal world with the motive to promote inclusion and respect for non-human actants is only sensible when we recognise our own agency and its consequences in the technical objects we deploy.

4. Concluding Remarks

Returning to the two types of AI discourse, I tried to make the case for the type of question that asks *how we do AI*. And that the question of *what AI does to us* misses the point in a way that renders AI into a Trojan horse, using AI as a proxy agent by proclaiming AI autonomous. But who is to gain from such an endeavour? — Some answers to this can be found in[6], where the authors meticulously show how the bias of the AI in question, a large language model chat-bot, favours the existing hegemonic viewpoint. Saying so as an employee of a big player in AI, cost one of the authors, Timnit Gebru, her job.

It has been suggested that my line of reasoning (that the linguistic nature of code exposes the fallacy of AI autonomy) may only apply to simple, if/then type of algorithms. In response I feel tempted to suggest that, respectively, linguistics, thence, could only apply to simple, if/then type of meaning making in language. Much to the contrary, even for quantum computing, sense-making and algorithmic processes rely on the binary oppositions that all speech or text has to articulate[26]. Let's re-iterate here that the complexity or limitation of meaning lies not in the binary opposition in sense-making, but in the complexity of the available signal-paths. This applies to code as much as to natural language. Actually, it is precisely for the most evident application of "AI" concepts — chatbots based on large language models like ChatGPT, where a linguistic approach is key to understand what the technology is actually doing: For AI developer Timnit Gebru and linguist co-author Emily Bender *et al.*, the realisation that no understanding takes place in chat-bots, and that hence the meaning making is entirely human, is key: "languages are systems of signs, i.e. pairings of form and meaning. But the training data for Language Models is only form; they do not have access to meaning."[6]

I hope I have shown here that the conception of machines — and computational machines in particular — as socio-material arrangements allows for an understanding of the technical as a social practice we actively perform. To think of machines deployed by us as cognising other is a logical fallacy in view of the linguistic nature of program languages and code. I am not disputing the potential dangers of technologies, nor questioning their possibilities. But I contest the deterministic conception of technology that attributes the dangers to the technology itself. The harms and dangers of AI are the direct result of (sloppy) human thinking, not something we can blame AI for in lieu. To attribute agency to an external, *deus ex machina* may be convenient for the hegemony that profits from doing so, but it is a cop-out from collective responsibility. So we need to rethink the way we develop AI. This insight resonates with Hannah Arendt's ethical imperative to "think what we're doing", to not rely on "artificial machines to do our thinking and speaking". Failing that, we are "thoughtless creatures at the mercy of every gadget which is technically possible, no matter how murderous it is"[29]. That the agency in AI is not autonomous but an extension of human cognition, provides a trojan horse for whoever is controlling the machine.

5. Acknowledgments

Many thanks to Dr Sofi Kurki, VTT Finland, whose input for this paper was essential and appreciated in discussions on the subject in another context. Further thanks go to Professor Jan Schacher, at the Department for Music & Technology, Uniarts Helsinki; and Professor Lily Diaz-Kommonen, Aalto University, for encouragement.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] N. Y. Times, AI poses 'risk of extinction,' industry leaders warn, 2023. URL: https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html.

[2] M. Egan, AI could pose 'extinction-level' threat to humans and the us must intervene, state dept.-commissioned report warns, 2024. URL: https://edition.cnn.com/2024/03/12/business/artificial-intelligence-ai-report-extinction/index.html.

[3] C. Vallance, Powerful artificial intelligence ban possible, government adviser warns, 2023. URL: https://www.bbc.com/news/technology-65779181.

[4] B. Z. K. . C. Vallance, Ai 'godfather' geoffrey hinton warns of dangers as he quits google, 2023. URL: https://www.bbc.com/news/world-us-canada-65452940.

[5] N. Omer, Sunak, musk and ai: what we learned from the bletchley park summit, 2023. URL: https://www.theguardian.com/technology/2023/nov/03/rishi-sunak-elon-musk-ai-summit-what-we-learned.

[6] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 610–623. URL: https://doi.org/10.1145/3442188.3445922. doi:10.1145/3442188.3445922.

[7] L. D. Milici, M. R. Milici, Is ai capable of generating an ethic to save the planet and contemporary society?, Proceedings 81 (2022). URL: https://www.mdpi.com/2504-3900/81/1/87. doi:10.3390/proceedings2022081087.

[8] D. J. Haraway, The Haraway Reader, New York Routledge, 2004.

[9] O. Pyyhtinen, S. Tamminen, We have never been only human: Foucault and latour on the question of the anthropos, Anthropological Theory 11 (2011) 135–152. URL: https://doi.org/10.1177/1463499611407398. doi:10.1177/1463499611407398.

[10] K. Barad, Meeting the Universe Halfway, Duke University Press, New York, USA, 2006. URL: https://doi.org/10.1515/9780822388128. doi:doi:10.1515/9780822388128.

[11] N. K. Hayles, How We Became Posthuman:Virtual Bodies in Cybernetics, Literature and Informatics, Chicago: University of Chicago Press, 1999.

[12] N. K. Hayles, My Mother Was a Computer: Digital Subjects and Literary Texts, University of Chicago Press, 2010.

[13] N. K. Hayles, How We Think: Digital Media and Contemporary Technogenesis, University of Chicago Press, 2012.

[14] N. K. Hayles, Unthought: The Power of the Cognitive Nonconscious, University of Chicago Press, 2017.

[15] R. Nath, R. Manna, From posthumanism to ethics of artificial intelligence, AI Soc. 38 (2021) 185–196. URL: https://doi.org/10.1007/s00146-021-01274-1. doi:10.1007/s00146-021-01274-1.

[16] D. Tüfekci Can, Reinterpreting human in the digital age: From anthropocentricism to posthumanism and transhumanism, Journal of Educational Technology and Online Learning 6 (2023) 981–990. doi:10.31681/jetol.1341232.

[17] J. Brusseau, Mapping ai avant-gardes in time: posthumanism, transhumanism, genhumanism, Discover Artificial Intelligence 3 (2023) 32. URL: https://doi.org/10.1007/s44163-023-00080-6. doi:10.1007/s44163-023-00080-6.

[18] T. Osborne, N. Rose, Against posthumanism: Notes towards an ethopolitics of personhood, Theory, Culture & Society 41 (2024) 3–21. URL: https://doi.org/10.1177/02632764231178472. doi:10.1177/02632764231178472.

[19] D. Schlienger, Developing Alps: Notes on Agency in Technology, University of the Arts Helsinki, 2023. URL: https://urn.fi/URN:ISBN:978-952-329-287-1.

[20] M. Mauss, Sociologie et Anthropologie, Paris: Presses Universitaires de France, 1968, pp. pp. 364–386. This lecture was given at a meeting of the Société de Psychologie, May, 7th, 1934 and published in the Journal de psychologie normal et patholigique, Paris, Année XXXII, 1935, pp. 271-93.

[21] G. Simondon, On the Mode of Existence of Technical Objects, Univocal, Mineapolis, 2017. Originally published Paris: Aubier, Editions Montaigne, 1958.

[22] Simondon, Gilbert simondon et l'amitié de la technicité, 1967. URL: https://www.youtube.com/watch?v=uqZSvK8jYXQ.

[23] B. Latour, An inquiry into modes of existence : an anthropology of the moderns / Bruno Latour ; translated by Catherine Porter, Harvard University Press Cambridge, Massachusetts, 2013.

[24] L. Suchman, L. A. Suchman, Human-Machine Reconfigurations: Plans and Situated Actions, Learning in Doing: Social, Cognitive and Computational Perspectives, Cambridge University Press,

2007.

[25] J. Derrida, G. Spivak, Of Grammatology, Of Grammatology, Johns Hopkins University Press, 2013.

[26] F. de Saussure, W. Baskin, P. Meisel, H. Saussy, Course in General Linguistics, Columbia University Press, 2011.

[27] U. Eco, A Theory of Semiotics, A Midland Book, Indiana University Press, 1979.

[28] M. Jeannerod, Motor Cognition: What Actions Tell the Self, Motor Cognition: What Actions Tell to the Self, OUP Oxford, 2006.

[29] H. Arendt, The Human Condition, Anchor Books, University of Chicago Press, 1958.