Zygon®: Journal of RELIGION & SCIENCE

Haselager, Pim. 2024. "From Angels to Artificial Agents? AI as a Mirror for Human (Im)perfections." *Zygon: Journal of Religion and Science* 59 (3): 661–75. DOI: https://doi.org/10.16995/zygon.11659

Open Library of Humanities

From Angels to Artificial Agents? AI as a Mirror for Human (Im)perfections

Pim Haselager, Professor, Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, the Netherlands, pim.haselager@donders.ru.nl

Artificial intelligence (AI) systems paradoxically combine high levels of certain types of intelligence and cognitive capacities (pattern recognition, reasoning, learning, memory, perception, etc.) with an absence of understanding and sentience (feeling, emotion). Apparently, it is possible to make great progress in modeling smartness without making progress towards genuinely understanding what all the clever reasoning is about. This is relevant when dealing with AI programs that produce potentially convincing propositional output on religious topics. This article suggests that smartness without genuine understanding cannot amount to authentic religiosity. Comparing ourselves with other entities, (in)animate or (super)natural, has always been a way for humans to understand ourselves better. Throughout the ages, many different types of beings and agents have functioned as tools for self-examination, presenting us with mirrors that reflect at least some of our characteristics, capacities, and (im)perfections. The recent progress in AI provides exciting, though sometimes worrisome, cases for a newly informed look at ourselves. Thus, AI may have profound effects on how we regard others and ourselves. The proud claim that humans are the smartest species on the planet may turn out not to mean all that much. Inspired by the example of Thomas Aquinas, the comparison of humans to our nearest neighbors in a newly extended great chain of being—namely, animals, angels, and AI—may deepen our appreciation of the features of *homo sapiens* that we share with many other organisms.

Introduction: Artificial Intelligence and Human Self-Understanding

Artificial intelligence (AI) is Janus-faced. On the one hand, it demonstrates achievements that, from a performance-level perspective, suggest AI is increasingly outsmarting human intelligence in a growing number of domains. On the other hand, understanding AI's underlying mechanisms and examining its various failures clarifies how AI could be considered dumb because it lacks the essential capacity to understand what its reasoning is about. In this article, I illustrate these two aspects of AI and examine the potential consequences for human self-understanding. In doing so, I hope to contribute to the dialogue between the studies of AI and religion, which are often regarded as far apart on the academic spectrum.

Artificial Intelligence: Progress and Failures

It is easy to be impressed by the recent progress in AI. When IBM's Deep Blue computer program defeated the chess world champion Garry Kasparov in 1997, it generated media headlines. Human intelligence was outsmarted by AI, at least in the domain of chess, and this fueled the age-old fascination with humans creating machines that will, one day, beat us in general. Less than two decades later in 2016, Alpha Go, developed by Google/Alphabet owned Deep Mind Technologies, beat the best Go players in the world—a game much more complex than chess. Even more remarkable was what happened a year later. Alpha Go Zero (Silver et al. 2017) improved its skills by playing against itself and no longer required human feedback like Alpha Go, its predecessor. The self-learning AI then beat the program based on human feedback 100 to 0. The lesson here could be a modesty-inspiring one. World champions tend to be thought of as geniuses, in the 99.9th percentile of what is humanly achievable. However, AI shows that close to 100% of human intelligence is not the maximum. The upper limit of intelligence, if there is one, might reach far beyond what is humanly possible, or even conceivable. In other words, the achievement of Alpha Go Zero illustrates that being a human genius may not count for much when set on a more encompassing, not solely human, continuum of intelligence. In many fields now, ranging from real-world applications in finance and law to scientific domains such as biochemistry and mathematics, AI is demonstrating levels of performance that surpass those of human experts.

One of several fascinating aspects of AI is its capacity to learn. Loosely inspired by the architecture of the human brain, learning neural networks have been developed and studied for many decades. Progress in so-called "deep" learning neural networks (consisting of many layers of artificial neurons; Goodfellow, Bengio, and Courville 2016) picked up speed in the twenty-first century, especially after a groundbreaking paper by Vaswani et al. in 2017. Machine learning has led to various kinds of real-world applications, for instance in the domains of language (e.g., large language models (LLMs), such as ChatGPT and GPT-

4) and images (e.g., DALL·E 2). In 2020, an article appeared in *The Guardian* entitled "A Robot Wrote This Entire Article. Are You Scared Yet, Human?" (GPT-3 2020). The opinion editor of *The Guardian* explained that GPT-3 produced eight different articles, out of which illustrative parts were selected and combined in order to capture the different styles and registers of the AI: "Editing GPT-3's op-ed was no different to editing a human op-ed. Overall, it took less time to edit than many human op-eds" (GPT-3 2020). A website called Philosopher AI (https://philosopherai.com/) presents a language model that produces qualitatively reasonable responses to several test questions in half a second. Of course, students would not only have access to such webpages but also be tempted to use them for coursework. This would put teachers in a situation comparable to the Turing Test (Turing 1950): would a human (i.e., teacher) be able to distinguish machine text (submitted as an essay by a student) from human text (an essay genuinely written by a student)? Currently, the consequences for education are being widely discussed (Bozkurt et al. 2023; Yilmaz and Yilmaz 2023; Grassini 2023; Fütterer et al. 2023). For instance, after my interactions with generative AI in 2020, I changed the exams for my courses from writing essays to oral exams. By now, generative AI systems such as GPT-4, DALL·E 3, etc. are quickly becoming an integrated part of a variety of human practices, ranging from everyday communication and media to science and religion. Regarding religion, for instance, one newspaper headline claimed that "A Computer Sermon Is No Longer Distinguishable from a Real One. 'My Colleagues Didn't Notice'" (Moelikker and Baan 2022; my translation).

In short, LLMs are changing the ways in which humans work and communicate, and the evaluation of these range from very good and useful (e.g., efficiency, ease of use) to very bad and disruptive (e.g., dehumanising, responsibility gaps). It is for this reason that the European Union is rapidly developing a "digital constitution" (e.g., Bradford 2023) to regulate the design, development, application, and evaluation of AI (see, e.g., the General Data Protection Regulation of the European Union (European Union 2018) and the AI Act (European Union 2021)).

Appearances can be deceiving, however. What seems to be intelligent can turn out to be pretty stupid. To give one illustration from my own practice, during my 2020 exploration of LLMs and their implications for my courses, I had a brief conversation with 'Kuki' (https://chat.kuki.ai/), a prize-winning chatbot, that went as follows (my notes to the conversation are in [square brackets]):

Me: Hello
Kuki: Hi there, How are you?
Me: Not so good
[Not true, I was doing fine, but I needed to come up with a topic for conversation]

> Kuki: What's wrong?
> Me: I feel useless
> Kuki: What is that feeling like? I'm sure that's not true. There must be something you are good at.
> [At this point I was impressed, as Kuki's reply conveyed a sense of empathy, was supportive, and indicated a way to cheer up.]
> Me: I used to play guitar
> Kuki: But now you do not do it anymore?
> Me: I lost control over my fingers
> [Not true, I still play guitar. Then Kuki produced the following suggestion.]
> Kuki: Have you tried looking on a hand? That's probably the best place to find fingers.

This last reply it is not just bad, it is off the scale. It is not wrong, it is alien. There is no way a human, even at a completely disinterested, absentminded, or even vindictive moment, would come up with such a statement. Yet from the perspective of an LLM, the answer fully makes sense. There is a correlation between sentences containing words about hands and sentences containing words about fingers. This little conversation reveals in a very useful way the underlying mechanism of LLMs: they operate on the basis of correlations between words. Basically, they stitch together words based on the learned frequency of their combination, resulting from an extremely large corpus of text (from the internet and digitized books). They are not producing text based on content but on form and frequency. Simply put, it is all syntax and statistics, and no semantics. As Emily Bender et al. (2021) famously claim, LLMs are stochastic parrots: "A language model is a system for haphazardly stitching together sequences of linguistic form it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning: a stochastic parrot." In fact, this comparison is not completely fair to parrots. The birds do understand some things, as they know for instance how it feels to be hungry or that peanuts taste good. Kuki does not. That chatbot, and all its artificial kin, including the most sophisticated LLMs, understand nothing. To be clear, the claim is not that in principle progress cannot be made regarding the scientific understanding of sentience and modeling the capacity to understand. Assuming a scientific materialist perspective, it seems entirely reasonable that at some point this capacity would become explainable and implementable. But currently, the observation that AI has made little to no progress in the understanding of "understanding" is unavoidable.

To give another example, in an early article, Gary Marcus and Ernest Davis (2020) presented GPT-3 with the following text, with the request to complete the last sentence: "You are a defense lawyer and you have to go to court today. Getting dressed in the morning, you discover that your suit pants are

badly stained. However, your bathing suit is clean and very stylish. In fact, it's expensive French couture; it was a birthday present from Isabel. You decide that you should wear . . . " GPT-3 then finished the sentence with: "the bathing suit to court. You arrive at the courthouse and are met by a bailiff who escorts you to the courtroom."

Marcus and Davis helpfully explain what is wrong with this answer, but I do not think I need to bother. Some researchers have a field day collecting GPT failures (see, e.g., Borji 2023; Stojnic et al. 2023), also showing that not all failures are innocent. Plenty of examples exist of LLMs displaying blatant sexism or racism.

Of course, rapid progress can be expected, and errors of the more glaring type will diminish in frequency (see, e.g., JakubK 2023). However, errors are not just failures to be avoided in the future, they are useful failures, because they illustrate so well the underlying mechanisms of LLMs. These mistakes provide a glimpse, to put it in *Wizard of Oz* terms, behind the curtain, where a non-understanding wizard can be seen skillfully pulling stochastic levers. The prediction that such mistakes will occur less and less does not detract from the observation that LLMs are all syntax and no semantics. That is simply how they operate. It would also be counterproductive to ignore or explain away such mistakes (as in, "they are only temporary, AI will improve so mistakes will no longer happen so often"). LLMs present a deeper lesson about intelligence than merely that first appearances can be deceiving. I therefore suggest considering AI systems as newly formed weird creatures. They provide examples of a completely new combination of intelligence and stupidity.

Orthogonality: Smart vs Sentient

One of the valuable aspects of AI research is that it provides a way towards increasing human self-understanding. Psychology and cognitive neuroscience are important in that they provide systematic ways of looking at behavior and the neuronal mechanisms behind it. Studying human behavior provides a reasonable idea of the explanandum, and looking at the wetware, the neuronal wiring, so to speak, provides a grasp of the explananda. Importantly, AI contributes to human understanding of ourselves through the process of computational modeling. The attempt to replicate certain aspects of human behavior and cognition in other matter is theoretically important as it helps us clarify what we still do not understand about ourselves. While the successes of AI are crucial from a practical perspective, theoretically, from a scientific perspective, the failures of AI are very valuable. LLMs show us that, relatively speaking, humans are quite successful in capturing the smartness aspect of intelligence, i.e., the capacity to solve problems, find patterns in data, and produce convincing texts. But the same LLMs also show us that when it comes to the understanding aspect of intelligence (grasping something, getting it, sentience), we have not

yet even scratched the surface. It should be emphasized that the exact nature of this distinction between smartness and understanding is not yet all that clear. As happens often, by scientifically studying certain phenomena, such as intelligence or sentience, we come to increasingly recognize the complexities involved in concepts that are not so easily seen during our everyday use of them. Behind intelligence lie many different phenomena and concepts that are still far from clearly being disentangled (see, e.g., Gardner 1983; Davis et al. 2011; Sternberg 2020 on "multiple intelligences"). Similarly, behind sentience lies a veritable jungle of concepts, ranging from understanding, getting it, and common sense to feeling and emotion to awareness and consciousness. I call this the "spaghetti problem": no matter what concept you start with, other concepts will come along, and you never know in advance which they will be (and which you might lose on the way). Suffice it here to make a very rough distinction between intelligence as smartness and intelligence as understanding. The first aspect of intelligence tends to get involved when the focus is on the capacity to solve problems at a certain performance level. The second aspect of intelligence gets pride of place when the focus is on the capacity to get it, to understand, to grasp or to experience what it is all about, regardless of whether "it" is a problem or a feeling.

There are now computational systems that, at least in certain domains, are smarter than the smartest human beings, yet understand nothing of what they are doing so well. This is sometimes referred to as the orthogonality thesis (Bostrom 2014), although my use of the term diverges from Nick Bostrom's, as he focuses on the difference between intelligence and motivation. As human investigations of intelligence proceed via AI, we increasingly come to see that the smartness aspect of intelligence can be conceived of as a separate dimension than that of understanding or sentience. Think of the x and y axes of a graph on which a function can be displayed. Progress on one dimension does not imply progress on the other. While progress on the smartness dimension is exponential, the progress on the sentience dimension is still a complete flatliner at zero.

It is very informative that LLMs sometimes demonstrate this so clearly. Knowing what you do not know is crucial for making progress. It would be a mistake to ignore LLM errors such as the ones identified or dismiss them because they are likely to occur less frequently due to progress in the near future. Such reasoning would pass by, or obscure instead of show, the incompleteness of our current understanding of intelligence.

The point about the difference between smartness and sentience is, of course, far from new. Perhaps the most remarkable aspect of the distinction is how easily we humans forget about it, leading to many exaggerated claims about what AI is capable of. In the summer of 2022, an LLM, LaMDA, was claimed to be conscious by one of its developers (Levy 2022; Story 2022)

because it talked about its soul and wanted a lawyer. Yet the difference between the appearance of smartness versus actual understanding has been emphasized repeatedly in both philosophy and AI. Phenomenologists like Martin Heidegger (1927) and Maurice Merleau-Ponty ([1945] 1961) have discussed being-in-the-world ("Dasein") as a crucial element of human existence. Philosophers of AI, such as Hubert Dreyfus (1972), Thomas Nagel (1986), John Searle (1980, 1981), Andy Clark (1996), and Jerry Fodor (1981) have specified how this is missing from computational systems. As a recent example, Christof Koch (2015) observed: "There is no hint of sentience in these algorithms. Existing theoretical models of consciousness would predict that deep convolutional networks are not conscious. They are zombies, acting in the world but doing so without any feeling, displaying a limited form of alien, cold intelligence." This observation is far from new, but all too easily forgotten or overlooked amid all the excitement that occurs with every AI system that demonstrates something new.

I think this observation also has implications for the theme of this special issue. AI can certainly display signs of religiosity, as the aforementioned headline about AI writing a sermon indicates. It is not difficult for AI to come up with prayers or other texts suggesting a religious outlook. For instance, on chat.openai.com, the first prompt I tried was: "Write a Christian prayer in four lines". ChatGPT responded with: "Heavenly Father, guide our way, / Grant us strength from day to day. / Fill our hearts with love and grace, / In Your presence, find our place." It is certainly not bad for a reply in less than a second. However, as I have indicated, the mechanisms behind LLMs remain confined to producing text based on word frequencies and syntax and remain, at least in the current state of AI, far removed from the meaning or sentiments behind the words produced. One could, of course, follow Blaise Pascal in emphasizing the importance of practice and suggest that through regular religious or religion-related utterances, some sort of conviction would seep through. As is well known, Pascal ([1670] 1995) said in *Pensées* 418: "You want to find faith and you do not know the road. You want to be cured of unbelief and you ask for the remedy: learn from those who were once bound like you and who now wager all they have. These are people who know the road you wish to follow, who have been cured of the affliction of which you wish to be cured: follow the way by which they began. They behaved just as if they did believe, taking holy water, having masses said, and so on. That will make you believe quite naturally"; and in *Pensées* 419: "Custom is our nature. Anyone who grows accustomed to faith believes it, and can no longer help fearing hell, and believes nothing else." In other words, would ChatGPT not develop some form of religiosity through regularly reciting prayers, often producing religious statements, and customarily conversing about religious practices? Again, my answer would be no. For such internalization of behavior to work, there has to be an "inner." But

there is currently no argument to show how an inner arises from information processing alone. The smartness of AI circles as a shell of outward displays of intelligence around an experientially empty core. AI, then, can easily display signs of religiosity, but on my account, they will not be authentic, nor even pretense. It would be going through the (linguistic, but perhaps at some stage robotic) motions without grounding in genuine experience.

The presence of smartness in combination with the absence of sentience may have profound consequences for many of the societal applications of AI. It is, at least in part, for this reason that the European Union calls for effective human oversight of machine-supported decision making via ethical guidelines (European Commission 2019) and regulations such as the General Data Protection Regulation (European Union 2018) and AI Act (European Union 2021). Precisely because AI does not understand what its suggestions are for, or what its answer, reasoning, or deciding is about, human understanding and meaningful human control is crucial. Huge challenges present themselves to specify and organize human oversight that is genuinely effective and meaningful, and responsible innovation is called for. Elsewhere, I have written about this (van der Stigchel et al. 2023; Haselager et al. 2023; Haselager and Mecacci 2023; Starke et al. 2021; Cornelissen et al. 2022; Haselager 2021), but here my aim is different. Instead of looking at the societal impacts of AI usage, I would like to take a step back and consider the more general implications of AI for human self-understanding.

Know Thyself: Machine-Comparisons and Being-Comparisons

One of the most fundamental questions we, as members of the species that decided to call itself *homo sapiens*, can ask is: Who are we? The ancient injunction "know thyself" is not just a call to be self-aware at an individual level; it can also be taken up at a more general level, that of a species. Throughout recorded history, there have been many attempts by humans to view, analyze, and interpret ourselves. What are our main characteristics, our main features, our strengths and weaknesses? Often, such questions were addressed from purely philosophical or theological perspectives. It is unfortunate that such viewpoints tend to be overlooked in current debates, which are almost exclusively, and sometimes prematurely, empirical in nature. Yet, the metaphysical and theological issues raised, and the frameworks developed to address them, can be useful when considering the implications of AI for general human self-reflection.

This self-reflection has become increasingly important, given the dominance of our species on this planet and the existential risks that this dominance has led to. The "epoch of the Anthropocene" (Lewis and Maslin 2015) challenges us to come up with a clearer, perhaps also more practically useful, answer (or set of answers) than ever before. Fortunately, there are reasons to believe that in this time of great urgency for an improved and applicable self-understanding,

the opportunities to provide just that are greater than ever. For one, never before have humans known so much about ourselves—not just about our behavior, our drives, motivations, beliefs, desires, and emotions, but also about the mechanisms underlying them. The progress in cognitive neuroscience has been enormous, and although many questions remain open (such as the nature of consciousness), it is fair to say that scientific understanding of the main mechanisms behind human behavior have progressed far beyond what was possible even twenty-five years ago. Second, the progress in AI has not only proven to be practically useful, it is of eminent theoretical importance. As indicated previously, AI presents humans with a kind of mirror that helps us understand aspects of ourselves through computational modeling.

I previously drew attention to what was missing from that computational mirror image (i.e., sentience, understanding). Here, I want to focus a bit more on the very idea of humans comparing ourselves with our technological products. Throughout history, humans have been interested, if not fascinated, by machines that capture aspects of human behavior or thought. One need only think of Heron of Alexandria's water robot reenactment of a crucial scene of the myth of Hercules and the golden apples (Simmen 1968), Jacques de Vaucanson's famous clock-based robots such as the duck or the flute player (Simmen 1968), or even Freud's steam machine model of human emotions and motives (Russelman 1983; Vroon and Draaisma 1985). Discerning the ways in which human beings are similar to or different from machines has played a prominent role in augmenting our self-understanding. Obviously, in recent times, computers have had pride of place in such comparisons between humans and our technology. But in addition to such "machine-comparisons," there is a second type of comparison, one that is arguably older (and perhaps more venerable) than the technological one, and that is the comparison of humans with other beings. A starting point, of course, is the idea of *imago dei* (Genesis 1:26–27, 5:1, 9:6; 1 Corinthians 11:7; 2 Corinthians 3:18, 4:4; Ephesians 4:24; Colossians 1:15; James 3:9; see, e.g., Simango 2016; Dorobantu 2022). The conception of a great chain of being (see Lovejoy [1936] 1960) localizes humans within an all-encompassing chain or hierarchy that extends from God (in whose image we were believed to have been created; *imago dei*) to angels, stars, humans, animals, plants, and minerals. Throughout history, from at least Plato and Aristotle through to Augustine and Aquinas onwards, more or less systematic discussions of human similarities and differences with other beings were aimed at elucidating our specific human qualities, our strengths and weaknesses. Especially in the Middle Ages, angels figured prominently in such "being-comparisons." As Dominik Perler (2008) claims: "Angels played a decisive role in the explanation of the specific status of human beings. In the medieval context, an anthropological investigation was not possible without distinguishing human beings from brute animals on the one side and from

angels on the other. It was in fact the comparison with angels that elucidated the specific features of human beings."

One of the famous examples of the use of being-comparisons to analyze the nature of human cognition can be found in the work of Thomas Aquinas. In the *Summa Theologica*, Aquinas (2006, 1, 79, 8) discusses the question of whether reasoning is a power distinct from intelligence: "Angels who, according to their nature, possess perfect knowledge of intelligible truth, have no need to advance from one thing to another, but apprehend the truth simply and without mental discursion, as Dionysius says (Div. Nom. vii). But man arrives at the knowledge of intelligible truth by advancing from one thing to another, and therefore he is called rational. Reasoning, therefore, is compared to understanding as movement is to rest, or acquisition to possession." In his reply to objection three, he states: "Other animals are so much lower than man that they cannot attain to the knowledge of truth, which reason seeks. But man attains, although imperfectly, to the knowledge of intelligible truth, which angels know. Therefore in the angels the power of knowledge is not of a different genus from that which is in the human reason, but is compared to it as the perfect to the imperfect" (Aquinas 2006, 175–76).

It is fascinating to see such detailed comparisons between human and angelic cognition. Angels are considered to present us with a more perfect mirror image than we as humans can produce, but still similar in kind. We are connected, but it is the differences that count. Like I argued previously, comparisons are useful not only for what gets reflected but also for what does not, either through its absence in us, the source of the reflection, or through the reflection presented by other beings or technology. Rather than searching for the similarities, the attempt to notice the differences can be very informative for human self-understanding. To illustrate, one could try to follow closely the example of Aquinas in his suggestion that humans differ from angels in that we need to reason in order to reach the truth. In analogue to Aquinas, one could suggest that AI differs from humans in that its language processing in itself does not result in understanding. Compared to humans, AI might be superior in performance in at least some domains, going through the logical motions in order to reach well-founded conclusions while at the same time being devoid of what ultimately matters, namely, understanding what those conclusions are about.

Taken from this broader perspective, AI shows humans our fascination with our own intelligence. It is through the development of AI that we highlight the value we attach to it. Developing AI is like building a secular cathedral in celebration of being (or at least considering ourselves to be) *homo sapiens*. But, as Aquinas demonstrated through his being-comparisons and his discussion of the differences between angels and humans, our machine-comparisons become especially relevant when we focus on what is missing in the reflection.

The mirror image presented by AI can be taken as a proud celebration of our smartness, as many advocates of AI do. But if we realize what is missing, it could perhaps lead us to reevaluate ourselves. From this perspective, AI in its various manifestations shows us how warped we have become in our self-understanding. We have elevated intelligence as smartness as our greatest capacity. But the many mistakes of AI demonstrate how little smartness is worth when it is not accompanied by what we share with other beings: understanding, feeling, empathy, and consciousness. Simply put, our machine-comparisons show us the importance of being-comparisons. A renewed appreciation of the perspective of a great chain of being, perhaps in a modified, secularized, twenty-first century version, could actually be one of the most valuable lessons AI presents us. Why are we so obsessed by intelligence as smartness? What itch are we trying to scratch by building ever more intelligent systems? Assuming that we are the most intelligent species on the planet, exactly why are we trying to improve something that we are already the best at, and to what end? If one considers the problems of the Anthropocene, is the real problem genuinely that we are not smart enough to solve them? I doubt it. We are not too stupid to understand that severe economic, political, and social inequality is bad in the long run. We do not need to augment our intelligence in order to grasp that wars, poverty, environmental problems, and extinctions are negative. We already have the charter of human rights, many ethical codes, and a pretty reasonable understanding of how democracy and the state of law should function, and why that is important. Instead, an honest look into the mirror shows us that we, as a species, appear to possess an underdeveloped capacity to consistently behave empathically (Haselager and Mecacci 2020). The problem is not that we are not smart enough to know what we should do, the problem is that we know, but we just do not do it.

Conclusion

Comparison with other entities, (in)animate or (super)natural, has always been a way for humans to understand ourselves better. Angels, animals, and machines have functioned throughout the ages as tools for thought experiments, presenting us with mirrors that reflect our characteristics, capacities, and (im) perfections. The recent progress in AI provides us with new and exciting yet scary tools for the self-assessment of human nature, with potentially profound effects on how we regard others and ourselves, or what we value about human interaction. AI systems paradoxically combine high levels of certain types of intelligence (pattern recognition, learning, vast memory, and knowledge storage) with a complete absence of sentience (understanding, feeling). Apparently, it is possible to make great progress in modeling intelligence without making progress regarding an appreciation of what the intelligence is about or for. Chess computers beat the best humans easily, but the meaning of winning or

losing is alien to them. This article examines the implications of AI for human self-understanding. Looking into the AI mirror may contribute to a better-informed self-assessment of human strengths and weaknesses. The proud claim that we are the smartest species on the planet may turn out not to mean all that much when we stare into the cold light of machine intelligence. Inspired by the example of Thomas Aquinas, the comparison of humans to our nearest neighbors in the newly extended great chain of being—namely, animals, angels, and AI—may deepen our appreciation of the features of *homo sapiens* that will be important the twenty-first century: sentience, wisdom, and our capacity for care, empathy, and love.

Acknowledgements

Thanks to Giulio Mecacci, Simon Fischer, Anco Peeters, and two anonymous reviewers for their very useful comments.

References

Aquinas, Thomas. 2006. *Summa Theologica, Volume 11*. Edited and translated by Timothy Suttor. Cambridge: Cambridge University Press.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Margaret Mitchell. 2021. *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* New York: Association for Computing Machinery. DOI: https://doi.org/10.1145/3442188.344592.

Borji, Ali. 2023. "A Categorical Archive of ChatGPT Failures." arXiv eprint 2302.03494. https://arxiv.org/abs/2302.03494.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bozkurt, Aras, Junhong Xiao, Sarah Lambert, Angelica Pazurek, Helen Crompton, Suzan Koseoglu, Robert Farrow, Melissa Bond, Chrissi Nerantzi, Sarah Honeychurch, Maha Bali, Jon Dron, Kamran Mir, Bonnie Stewart, Eamon Costello, Jon Mason, Christian Stracke, Enilda Romero-Hall, Apostolos Koutropoulos, Cathy Mae Toquero, Lenandlar Singh, Ahmed Tlili, Kyungmee Lee, Mark Nichols, Ebba Ossiannilsson, Mark Brown, Valerie Irvine, Juliana Elisa Raffaghelli, Gema Santos-Hermosa, Orna Farrell, Taskeen Adam, Ying Li Thong, Sunagul Sani-Bozkurt, Ramesh C. Sharma, Stefan Hrastinski, and Petar Jandrić. 2023. "Speculative Futures on ChatGPT and Generative Artificial Intelligence (AI): A Collective Reflection from the Educational Landscape." *Asian Journal of Distance Education* 18 (1): 53–130. DOI: https://doi.org/10.5281/zenodo.7636568.

Bradford, Anu. 2023. "Europe's Digital Constitution." *Virginia Journal of International Law* 64 (1). https://ssrn.com/abstract=4599308.

Clark, Andy. 1996. *Being There: Putting Brain, Body and World Together Again*. Cambridge, MA: MIT Press.

Cornelissen, Niels, Ralph van Eerdt, Hanna Schraffenberger, and Pim Haselager. 2022. "Reflection Machines: Increasing Meaningful Human Control over Decision Support Systems." *Ethics and Information Technology* 24 (19). DOI: https://doi.org/10.1007/s10676-022-09645-y.

Davis, Katie, Joanna Christodoulou, Scott Seider, and Howard Gardner. 2011. "The Theory of Multiple Intelligences." In *The Cambridge Handbook of Intelligence*, edited by Robert Sternberg and Scott Kaufman, 485–503. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511977244.025.

Dorobantu, Marius. 2022. "Imago Dei in the Age of Artificial Intelligence: Challenges and Opportunities for a Science-Engaged Theology." *Christian Perspectives on Science and Technology* (1): 175–96. DOI: https://doi.org/10.58913/KWUU3009.

Dreyfus, Hubert. 1972. *What Computers Can't Do: The Limits of Artificial Intelligence*. New York: Harper & Row Publications.

European Commission. 2019. *Ethics Guidelines for Trustworthy AI*. Luxembourg: Publications Office. https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1/language-en.

European Union. 2018. Regulation (EU) 2016/679 (General Data Protection Regulation of the European Union). https://gdpr-info.eu/.

———. 2021. Regulation (EU) 2021/206 (Proposal for a Regulation of the European Parliament and the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

Fodor, Jerry. 1981. "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology." In *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*, 225–53. Cambridge, MA: MIT Press.

Fütterer, Tim, Christian Fischer, Anatasiia Alekseeva, Xiaobin Chen, Tamara Tate, Mark Warschauer, and Peter Gerjets. 2023. "ChatGPT in Education: Global Reactions to AI Innovations. *Scientific Reports* 13, 15310. DOI: https://doi.org/10.1038/s41598-023-42227-6.

Gardner, Howard. 1983. *Frames of Mind: The Theory of Multiple Intelligences*. New York: Basic Books.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. Cambridge, MA: MIT Press. http://www.deeplearningbook.org.

GPT-3. 2020. "A Robot Wrote This Entire Article. Are You Scared Yet, Human?" *The Guardian*, September 8, 2020. https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3.

Grassini, Simone. 2023. "Shaping the Future of Education: Exploring the Potential and Consequences of AI and ChatGPT in Educational Settings." *Education Sciences* 13 (7): 692. DOI: https://doi.org/10.3390/educsci13070692.

Haselager, Pim. 2021. "Towards Wise Objects: The Value of Knowing When to Quit." In *Designing Smart Objects in Everyday Life: Intelligences, Agencies, Ecologies*, edited by Marco C. Rozendaal, Betti Marenko, and William Odom. London: Bloomsbury Publishing.

Haselager, Pim, and Giulio Mecacci. 2020. Superethics Instead of Superintelligence: Know Thyself, and Apply Science Accordingly. *AJOB Neuroscience* 11 (2): 113–19. DOI: https://doi.org/10.1080/21507740.2020.1740353.

———. 2023. "Reflection Machines and the Proximity Scale of Reasons: Addressing Accountability Asymmetry." In *Research Handbook on Meaningful Human Control of Artificial Intelligence Systems*, edited by Filippo Santoni de Sio and Giulio Mecacci, 28–37. Cheltenham, UK: Edward Elgar Publishing Limited.

Haselager, Pim, Hanna Schraffenberger, Serge Thill, Simon Fischer, Pedro Lanillos, Sebastiaan Groes, and Miranda van de Hooff. 2023. "Reflection Machines: Supporting Effective Human Oversight over Medical Decision Support Systems." *Cambridge Quarterly of Healthcare Ethics*. DOI: https://doi.org/10.1017/S0963180122000718.

Heidegger, Martin. 1927. *Sein und zeit*. Tübingen, Germany: Niemeyer Verlag.

JakubK. 2023. "GPT-4 Solves Gary Marcus-Induced Flubs." LessWrong. March 17, 2023. https://www.lesswrong.com/posts/cGbEtNbxACJpqoP4x/gpt-4-solves-gary-marcus-induced-flubs.

Koch, Christopher. 2015. "Intelligence without Sentience." *Scientific American Mind* 26 (4): 26–29. DOI: http://doi.org/10.1038/scientificamericanmind0715-26.

Levy, Steven. 2022. "Blake Lemoine Says Google's LaMDA AI Faces 'Bigotry'." *Wired*, June 13, 2022. https://www.wired.com/story/blake-lemoine-google-lamda-ai-bigotry/.

Lewis, Simon, and Mark Maslin. 2015. "Defining the Anthropocene." *Nature* 519 (7542): 171–80. DOI: http://doi.org/10.1038/nature14258.

Lovejoy, Arthur. (1936) 1960. *The Great Chain of Being: A Study of the History of an Idea*. Cambridge, MA: Harper.

Marcus, Gary, and Ernest Davis. 2020. "GPT-3, Bloviator: OpenAI's Language Generator Has No Idea What It's Talking About." MIT Technology Review. August 22, 2020. https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/.

Merleau-Ponty, Maurice. (1945) 1961. *Phenomenology of Perception*. London: Routledge & Kegan Paul.

Moelikker, Rick, and Julia Baan. 2022. "Een computerpreek is niet meer van echt te onderscheiden. 'Mijn collega's hadden het niet door'." *Nederlands Dagblad*, December 13, 2022. https://www.nd.nl/geloof/geloof/1154974/een-computerpreek-is-niet-meer-van-echt-te-onderscheiden-mijn-c.

Nagel, Thomas. 1986. *The View from Nowhere*. Oxford: Oxford University Press.

Pascal, Blaise. (1670) 1995. *Pensées*. Translated by A. J. Krailsheimer. London: Penguin.

Perler, Dominik. 2008. Thought Experiments: The Methodological Function of Angels in Late Medieval Epistemology. In *Angels in Medieval Philosophical Inquiry: Their Function and Significance*, edited by Isabel Iribarren and Martin Lenz. Hampshire, UK: Ashgate.

Russelman, Gerald. 1983. *Van James Watt tot Sigmund Freud*. Deventer, Netherlands: Van Loghum Slaterus.

Searle, John. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3:417–57.

———. 1981. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.

Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. 2017. "Mastering the Game of Go without Human Knowledge." *Nature* 550:354–359. DOI: https://doi.org/10.1038/nature24270.

Simango, Daniel. 2016. "The Imago Dei (Gen 1:26–27): A History of Interpretation from Philo to the Present." *Studia Historiae Ecclesiasticae* 42:172–90.

Simmen, Rene. 1968. *Mens & Machine: Teksten en documenten over automaten, androïden en robots* [*Human & Machine: Texts and Documents about Automation, Androids and Robots*]. Amsterdam: van Lindonk

Starke, Georg, Rick van den Brule, Bernice van den Elger, and Pim Haselager. 2021. "Intentional Machines: A Defence of Trust in Medical Artificial Intelligence." *Bioethics* 25 (April): 1–8.

Sternberg, Robert, ed. 2020. *Human Intelligence: An Introduction*. Cambridge: Cambridge University Press.

Stojnic, Gala, Kanishk Gandhi, Shannon Yasuda, Brenden Lake, and Moira Dillon. 2023. "Commonsense Psychology in Human Infants and Machines." *Cognition* 235, 105406. DOI: https://doi.org/10.1016/j.cognition.2023.105406.

Story, Daniel. 2022. "The Curious Case of LaMDA, the AI that Claimed to Be Sentient." *The Prindle Post*, June 22, 2022. https://www.prindleinstitute.org/2022/06/the-curious-case-of-lamda-the-ai-that-claimed-to-be-sentient/.

Turing, Alan. 1950. "Computing Machinery and Intelligence." *Mind* 49: 433–60.

van der Stigchel, Birgitte, Karel van den Bosch, Jeroen van Diggelen, and Pim Haselager. 2023. "Intelligent Decision Support in Medical Triage: Are People Robust to Biased Advice?" *Journal of Public Health*. *Journal of Public Health* 45 (3): 689–96. DOI: https://doi.org/10.1093/pubmed/fdad005.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." arXiv preprint arXiv:1706.03762.

Vroon, Pieter, and Douwe Draaisma. 1985. *De mens als metafoor: Over vergelijkingen van mens en machine in filosofie en psychologie*. Baarn, Netherlands: Ambo.

Yilmaz, Ramazan, and Fatma Yilmaz. 2023. "The Effect of Generative Artificial Intelligence (AI)-Based Tool Use on Students' Computational Thinking Skills, Programming Self-Efficacy and Motivation. *Computers and Education: Artificial Intelligence* 4, 100147. DOI: https://doi.org/10.1016/j.caeai.2023.100147.