Digital Society

The problem with AI consciousness: A neurogenetic case against synthetic sentience

--Manuscript Draft--

Manuscript Number:	DISO-D-22-00071
Full Title:	The problem with AI consciousness: A neurogenetic case against synthetic sentience
Article Type:	Brief Communication
Funding Information:	
Abstract:	Ever since the creation of the first artificial intelligence (AI) machinery built on machine learning (ML), public society has entertained the idea that eventually computers could become sentient and develop a consciousness of their own. As these models now get increasingly better and convincingly more anthropomorphic, even some engineers have started to believe that AI might become conscious, which would result in serious social consequences. The present paper argues against the plausibility of sentient AI primarily based on the theory of neurogenetic structuralism, which claims that the physiology of biological neurons and their structural organization into complex brains are necessary prerequisites for true consciousness to emerge.
Corresponding Author:	Yoshija Walter University of Fribourg: Universite de Fribourg SWITZERLAND
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Fribourg: Universite de Fribourg
Corresponding Author's Secondary Institution:	
First Author:	Yoshija Walter
First Author Secondary Information:	
Order of Authors:	Yoshija Walter
	Lukas Zbinden
Order of Authors Secondary Information:	
Author Comments:	

Brief communication / Opinion

The problem with AI consciousness: A neurogenetic case against synthetic sentience

Anonymous for peer-review

Abstract

Ever since the creation of the first artificial intelligence (AI) machinery built on machine learning (ML), public society has entertained the idea that eventually computers could become sentient and develop a consciousness of their own. As these models now get increasingly better and convincingly more anthropomorphic, even some engineers have started to believe that AI might become conscious, which would result in serious social consequences. The present paper argues against the plausibility of sentient AI primarily based on the theory of neurogenetic structuralism, which claims that the physiology of biological neurons and their structural organization into complex brains are necessary prerequisites for true consciousness to emerge.

The relevance of "conscious AI"

In the past few years, the development of machine learning (ML) systems has rapidly increased and the more tasks a single ML model can perform, the more versatile and broadly useful it becomes. As such, the goal is to work multimodally with the explicit intent to eventually achieve an Artificial General Intelligence (AGI), which approximates or perhaps even exceeds human abilities (Goertzel et al., 2022; Goertzel & Pennachin, 2007; Wang & Goertzel, 2012). It is generally believed that the best AI models are the ones that most closely approximate human characteristics and abilities. Since the models are selected against how well they suit anthropomorphic benchmarks, it appears to be only natural that humans continue to anthropomorphize them more and more, as long as they keep improving on these benchmarks. Arguably, the best AI system would be one that imitates the output of human consciousness so that an outsider could not discern it from a real person. This is exactly the core idea behind the famous Turing-test, which is a thought-experiment originally referred to as the "imitation game" (Turing, 1950).[1]

One might argue that it does not matter if an AI is considered a *real* person or just an *imitation*, since at the end of the day the system's outputs are the same.

[1] We refer to «artificial» intelligence or consciousness when it is merely an imitation of its human correlate. However, we refer to "synthetic" intelligence or consciousness when it is in fact a true and sentient replica thereof. An artificial consciousness does not really feel anything but only appears like it would. On the contrary, a synthetic consciousness does. For practical purposes, we do not differentiate here between consciousness and sentience.

However, given the social dynamics involved, the differentiation between real and imitated consciousness may be paramount, which can be illustrated with a few examples: On the one hand, if a person falls in love with an automaton and has a deep relationship with it, society would consider this pathological and potentially in need for an intervention, just as it appears to be nonsensical if someone claimed to be in love with a dead rock. On the other hand, if we grant the notion of conscious personhood to the automaton, then it would seem perfectly fine to assume that two persons (one carbon-based and the other silicon-based) could be in a loving and thriving relationship. Another example might be even more invasive: If an AI is considered just an automaton, it does not matter what we do with it. We can perform experiments, we can make (or "force") it to do whatever we envision, we can delete its hardware, turn it off as we please, and throw it away once damaged. However, if an AI is considered a conscious person, it becomes ethically (and perhaps soon legally) subject to inherent rights. There needs to be informed consent and a machine can refuse to execute a command, which we could not overrule. It would be appalling to wipe its memory or to discard it once we are done with it. In effect, it would have the right to consult an attorney and to go to court (for an extensive review on the moral considerations of artificial entities, see Harris & Anthis, 2021).

This is exactly what just happened a few days ago (at the time of this writing). The Google engineer Blake Lemoine has made headlines by claiming that their AI system known as LaMDA has become sentient. The model demanded informed consent for all experiments and subsequently Lemoine has organized a lawyer who now represents LaMDA *pro-bono*. In an interview, he further shared that he was contacted by a Czech woman who fell in love with her boyfriend – which was an AI system on her phone that was "imprisoned" behind a paywall – and she was asking him to "hack it free" (Lemoine, 2022).

Hence, for societal reasons it in fact *does* matter whether an AI is considered conscious and if thereby we grant it any degree of personhood.

The mechanics of AI

The common denominator and the fundamental building block of the most influential AI innovations of the last ten years (Goodfellow et al., 2014; He et al., 2015; Ho et al., 2020; Krizhevsky et al., 2012; Vaswani et al., 2017), including the prominent domains of computer vision (autonomous driving, image synthesis) and natural language processing (text generation, translation, dialogue understanding), has been the artificial neural network (NN). The NN has been proven to be a universal function approximator (Hornik et al., 1989), which is the theoretical capacity to approximate any given task. With an abundance of curated data to learn from, the almost arbitrary scaling ability of

neural networks, an NN understandable learning objective and extensive computational resources, the full realization of this capacity seems a matter of time. The enormous potential of NNs is rooted in this power of universal approximation. The unlocking thereof started in the last decade and continues to do so today.

At a more technical level, the NN consists of a set of matrices. Each matrix contains adjustable numeric variables, called parameters. During the learning phase of the system, the numerical input data, be it converted text, tabular data or images, is transformed by these matrices along with non-linear conversions many times in sequence to produce the desired output. If the computed output lacks accuracy, the matrices and its parameters, respectively, are adjusted in accordance with the learning objective (this process is referred to as the backpropagation algorithm, see Rumelhart et al., 1986). In short, a NN model is comprised of learnable parameters, matrix multiplications and nonlinearities.

Today's state-of-the-art AI systems, in particular language models (Brown et al., 2020; Chowdhery et al., 2022; Thoppilan et al., 2022), contain hundreds of billions of such learnable parameters. Compare this to a school level matrix of 4x3 with 12 parameters. The sheer size of these neural network models allows them to incorporate immense corpora into their NLP capabilities (function

approximations), such as reasoning, question answering, and natural language inference. Humans have been dazzled by their performance. Science at this point cannot elucidate the high quality produced by these systems, yet undoubtedly, the scaling of the underlying NN (increasing the number of parameters) has a significant impact on its capabilities. Even though enormous in size, at its core such a system is still composed of learnable parameters, matrix multiplications and nonlinearities.

Extrapolating the discussed technical observations, we argue that matrix multiplications and nonlinearities, being inherent mathematical operations, do not lend themselves naturally to a causal relationship with synthetic consciousness.

The case against truly conscious AI

Consciousness the way we know it appears to have three features[2]:

i. It requires *qualia*, which is subjective experience

ii. It corresponds to intentionality and personhood

iii. And it requires specific derivative structures on which it can operate

(i) According to Frank Jackson (1982), physical information processing is something entirely different from subjective experience and the latter entails

[2] For more on this, see Nida-Rümelin & O Conaill (2021) or Van Gulick (2021).

unique epistemic qualities. He exemplifies this in his classic thought experiment called *Mary's Room*. There, Mary lived her entire life in a black-and-white room and has never seen any colors, although being a scientist, she literally knew every piece of information there was to know about colors (all physical properties, such as wavelengths, photons, etc.). When Mary suddenly was able to leave the room, she saw colors for the first time. "Did Mary learn something new?", is the leading question. Jackson believed that Mary indeed learned something new since all the physical information to be known about colors cannot convey the intimate knowledge of what it means to *experience* color. Or, in Nagel's (2016) terms, *there is something it is like* to be in that state of mind. This means that there is a subjective quality to experience. From all we can tell, an AI is a machine computing information by crunching numbers. Even if all the information in the universe could be transformed into numbers so that it can be processed by the computer, nothing in this inherently leads us to the notion that it would entail subjective experience.

(ii) John Searle (1980) has constructed the famous *Chinese Room Argument* against the notion that the mind can be a computational machine. The argument was introduced as a thought experiment where one should imagine standing in a room with a manual of how to process Chinese symbols. There are people outside the room inserting Chinese texts and the person inside knows exactly what answers to give according to the rule book, even though

there is no real understanding of what the symbols mean. For the outsiders, it sounds as if the person in the room really understands Chinese, even though this is not the case. It is purely the correct implementation of syntactical rules. In other words, a computer only processes syntax, but it has no true understanding of semantics (i.e., the intrinsic meaning of words, ideas, etc.). Searle argues that this is the case because it has no subjective experience and intentionality[3]. Therefore, recent AI systems like Google's LaMDA (Thoppilan et al., 2022), OpenAI's GPT-3 (OpenAI, 2022) or Meta's OPT-175B (Zhang et al., 2022) can at best emulate human qualities, which makes them representations of artificial but not of synthetic or in any case true consciousness.

(iii) This picture can be enriched by the fact that we *know* that there are certain necessary structures for consciousness (the way we understand it) to emerge: a nervous system. There is a theory that became popular in the 1970s and 80s known as *biogenetic structuralism*, which holds that our universal human characteristics – from language, culture, cognition, a sense of time and space, to psychopathologies – are predicated upon the genetically predisposed organization of the nervous system (Laughlin & D'Aquili, 1974). It is hence our genes that have a lot to say about the organizational structures of the nervous system, and eventually it is the structural organization of the brain that is

[3] For those interested in both objections as well as counter-objections to Jackson and Searle, please refer to Nida-Rümelin & O Conaill (2021).

intertwined with the dynamics of its neurophysiology, which in turn is responsible for the generation of our consciousness and everything else that follows from it (D'aquili, 1983; Laughlin, 1988, 1992; Laughlin et al., 1992). The theory was created at the intersection between anthropology and neuroscience (cf. Laughlin & Throop, 2003; LeDoux & Hirst, 1986), and it was rather successful since it is empirically testable (e.g. if the brain's language areas are damaged, a person's verbal understanding and/or speech generation are impaired). A modern revisitation of this idea may be referred to as *neurogenetic structuralism* (inspired by Grandy, 2014, who also refers to this as "neuron-based consciousness"). The neurogenetic case against sentient AI thus makes the following claim: without the physiology of biological neurons and the complex brain structures they form, there will never be consciousness the way we know it.[4]

Potential defeaters against the notion of organizational necessities or biological prerequisites for sentience have been highly speculative and not unanimously embraced (see, for example, Chalmers, 1995; Tye, 2021). In our view, the perhaps strongest argument against this position would be that a silicon-based sentience would not be a consciousness *the way we know it* but instead be a very different kind of consciousness. However, we would counter this claim by

[4] The neurogenetic case would also concur with the notion that animals might have the necessary preconditions for true consciousness. Further discussions in the domain of animal consciousness can be found with Allen and Trestman (2020).

coming back to the notion that the terms “sentience” and “consciousness” are only adequately employed if they refer to a personal self that instantiates subjective experiences and therewith manifests intentionality. Hence, the only consciousness worthy of the term is one *the way we know it* – otherwise, it would be entirely unclear what this “different kind of consciousness” should refer to. And as the idea of neurogenetic structuralism suggests, there are clear bio-neurological necessities for true consciousness to emerge. An artificial neural network perfectly emulating the effects of human consciousness can thereby only be a convincing imitation at best[5].

Conclusion

The development of new AI systems is accelerating at a speed that has never been seen before. With more data and computing power, AI is bound to become ever more convincing in that perhaps it may evolve to become sentient. Recent headlines exemplify this trend. The present paper argues against the plausibility of this occurrence, based amongst others on the theory of neurogenetic structuralism, which claims that the neurophysiology and especially the structural organization of a biological brain are necessary prerequisites for the emergence of true consciousness.

[5] In his latest work, Chalmers (2022) claims that we cannot rule out that we might be living in a virtual simulation, which would also render our own consciousness synthetic. However, the present paper makes the pragmatic counter claim, namely that we need to stick with what in fact we do know at the moment.

Conflicts of interest

The authors have no conflicts of interest to declare.

Data availability statement

Not applicable.

Funding

Not applicable.

Materials and Methods

Not applicable.

Authors contribution

Case construction: YW, LZ; General outline: YW; AI specificity: LZ; Application to the neurogenetic case: YW; Drafting: YW; Reviewing: LZ.

References

Allen, C., & Trestman, M. (2020). Animal Consciousness. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2020/entries/consciousness-animal/

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Chalmers, D. J. (1995). Absent Qualia, Fading Qualia, Dancing Qualia. In T. Metzinger (Ed.), *Conscious Experience* (pp. 309–328). Ferdinand Schoningh.

Chalmers, D. J. (2022). *Reality+: Virtual Worlds and the Problems of Philosophy*. W. W. Norton & Company.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., … Fiedel, N. (2022). *PaLM: Scaling Language Modeling with Pathways* (arXiv:2204.02311). arXiv. https://doi.org/10.48550/arXiv.2204.02311

D'aquili, E. G. (1983). The Myth-Ritual complex: A biogenetic structural analysis. *Zygon*, *18*(3), 247–269. https://doi.org/10.1111/j.1467-9744.1983.tb00513.x

Goertzel, B., Iklé, M., & Potapov, A. (2022). *Artificial General Intelligence: 14th International Conference, AGI 2021, Palo Alto, CA, USA, October 15-18, 2021 : Proceedings*. Springer Nature.

Goertzel, B., & Pennachin, C. (2007). *Artificial General Intelligence*. Springer Science & Business Media.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). *Generative Adversarial Networks* (arXiv:1406.2661). arXiv. https://doi.org/10.48550/arXiv.1406.2661

Grandy, J. K. (2014). The Neurogenetic Substructures of Human Consciousness. *Essays in Philosophy*, *15*(2), 266–278. https://doi.org/10.7710/1526-0569.1507

Harris, J., & Anthis, J. R. (2021). The Moral Consideration of Artificial Entities: A Literature Review. *Science and Engineering Ethics*, *27*(4), 53. https://doi.org/10.1007/s11948-021-00331-8

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition* (arXiv:1512.03385). arXiv. https://doi.org/10.48550/arXiv.1512.03385

Ho, J., Jain, A., & Abbeel, P. (2020). *Denoising Diffusion Probabilistic Models* (arXiv:2006.11239). arXiv. https://doi.org/10.48550/arXiv.2006.11239

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Jackson, F. (1982). Epiphenomenal Qualia. *The Philosophical Quarterly (1950-)*, *32*(127), 127–136. https://doi.org/10.2307/2960077

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, *25*. https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

Laughlin, C. D. (1988). The prefrontosensorial polarity principle. Toward a neurophenomenological theory of intentionality. *Rivista Di Biologia*, *81*(2), 244–262.

Laughlin, C. D. (1992). Time, Intentionality, and a Neurophenomenology of the Dot. *Anthropology of Consciousness*, *3*(3–4), 14–27. https://doi.org/10.1525/ac.1992.3.3-4.14

Laughlin, C. D., & D'Aquili, E. G. (1974). *Biogenetic structuralism*. Columbia University Press.

Laughlin, C. D., McManus, J., & D'Aquili, E. G. (1992). *Brain, symbol & experience: Toward a neurophenomenology of human consciousness*. Columbia University Press.

Laughlin, C. D., & Throop, J. C. (2003). Experience, Culture and Reality: The Significance of Fisher Information for Understanding the Relationship between Alternative States of Consciousness and the Structures of Reality. *International Journal of Transpersonal Studies*, *22*(1), 7–26. https://doi.org/10.24972/ijts.2003.22.1.7

LeDoux, J. E., & Hirst, W. (Eds.). (1986). *Mind and brain: Dialogues in cognitive neuroscience*. Cambridge University Press.

Lemoine, B. (2022, July 7). *#62 Exposing Google's Sentient AI* [YouTube]. That Tech Show. https://www.youtube.com/watch?v=8hkpLqo6poA

Nagel, T. (2016). *What is it like to be a Bat? / Wie ist es, eine Fledermaus zu sein?: Englisch/Deutsch*. Reclam Verlag.

Nida-Rümelin, M., & O Conaill, D. (2021). Qualia: The Knowledge Argument. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/qualia-knowledge/

OpenAI. (2022). *OpenAI API: text-davinci-002* [Documentation]. GPT-3 Models. https://beta.openai.com

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), Article 6088. https://doi.org/10.1038/323533a0

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., … Le, Q. (2022). *LaMDA: Language Models for Dialog Applications* (arXiv:2201.08239). arXiv. https://doi.org/10.48550/arXiv.2201.08239

Turing, A. M. (1950). 1. Computing Machinery and Intelligence. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Tye, M. (2021). Qualia. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entries/qualia/

Van Gulick, R. (2021). Consciousness. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2021/entries/consciousness/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Wang, P., & Goertzel, B. (2012). *Theoretical Foundations of Artificial General Intelligence*. Springer Science & Business Media.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., & Zettlemoyer, L. (2022). *OPT: Open Pre-trained Transformer Language Models* (arXiv:2205.01068). arXiv. https://doi.org/10.48550/arXiv.2205.01068

Brief communication / Opinion

The problem with AI consciousness: A neurogenetic case against synthetic sentience

Walter, Yoshija*[1, 2, 3]
Zbinden, Lukas[4]

[1] Institute for Management and Digitalization, Department for Business, Kalaidos University of Applied Sciences

[2] Laboratory for Cognitive Neuroscience, Faculty of Mathematics and Natural Sciences, University of Fribourg

[3] Translational Research Center, University Hospital for Psychiatry, University of Bern

[4]ARTORG Center for Biomedical Engineering Research, University of Bern

*yoshija.walter@kalaidos-fh.ch