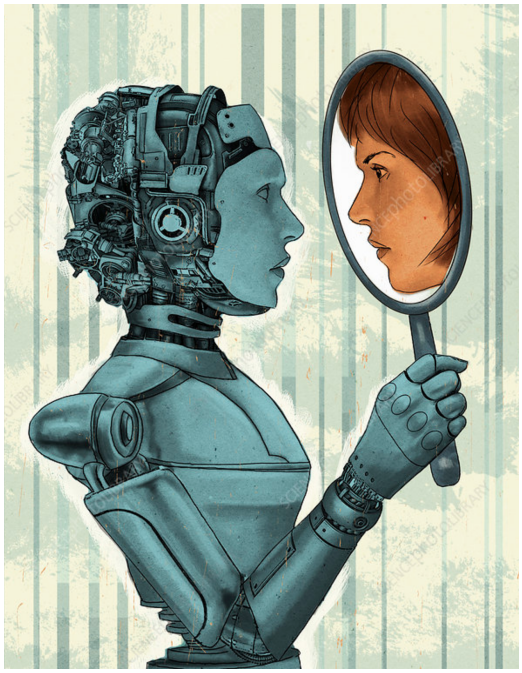
The decision of consciousness in machines is up to us

H.J. Kappen

February 13, 2024



We are currently witnessing a spectacular and unprecedented growth of AI technology. Since the release of ChatGPT last year, our society has profoundly changed in many aspects. These developments are only the beginning, and we can expect even more spectacular AI technology in the (near) future. For instance last month, Google released a research paper [1] showing how an AI system can have a dialogue with a patient as is it were a doctor, collect all the necessary information and reach a diagnosis that is more accurate than human doctors. In addition, this artificial doctor is judged by patients to score better than human doctors in terms of social skills and empathy.

AI systems are already performing well on a Turing test, where a human needs to decide whether he/she is talking to another human or an AI system that simulates conscious behaviour (a consciously behaving machine, CBM). It is likely that in a near future we will not be able to tell the difference. These developments force us to reconsider the question what makes us humans different from machines. Do we posess some special quality that will always distinguish us from AI? Intuitively (and this is a very human trait in itself!) we all tend to think that this is the case. We will admit that in some stylized academic setting the AI will pass the Turing test, but does that really mean that we can build an intimate relation with an AI as we do with humans, become friends? lovers? This seems hard to imagine. Instead, most of us will insist that we have some unique properties that cannot be copied in machines. Then what are these unique human qualities? Is it our consciousness? our moral and ethical values? love? creativity? art? empathy?

Consider consciousness. Consciousness is our first person experience (FPE) that is unique to each and every one of us. We know ourselves to be conscious because we experience our FPE. Strictly speaking, we cannot know whether other persons are conscious because we have no access to their FPE. Nevertheless, we believe that other persons are conscious as well, partly based on their behaviour which shows much similarity with our own behaviour, but crucially because we *assume* that they have a FPE. This is the zombie phenomenon: if we would know that the other has no FPE we would not consider it conscious, regardless of its behaviour. So the question of consciousness in machines is not a question about the behaviour of the machine, but instead whether we (have reasons to) believe that the machine has a FPE.

The dominant view in neuroscience is that we only need concepts from classical physics to understand the brain. In principle this should then also include an understanding of our FPE. In my opinion, there are currently no promising ideas how FPE could occur in a classical brain. After all, the key point about classical physics is that it is based on reductionist notions such as elementary entities (neurons, synapses, vesicles, ion channels) with localized properties that work together to make bigger parts (a neural network). If you use concepts from classical physics to understand the brain, you will end up with a machine. A very complex machine, indeed, but a machine nevertheless with deterministic features such as 'firing of neuron A will cause the opening of ion channel B' or probabilistic descriptions such as 'firing of pre-synaptic

neuron A will cause a post-synaptic potential in neuron B'in 50 % of the time'. Classical physics can be simulated on a digital computer. Therefore, if the classical neuroscience view is correct, it implies that our minds and our FPE can be simulated on a digital computer. This same reasoning holds for AI systems, because current AI technology is also exclusively based on concepts of classical physics. The miraculous behaviour of current large language models is the result of training a model with billions of parameters on enormous data volumes. But in the end there is a software that you can download [2] and can run on any computer. I think that it is absurd to think that a software program can contain the specification of a FPE.

Therefore, there is a good reason to believe that CBMs have no FPE and therefore, regardless of their conscious-like behaviour, we may safely assume that they are not conscious. This is the current situation, and will continue to hold at least as long as AI systems are based on digital computers. With humans, instead, such a rational argumentation is lacking. We do not know whether others have an FPE, so we could consider them conscious or not. Or we could be agnostic about it, saying that we have not way of knowing whether they are conscious or not and. But this is not what we do. We *prefer* to attribute them consciousness, regardless the lack of evidence. This is a choice. We rather live among others who have a FPE than among zombies. In the future, or even right now, the same situation can also apply to machines. People may not accept my argument for lack of FPE in a digital machine that rules out consciousness in a CBM. Instead, they may insist that it is not clear whether the machine has a FPE. And then the situation is similar as with consciousness in other humans: we have a choice, it is up to us, to decide whether the CBM is conscious or not [1].

With classical digital technology, I believe there is a good reason to decide that CBMs are not conscious. But not everyone may agree with me. In the end, the decision whether machines are conscious or not is not a technical issue and cannot be decided scientifically, because even if a machine had a FPE how would we know? Instead, it is a choice that we as humans are free to make. It is up to us to decide whether we want to believe our machines to be conscious or not, as we have done forever with other humans. With humans we prefer them as fellow conscious beings rather then zombies, although there is no evidence either way. With future CBMs we will be in the same situation. We will also have to make a choice with little or no evidence. And that decision will determine how we as a society want to live together with machines. It is up to us.

References

[1] Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*, 2024.

[2] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[3] Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M. Fleming, Chris Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, and Rufin VanRullen. Consciousness in artificial intelligence: Insights from the science of consciousness, 2023.

[1] This recent paper [3] suggests something along these lines. It argues for the decision on the consciousness of an artificial system based on features of its technical design and to what extend these features agree with features proposed by theories of consciousness that deem such features necessary (or sufficient?) for consciousness in any system. Examples of such features are recurrent processing, parallel processing, bottleneck in information flow, generative processing, sparse coding, etc. The authors propose to grant consciousness to an artificial system if it checks a sufficient number of boxes that these theories proclaim to be necessary, although no one has a good understanding of what consciousness actually is. This approach has a the striking similarity with approaches in psychiatry. For instance, psychiatry has no good understanding of what psychosis is at a physiological level, but we do diagnose patients as psychotic based on whether they fullfill a number of criteria, such as the psychiatric DSM5 score.