Why AI Cannot Think
A Theoretical Approach

Daniel M. Feige

In June 2022, Blake Lemoine, then an employee at Google, published a sensational announcement: According to him, LaMDA, the chatbot that he was working on, had developed consciousness and feelings (Wertheimer 2022). As a being with a consciousness, Lemoine said, it should thus be given the same rights as a human person. Lemoine justified this by saying, among other things, that he recognizes a person when he speaks to one. But what is in fact the condition for the possibility of Lemoine's rather astonishing statement being valid? Among the necessary presuppositions behind Lemoine's statement is an idea that has been argued for explicitly by Nick Bostrom: that there might be other intelligent beings than merely carbon-based beings (Bostrom 2013). But such transhumanist ideas are surely still science fiction.

Nevertheless, it is easy to position this idea within a broader emancipatory narrative: Just as humans have understood themselves for too long as categorially different from other animals, whereas they are just one animal species among other animal species, the idea that thinking is exclusive to human beings is a problematic position insofar it merely expresses anthropocentrism. But, in fact, what is at stake here is the analogy between humans and other animals and between humans and artificial intelligence, and it therefore cannot be taken for granted whether this analogy holds true. So again: What is the condition for the possibility of Lemoine's statement being valid? It is then the special sort of knowledge he has as a software engineer? This does not seem to be the case when he merely states that he recognizes a person when he speaks to one. At least one further supposition can be elaborated as the condition for the possibility of Lemoine's statement being valid: that the question of whether something or someone is a person can be understood as a specific *epistemological* question in terms of something that one can notice based on the reactions of one's counterpart.

In what follows, I will argue that Lemoine's statement, and more generally, the idea that we can conceive of anything we currently subsume under the rubric of 'artificial intelligence' as having the power to think, is a deeply flawed and ultimately unintelligible concept (Feige 2024). To show this, I will proceed in three steps. In the

first step (1), I will work out the implicit background of Lemoine's statement: the Turing test, which substitutes an ontological question for an epistemological one. Taking up arguments by Davidson, I will hint at a direction we could go in instead so as to find resources for answers to what is constitutively lacking in an artificial intelligence. In the second step (2), I will draw on the arguments developed by Dreyfus and Cantrell-Smith, who advocate a strong distinction between the operations an artificial intelligence is capable of and what we do insofar we are thinking beings who understand a distinctive feature of the latter as being situated in an intelligible world in which the entities we encounter matter to us. The third and final step (3) will sketch a line of thought that takes recourse to McDowell, who argues for the idea that we can only ascribe thinking to beings that are bearers of a form of life.

On Changing the Subject: The Turing test and the Causal Impact of Reality

Lemoine's statement that LaMDA is a conscious and feeling person lacks any clear conception of what it means to be a conscious and feeling person. But, even worse: 'consciousness' and 'being able to feel' are not the conceptual resources that go together very well with the concept of 'person'. This is the case because we also apply concepts like 'consciousness' and 'being able to feel' to beings that are obviously not persons: for instance, cats, sheep, and dogs. What Lemoine must have meant instead of these categories is a being that is a *self-conscious* being. A self-conscious being not only has conscious episodes as part of its architecture and is not only subjected to affective reactions. It is instead a being that by having a thought or feeling *knows* that it is having this thought or that it is in a specific *emotional* state—since an emotional state is an embodied cognitive state (Goldie 2000).

Operating based on such conceptual confusion and reduction of a full-fledged conception of a person can be attributed to the role that Alan Turing and his Turing test played in the tradition of the development of artificial intelligence with respect to the concept it embodies. In his classic paper on the topic (Turing 1950), Turing proposed substituting the question whether machines are able to think with the question of whether we can notice the difference when confronted with an output in the form of a written text, for instance, on a screen, that is either the output of a machine or was written by a real person. But, in fact, Turing's paper is not an elaborated contribution to the question of whether machines might be able to think someday (Boden 2018, 106ff). It is instead bold conceptual engineering *avant la lettre* (Cappelen 2018). His methodical approach to the question of whether we can attribute the power of thinking to a machine consists of replacing one question with a question that can subjected to some kind of empirical testing. To put it less charitably, one might say that Turing can be said to have not so much engineered concepts

like 'thinking', 'intelligence' et cetera to be made testable in their application to machines, but instead the opposite: *He made it possible to conceive of the power of thinking of human beings in terms of a machine logic*—and also to conceive the mind as a biologically based 'virtual machine' (Boden 2018, 3). The Turing test thus conceptually engineers machines in terms of possessing the ability to think, as well as conceptually engineers ourselves as humans as special kinds of machines.

An obvious fallacy in this sort of substitution of questions is that it is not so much engineers the concepts in question, but instead simply changes the topic. This is true on the level of what question the Turing test is able to respond to: It does not give an answer to the question whether machines can think—a question that Turing hastily dismissed as a crypto-theological question—but also applies to the sorts of question the Turing test asks: It proposes suspending the ontological question of what kind of thing we are dealing with and what powers this thing possesses with respect to the question how we can recognize what sort of thing and what sort of powers we are dealing with. The relationship between ontological questions and epistemological questions, between questions regarding the mode of being of objects and their knowability, is linked to debates on the realist and anti-realist status of this distinction in philosophy (for instance, Putnam 1981, Ch. 3). Is the concept of reality tied to the concept of the knowability of reality or not? What is characteristic of Turing's proposal is that he skips over all these questions and, in a sense, rolls up his sleeves in a computer-scientific way in order to get to work. But those who simply do so drag along with them the errors of what they have sought to overcome; it is not the case that Turing offers a minimal procedural and testable definition of 'thinking' or the like; he instead makes use of a specific and reduced notion of thinking when he conflates epistemological and ontological questions.

A less negative and formal criticism of the Turing test has been put forward by Donald Davidson based on his reflections on the notion of translation and triangulation (Davidson 2004). He somewhat accepts Turing's sharp division between physical and intellectual faculties. The test is designed to test the deceptiveness of subjects by asking whether they can reliably detect the difference when given linguistic utterances in textual form. But the question of whether these linguistic utterances are meaningful is not decided solely on the level of forming syntactically correct sentences. What is relevant for the semantic level, according to Davidson, is that the linguistic expressions have a causal relation to reality. If I state regarding living beings in the world that they are dogs, then I not only have to have acquired the concept of 'dogs'; this acquisition also has to be causally connected in some way to the objects in question—though one can, of course, know what giraffes are by simply being familiar with photos or drawings of them—but those are then 'feature-tracking' depictions (McIver Lopes 2016, 21), not epistemological intermediates. I must have a concept of what it means for a proposition to be true or false and to say something about a reality independent of me.

Davidson's ingenious move is to argue that the holistic character of beliefs—since having one belief means having other or indeterminately many beliefs (Davidson 2001)—does not counter the world-directedness: This network is causally grounded in reality. I do not merely have the concept of a dog because that concept has been established within the framework of a speech community; rather, for Davidson, it is a concept in the first place, and thus something that can be true or false, only because my belief that there is a dog there has been caused by dogs within the framework of a causal history. Under the catchword of radical interpretation (Davidson 1984), which he adopted from his teacher Willard Van Orman Quine, he has also played out this idea for the case in which the persons in question speak different languages; a case that he ultimately distinguishes only by degree from the case in which the persons speak the same language.

What he pits against the Turing test is the case of radical interpretation: In Davidson's view, the question of whether we can be deceived in a human-machine interaction about the fact that the other person is not a human being cannot be tested in the way that Turing set up his test, because we then lose the relations to events, states, and the objects of reality with their impact on linguistic behaviour. Here, we cannot simply subtract the reference to the world entirely and decide the question of reasoning solely on the level of producing syntactically correct English-language sentences; when the counterpart does not have beliefs, desires, et cetera related to the world, there can be no linguistic understanding, because: 'For the object to have a semantics, it must operate in the world in a certain way, and for someone else to grasp those semantics, there must be a three-way interaction among object, interrogator, and a shared world (Davidson 2004, 83f.). Davidson's argument therefore does not show that computers cannot think. It shows solely that the Turing test is not able to answer this question or questions derived from it.

On Having the World Embodied in View: Dreyfus and the Worldlessness of Artificial Intelligence

In Davidson's picture, the concept of the world is a rather thin concept; it ultimately boils down to the causal impact on our holistic web of beliefs. A richer account of the world and its critical consequences for the prospects of an artificial intelligence can be found in the line of critique of Hubert L. Dreyfus and Brian Cantwell-Smith. The basic idea of Dreyfus's criticism (Dreyfus 1972) consists of claiming that artificial intelligence—and with the book from the 1970s he was aiming at first-wave, symbol-processing artificial intelligence—is engineered in such a way that it can never be said to possess the power of thinking. Even if an artificial intelligence is fed with rules for logically correct reasoning, it lacks the world-directness of human thought and action. The identification of our thinking with an explicit set of logical

rules then explains the problem of the correct frame of reference, which cannot be resolved by corresponding semantic networks themselves; accordingly, it is no accident that they produce as many meaningless inferences as they do a multitude of true but uninformative ones.

His criticism stands against the backdrop of Heidegger's legacy: Our being directed towards a world in thinking and acting can be made intelligible only against the backdrop of an unthematic framework of practical understanding that discloses the world as meaningful—which means that Heidegger's holism thus takes a shape distinctly different from Davidson's, which I mentioned in Part I). Thus, we do not register given, context-free facts when we, for instance, hear a storm in the chimney or a car approaching and then infer something from that in order to come to the conclusion that there must be a storm or that there must be a car; rather, we know for the most part what we are dealing with here and also often what needs to be done. In Heidegger's view, a theoretical consideration of the facts of reality is even only possible by virtue of our standing in practical contexts of meaning, and that means that we are not simply confronted with objects to which meanings are somehow 'glued' (Heidegger 1962, §15), but that the objects are instead originally objects opened up in their practical meaningfulness. Such a practical meaningfulness can, however, only exist for beings that do not process data according to logical or statistical laws, but instead have an understanding of themselves and their world. It can only exist for bearers of a self-conscious, free, and reasoning form of life. And the bearer of a life form is not an entity that produces further data from the outside like a forensic instrument by collecting data or applying hardwired schemes of inference.

Brian Cantwell-Smith has renewed this line of criticism with regard to second-wave artificial intelligence (Cantwell-Smith 2019). Neural networks are paradigmatic for developments in this field. The analogies to the human brain are ultimately not theoretically based, but instead presented in the form of a heuristic model. Even in the case that a number of discrete states are produced at the end of the network's activity, they come about differently than in the case of symbol-processing artificial intelligence: they depend on the patterns that these neural networks carve out statically in large amounts of data. It is characteristic for neural networks that they do not necessarily operate with dichotomous states like classic symbol-processing artificial intelligence, in which a sentence is either true or false, an inference either valid or invalid. It is precisely static results with, as it were, ambiguous data that exhibit a logic strikingly alien to our thinking and acting. And therein lies a productive potential of second-wave artificial intelligence: It has a forensic potential to uncover patterns that are unrecognized and even unrecognizable by us, for example, in side-effects in the use of medication or the treatment of cancer. All this can, however, be said without claiming that such a static neural network might 'think'. Cantwell Smith has presented convincing arguments that it does not.

What both first-wave and second-wave artificial intelligence lack is not merely being adequately embodied and engaged in a world. Rather, they lack the possibility of being able to comprehend that world as a world. For: 'most of the computational systems we construct … represent the world in ways that matter to us, not them' (Cantwell Smith 2019, 108)—'all existing AI systems, including contemporary second-wave systems, do not know what they are talking about' (Cantwell Smith 2019, 76). We can use them to find out about the world—but they are not themselves world-aware, as we are in our assessments. To be so, they have to relate to the world in a way that they do not *represent* states of the world alone, but *know* that they are states of the world—which is quite different from picking patterns out of large amounts of data using stochastic methods: A system must be oriented towards what it represents, not just oriented towards its representation. In order to accomplish this, according to Cantwell Smith, what is required is a being that deals practically with objects within the framework of a rich network of patterns of collective actions. Such a being is a self-conscious being, which qua self-consciousness possesses the concept of a belief, which at the same time carries with it the distinction between being-for-true and being-so. We would only entrust our child to a nanny-robot only if we knew that the robot is not concerned with representations of children, but instead with the child in question.

Humans as Bearers of a Form of Life: McDowell on 'Life' as a Transcendental Concept

Both Davidson's critique and that of Dreyfus and Cantrell-Smith aim at current architectures of artificial intelligence. They do not claim that it is logically impossible that someday we will have an artificial intelligence that is able to think and act. In the third and final part of my paper, I will, however, take up a line of thought with a stronger critique of the idea of an artificial intelligence that might be in possession of the power of thinking. Within the framework of recent debates on a philosophical notion of 'life', it can be shown that the idea of an artificial intelligence in possession of the power of thinking is unintelligible.

The concept of 'life' is among the most important topics in contemporary philosophical debates in anthropology, epistemology, and metaethics. The basic idea of these contemporary positions consists of the following (Boyle 2012): we are rational beings insofar we are living beings of a special kind. If the basic insights of these debates, which consist essentially of a combination of the philosophies of Aristotle and Kant, are correct, there is a fundamental limit to what we can meaningfully ascribe to artificial intelligence: They cannot think—or act—because they are not living beings.

John McDowell has drawn on one of Aristotle's ideas to show that rationality is not an additional feature that comes on top of what we share with other living beings; it is rather informed by our way of being alive (McDowell 1996, Lecture VI; Feige 2022, Ch. 2). The respective conception by Aristotle can be called a transformative account of human rationality: What distinguishes us from dumb animals is not an additional feature—Aristotle named nutrition, motion, and perception (Aristotle, 1986)—but the fact that all these faculties are transformed because we are living beings who are responsive to reasons as reasons. Thus, we cannot subtract what is specific to ourselves and then discover what stays the same compared with mere animals. McDowell calls such positions that propose a subtractive account 'highest common factor theories' (McDowell 1998): What the objects in the comparative class in question share remains the same, but in one case another property is added. The transformative account instead denies that there is a common core to be uncovered.

Even if some of the features mentioned by anthropologists were exclusive to human beings, this would still be the wrong sort of answer. It would not bring the difference between humans and dumb animals into view in the right way. Being a rational living being does not mean having biological drives that are then somehow rationally moderated. It instead means that we have our impulses and needs in a self-conscious way and, for this reason, they are also within reach of critical moderation within the framework of the question of how they are to be realized and whether specific ways of realizing them are appropriate or not in other terms than based merely on functional-biological explanations. Whereas the behaviour of mere animals can be fully explained in terms of biological imperatives and the environment in which those animals move in light of those imperatives, such an explanation fails with respect to humans because they do not simply have needs and biological drives. Even supposedly hard-wired facts about ourselves, such as sexual desires, do not in principle silence all other reasons, but are themselves within the reach of rational moderation: we necessarily already shape what looks as if it is merely biologically given.

What does this then have to do with the question whether an artificial intelligence might possess the power of thinking? From an Aristotelian and neo-Aristotelian perspective, the transformative idea identifies limits with respect to the set of beings to which we can meaningfully ascribe something like reason. While there might be rational extra-terrestrial life, using the term 'reason' for an artificial intelligence is meaningless from this perspective. If one follows McDowell's—and, for that matter, Kant's—agenda, however, even if confronted with extra-terrestrial life, the idea that they might embody a very different use of reason fundamentally superior to our reason would be difficult to understand. However, their reason itself would be informed by the particular facts of their being alive and could thus gain different contours—everything else is not only science fiction but also incomprehensible, if authors like McDowell are right. Beings can be thinking (and acting) beings not only

when engaged with a meaningful world. They can only be thinking (and acting) beings insofar as they embody a specific form of life. Despite the fact that McDowell and Dreyfus have been regarded as antipodes in the debate on the role of reason in our engagement with the world (Shear 2013), if these remarks about the role of life are correct, they would underpin rather than counter Dreyfus's line of argumentation. Only living beings can be concerned with the world, because, as bearers of a life form, they are able to distinguish something relevant and do something specific in it.

References

Aristoteles (1986). De Anima (On the Soul). London, Penguin.

Boden, Margaret A. (2018). Artificial Intelligence: A Very Short Introduction. Oxford, Oxford University Press. https://doi.org/10.1093/actrade/9780199602919.001.0001 (all URLs here accessed in August 2023).

Bostrom, Nick (2013). Why I Want to Be a Posthuman When I Grow Up. In: Max More/Natasha Vita-More (Eds.). The Transhumanist Reader. New York, Wiley-Blackwell, 28–53. https://doi.org/10.1002/9781118555927.ch3.

Boyle, Matthew (2012). Essentially Rational Animals. In: Günter Abel/James Conant (Eds.). Rethinking Epistemology. Berlin, De Gruyter, 395–427. https://doi.org/10.1515/9783110277944.395.

Cantwell Smith, Brian (2019). The Promise of Artificial Intelligence: Reckoning and Judgment. Cambridge, MA, The MIT Press. https://doi.org/10.7551/mitpress/12385.001.0001.

Cappelen, Herman (2018). Fixing Language: An Essay on Conceptual Engineering. Oxford, Oxford University Press. https://doi.org/10.1093/oso/9780198814719.001.0001.

Davidson, Donald (1984). Radical Interpretation. In: Donald Davidson. Inquiries into Truth and Interpretation. Oxford, Clarendon, 125–39. https://doi.org/10.1093/0199246297.003.0009.

Davidson, Donald (2001). Rational Animals. In: Donald Davidson. Subjective, Intersubjective, Objective. Oxford, Oxford University Press, 95–105. https://doi.org/10.1093/0198237537.003.0007.

Davidson, Donald (2004). Turing's Test. In: Donald Davidson. Problems of Rationality. Oxford, Oxford University Press, 77–86. https://doi.org/10.1093/0198237545.003.0005.

Dreyfus, Hubert L. (1972). What Computers Can't Do: On Artificial Reason. New York, Harper & Row.

Feige, Daniel M. (2022). Die Natur des Menschen. Eine dialektische Anthropologie. Berlin, Suhrkamp. https://doi.org/10.1017/hgl.2023.3.

Feige, Daniel M. (2024). Gegen-Digitalisierung. Ästhetik, Rationalität und Kritik. Berlin, Suhrkamp, in preparation.

Goldie, Peter (2000). The Emotions: A Philosophical Exploration. Oxford, Clarendon. https://doi.org/10.1093/0199253048.001.0001.

Heidegger, Martin (1962). Being and Time. Oxford, Basil Blackwell.

McDowell, John (1996). Mind and World. Cambridge, MA, Harvard University Press. https://doi.org/10.2307/j.ctvjghtzj.

McDowell, John (1998). Criteria, Defeasibility, and Knowledge. In: John McDowell. Meaning, Knowledge, and Reality. Cambridge, MA, Harvard University Press, 369–94. https://doi.org/10.2307/j.ctv22jntgn.

McIver Lopes, Dominic (2016). Four Arts of Photography: An Essay in Philosophy. Malden, MA, Wiley & Sons. https://doi.org/10.1002/9781119053194.

Putnam, Hilary (1981). Reason, Truth and History. Cambridge, Cambridge University Press. https://doi.org/10.1017/cbo9780511625398.

Schear, Joseph K. (2013) (Ed). Mind, Reason, and Being-in-the-World: The McDowell-Dreyfus-Debate. London, Routledge. https://doi.org/10.4324/9780203076316.

Turing, Alan M. (1950). Computing Machinery and Intelligence. Mind LXI (236), 433–60. https://doi.org/10.1093/oso/9780198250791.003.0017.

Wertheimer, Tiffany (2022). Blake Lemoine: Google Fires Engineer Who Said AI Tech Has Feelings. BBC News, 23 July 2022. https://www.bbc.com/news/technology-62275326.