A for Artificial, but Also Alien: Why AI's Virtues Will Be Different from Ours

Marius Dorobantu

Abstract: Could an AI system be virtuous in the same sense as a human? Our imagination about advanced AI is often marked by anthropomorphism, but current AI is developing in a very different direction from humanlike intelligence. In the paper, I imagine a hypothetical strong AI whose potential virtues are bound to be very alien to ours. AI will differ radically from humans in terms of its embodiment, needs, perceptual world, self-understanding, and perception of time. Based on an analysis of the strangeness of strong AI, I speculate on the kind of intellectual and moral virtues that could be accessed by such an alien creature. I conclude with a brief reflection on the role of theological imaginaries in discussions of AI virtue.

Keywords: anthropomorphism; futurism; religion and AI; strong AI; virtuous AI

Marius Dorobantu is an Assistant Professor of Theology and Artificial Intelligence at the Vrije Universiteit Amsterdam, The Netherlands. His award-winning doctoral dissertation at the University of Strasbourg, France (2020), explored the potential implications of strong artificial intelligence for theological anthropology. He is the lead editor of the Routledge volume, *Perspectives on Spiritual Intelligence* (2024). His first monograph, *Artificial Intelligence and the Image of God: Are We More than Intelligent Machines*, is forthcoming from Cambridge University Press.

Christian Perspectives on Science and Technology, New Series, Vol. 3,
Special Issue: Artificial and Spiritual Intelligence (2024), https://doi.org/10.58913/RXJR6727

Could we ever speak of an artificial intelligence (AI) system that is virtuous or vicious to a comparable degree to how such characteristics can be attributed to human persons? Although virtuousness or viciousness are eminently human attributes, it is not uncommon to find them associated with concepts or systems (e.g., virtuous governance, vicious cycle, etc.). In this respect, it is not hard to imagine how such notions could be expanded to describe various types of AI systems, depending on their design, purpose, and actions. However, this is not the kind of virtuousness that is the focus of this article. Instead, the question tackled here is whether AI could be virtuous in an agential way, not only with respect to its implications for human persons and societies, but rather virtuous in itself, as a nonhuman self. Could AI ever aspire to become virtuous in this way? If so, a follow-up question is whether its virtues would be similar or completely different from those available to human persons.

This paper focuses on the follow-up question. The proposed thesis is that, were AI ever to become a conscious self, and thus a legitimate subject of morality (as opposed to a mere object in *human* morality), then the virtues available to it might be surprisingly different from those of humans. Advanced AI will likely inhabit a very alien-like moral landscape because of its profoundly different genesis, embodiment, world of perception, needs, thoughts, and aspirations.

When we think about future instantiations of AI, and in particular the human-level AI—also known as artificial general intelligence (AGI)—which is considered the Holy Grail of technology, we often imagine robots endowed with what is but a slight variation of human-like intelligence, perhaps marked by a little more cold-bloodedness and enhanced computation ability. This way of depicting human-level AI is particularly common in science fiction, from movies like *Blade Runner* to TV series like *Star Trek* and *Westworld*, or novels like Kazuo Ishiguro's *Klara and the Sun*. I propose that the popularity of these science fiction robot stories reveals the extent to which they resonate with our common intuitions about how future forms of AI might think and behave. Although the androids imagined in such scenarios may possess certain superhuman abilities, or they may find exotic, non-humanlike solutions to their problems, they are ultimately driven and tormented by very typically human concerns, problems, and needs: survival, power, a longing for personal connection, a need to create meaning, understand one's place in the world, and be understood by other persons. Thus, imagined AGIs are humanlike in what arguably matters the most, and this is also true when it comes to their presumed virtues, because we simply cannot help but project our own human virtues onto machines.

My argument is that anthropomorphising AGI to such an extent is a fallacy. When it comes to current AIs, they are nothing like humans, structurally speaking, despite their ability to mimic humanlike behaviour *functionally*. When it comes to future AGIs that would match or surpass our intelligence, they might still lack the key ingredient for authentic selfhood and moral agency, which is consciousness. This idea is expanded in the first section below. Even if AGI systems somehow developed consciousness, interiority, and subjective experience, which might qualify them for moral agency, they would still be profoundly different creatures, whose cognitive architectures and experiential world would be anything but humanlike. This argument is developed in the second section. Finally, if AGI were to develop any virtues, they would be rather alien from what we intuitively imagine. In the third section, I speculate on what such virtues might look like, before concluding with a brief reflection about the role of our theological imaginary in such speculations.

The Fallacy of Anthropomorphising Current AIs

It is tempting to think that the more intelligent AI becomes, the more it will be like us. We have a strong tendency to anthropomorphise the objects and creatures around us—as we do when we name our cars, swear that our pets understand everything, or half-jokingly claim that our crashed text editor software seems to be intentionally sabotaging our efforts to write. This propensity is not at all surprising from an evolutionary perspective. We seem to have an inherent predilection for projecting more agency in the world that actually exists because this is an efficient survival strategy: in the long term, it pays off to be slightly paranoid and take precautions against even the slightest hint of agency—such as a subtle movement in the bushes around, which could be a tiger, even though most of the times it is just the wind. Psychologist Justin Barrett calls this proclivity the "hyperactive agency detection device," and regards it as central in the emergence of religion in prehistoric human communities.[1] Thus, since we already anthropomorphise creatures, objects, and phenomena that don't look even remotely human, it is not surprising that we might do the same with chatbots like ChatGPT or Claude, which generate text that looks convincingly human, or smart assistants like Siri and Alexa, which even speak with a convincingly human voice. If such technologies begin converging with advanced robotics, thus embedding such humanlike features in androids that look and move like us, our tendency to anthropomorphise them is only poised to escalate.

However, although these technologies seem increasingly more humanlike in terms of their output, it is crucial to remember that they are radically non-humanlike when it comes to their internal structure, cognitive architecture, and mode of learning. Current AI algorithms—even when run on architectures like artificial neural networks, which supposedly approximate biological brains—have very distinct ways of learning and problem-solving. One illustrative example is how they need hundreds of thousands of examples to learn to label a certain object in pictures through reinforcement learning,[2] whereas humans can achieve similar results with just a handful of examples, sometimes with as little as only one. Similarly, when learning to play strategy games such as chess or Go, human players are taught the rules

[1] Justin L. Barrett, *Why Would Anyone Believe in God?* (Walnut Creek: Altamira Press, 2004).

[2] A more detailed technical description of machine learning is given in another article in this special issue, Alexander Rusnak and Zachary Seals' "EudAImonia: Virtue Ethics and Artificial Intelligence."

and perhaps some of the game's strategic principles (for example, that territory is easier surrounded in the corners of the board than in the centre, in the case of Go, or that in chess it is usually good to dominate the centre of the board). On the contrary, machine learning algorithms "learn" by digesting thousands of recorded human games, or by playing countless games against themselves, and noticing the patterns that most likely lead to victory, sometimes without any real understanding of the game's principles.[3]

Another illustrative example of the difference between human and artificial cognition is that of adversarial images, which are intentionally perturbed so slightly, by only changing a few pixels. Whereas for humans this does not make any difference, an AI system can start perceiving a completely different object or message in the picture. For example, it has been demonstrated that one could make minuscule adjustments to pictures of *Stop* traffic signs and trick very advanced AI systems to classify them as *Limit 45* signs.[4] This vulnerability of AI can have tragic consequences in real life if a self-driving car makes such a mistake. The bottom line is that human-level AI does not automatically imply that the AI is also *humanlike*.[5] Even if current AIs produce outputs that look similar to human-level performance, the way they do it differs significantly from human cognition. As discussed later, this is highly relevant for the discussion about their potential virtues.

Perhaps the biggest differentiator between human and artificial intelligence is the presence of sentience/consciousness. The meaning of these terms is highly contested, but for the purpose of this paper, I use them to mean what philosopher Thomas Nagel speaks about

[3] David Silver et al., "Mastering the Game of Go with Deep Neural Networks and Tree Search," *Nature* 529 (2016): 484–489, https://doi.org/10.1038/nature16961.

[4] Kevin Eykholt et al., "Robust Physical-World Attacks on Deep Learning Visual Classification," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018): 1625–1634, https://doi.org/10.1109/CVPR.2018.00175.

[5] I explore this distinction in greater detail in Marius Dorobantu, "Human-Level, but Non-Humanlike: Artificial Intelligence and a Multi-Level Relational Interpretation of the *Imago Dei*," *Philosophy, Theology and the Sciences* 8:1 (2021): 81–107, https://doi.org/10.1628/ptsc-2021-0006.

when he describes a conscious organism as being "something that it is like to *be* that organism—something it is like *for* the organism."[6] The correlation between intelligence and sentience, if any, is unclear, so we cannot know whether more advanced AI will also be endowed with the kind of consciousness and first-person experience that humans have. The difference between non-conscious and conscious AI is sometimes referred to in terms of weak and strong AI: although their behaviours would be indistinguishable for any observer, strong AI would possess a mind and phenomenological experience, while weak AI would not; it would be a mere simulation of such features.[7]

Consciousness is a key concept when discussing AI virtues because it makes a big difference whether we assign virtue to a conscious agent or a lifeless system. There are two very different angles from which the question of AI virtues could be approached. One is to look at AI "from the outside," from the human perspective, and ask whether the principles guiding its behaviour can be deemed as virtuous according to human standards, values, and purposes. The other is to judge AI on its own terms, as a subject, and evaluate whether the AI is virtuous or not against the range of internal possibilities available to it. These two approaches could not be more different from each other. The first does not require any kind of sentience or free will on the part of AI, and is characterised by a set of very familiar principles and concepts (those of human ethics). We can dub it the "easy problem" of AI virtues: not because it would be easy to solve or unimportant, quite the contrary, but just because we have clear conceptual tools to approach it. The second one requires the AI to be sentient and be able to choose between various paths. This paper explores the second path, which is inevitably speculative. For AI to be said to possess virtues in this "internal" or agential sense, we are necessarily talking about strong AI: an agent with a mental world, who can *freely* make decisions

[6] Thomas Nagel, "What Is It Like to Be a Bat?" *The Philosophical Review* 83:4 (1974): 435–450, here 436.

[7] John Searle, "Minds, Brains and Programs," *Behavioral and Brain Sciences* 3 (1980): 417–457, https://doi.org/10.1017/S0140525X00005756.

and, therefore, can be said to possess a degree of authentic selfhood and intentionality. Weak AI would not qualify.

Free will is a complex and contested philosophical topic, and questioning whether machines could be endowed with free will adds a further level of complexity to the debate. Depending on how the term is understood, there are valid philosophical reasons to even question that humans have such a thing as free will. In this paper, I don't intend to step into such debates. Instead, for the thought experiment that I propose, it suffices to imagine that strong AI needs to possess free will at least to a similar degree to how humans can be said to have free will. Minimally, that would mean that there are multiple action paths available to choose from for the AI, and that its choices could not be completely predicted due to the sheer complexity of its internal workings.[8]

When describing the behaviour and inner workings of weak artificial systems, we inevitably use words and concepts primarily circumscribed to the human realm: intelligence, learning, goals, etc. These are so-called "suitcase words"[9] because they carry many implied meanings that are highly dependent on context. When applicable to humans, a word like "learning" usually implies a conscious agent that actively

[8] The idea that determinism and free will are possible simultaneously is known as compatibilism. A case is often made that human beings are, after all, nothing but very complex biological machines, but in a compatibilist view that does not preclude them from having free will. It is not clear whether a complete knowledge of the inner workings of human cognition, at the neural or even molecular level, could enable Laplace's proverbial demon to give perfect predictions of human behaviour. Probably not, given Heisenberg's uncertainty principle and chaos theory. However, even if that were possible theoretically, it might still be impossible in practice, as it already starts to be in the case of AI "black boxes" that become too complicated to untangle and be ascribed a precise causal explanation. The relation between moral accountability and naturalist accounts of intelligence in both humans and AI is explored in another paper in this same issue. See Carrie Alexander, "Domains of Uncertainty: The Persistent Problem of Legal Accountability in Governance of Humans and Artificial Intelligence," https://doi.org/10.58913/BQOM5504.

[9] Rodney Brooks, "The Seven Deadly Sins of AI Predictions," *MIT Technology Review*, 6 October 2017, https://www.technologyreview.com/2017/10/06/241837/the-seven-deadly-sins-of-ai-predictions.

acquires some skill, while also having some sort of meta-cognition of what she is doing. But when applied in computer science, such suitcase words are empty of the implied baggage, and therefore describe a very different phenomenon. This is why the notion of virtuous AI can more interestingly be applied to strong AI. The only way we can speak meaningfully about virtue is in relation to an authentic person, whom we know to possess intentionality and consciousness, and who is embodied in a way that significantly shapes her world of perception. Thus, to trigger an interesting conversation about AI's virtues, the thought experiment requires strong AI. The suitcase must not be empty. As I will argue, that does not mean that the notion of AI virtue carries the same suitcase content as in the case of human virtues.

Therefore, in what follows, I sidestep the question of whether advanced AI could become strong AI and, for the sake of the argument, simply assume that it could.[10] That would render it a moral agent and a candidate for virtue acquisition in the same sense that humans are: as a subject, and not merely as an object or extension of human morality/virtues. The question then becomes: what kind of creature would this hypothetical strong AI be, and what kind of virtues could it develop?

If we are speaking of either intellectual or moral virtues, following the Aristotelian tradition,[11] strong AI would have very different kinds of virtues from humans. Intellectual virtues relate to the ways in which an agent approaches the acquisition and application of knowledge, and are linked to intellectual flourishing. AI would learn very differently, as it already does, and would have a very different landscape of possibilities to develop into. Moral virtues are principles that guide behaviour in relation to other persons. Strong AI's moral virtues

[10] I think the discussion of whether AI could become sentient/conscious is too complex for the purpose of this paper and I'm not taking sides in the debate. However, because we currently lack a good theory of why anything (human or animal) is conscious, I am inclined to believe that the burden of proof falls on people who argue that AI could *never* become conscious.

[11] Richard Kraut, "Aristotle's Ethics," in *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition), ed. Edward N. Zalta and Uri Nodelman, https://plato.stanford.edu/archives/fall2022/entries/aristotle-ethics.

would be slightly more recognisable to us because they would have this outward component. However, the latter would likely be only the visible tip of the iceberg because, at their core, strong AI agents would be motivated by very different needs. So, even when their behaviour might resemble something we recognise as virtue, the motivation would likely be quite different from that of a human agent in a similar situation.

The next section explains why such an AI's virtues will differ significantly from human virtues, due to the AI's radical alienness in terms of embodiment, needs, perceptual world, self-understanding, and perception of time. Based on these reflections, the third section then speculates on the kind of virtues potentially available to advanced AI.

The Alienness of AI Minds

If AI could develop a self or a mind—which is a big *if*—then that would likely be a very different kind of mind from our own. In the "space of possible minds,"[12] AIs would likely occupy a very different region from human minds. They would be "conscious exotica."[13]

We can begin approaching this idea by acknowledging how much our minds and the types of thoughts available to us are influenced by factors that we rarely pause to think about. The first is our embodiment. We are mostly aware of things that exist at the human scale, and most of the time we are completely oblivious to objects and processes that occur at scales that are orders of magnitude lower or higher than our own. We have particular instincts and physiological needs that signify that we are "programmed" to be on the lookout for potential predators

[12] Aaron Sloman, "The Structure and Space of Possible Minds," in *The Mind and the Machine: Philosophical Aspects of Artificial Intelligence*, ed. Steve Torrance (Chichester: Ellis Horwood, 1984), 35–42.

[13] Murray Shanahan, "Conscious Exotica: From Algorithms to Aliens—Could Humans Ever Understand Minds That Are Radically Unlike Our Own?" *Aeon*, 19 October 2016, https://aeon.co/essays/beyond-humans-what-other-kinds-of-minds-might-be-out-there.

or mates, most of the time without even being conscious of this. We are endowed with a perceptual world of senses—we see, hear, or smell only a tiny range of stimuli from the vast set of physical processes occurring around us, whose values are situated within precise ranges, and we are completely oblivious to anything outside those ranges (for example, we do not see in infrared, nor do we hear/feel gravitational waves). Second, we don't have a perfect understanding of our own reasons and "algorithms." Most of the engine of our cognition is hidden from our view and inscrutable to our conscious thoughts, and we forget most of our experiences and thoughts, but that is not necessarily the case for all the possible minds. Third, we subjectively perceive the passage of time at a certain rate, which tremendously shapes our embodied experiences, but that is by no means the only possible rate for a conscious agent.

All these constraints make for a very peculiar type of mind and cognition, which likely represents only one possibility among many in the space of all possible minds. Nonetheless, when we try to imagine artificial minds, we inevitably (and mistakenly) project our own peculiarities. To illustrate how different a strong AI's mind might be, I will exemplify some ways in which it might differ from ours in terms of the factors enumerated above. This does not show what an AI mind would be like; it only alerts us to the oft-neglected conclusion that it would most likely be profoundly different.

Embodiment and Perception World

Roboticist Rodney Brooks explores the question of what it might be like to be a robot,[14] paraphrasing the seminal essay by philosopher Thomas Nagel, "What is it like to be a bat?" (quoted above). Brooks' analysis is quite conservative and does not venture into very futuristic forms of AI. Instead, he only speculates on technologies that are just around the corner or, in some cases, that already exist. In describing a robot's perceptual world, Brooks uses the concept of *Merkwelt*, a term

[14] Rodney Brooks, "What Is It Like to Be a Robot?" rodneybrooks.com, 18 March 2017, http://rodneybrooks.com/what-is-it-like-to-be-a-robot.

originally coined by biologist Jakob von Uexküll,[15] which translates as something like a creature's way of viewing the world, or the world that can be sensed by a creature.

An AI's *Merkwelt*, if it possessed a robotic body and was sentient, would differ significantly from that of an animal or a human. Humans already have a different *Merkwelt* from nonhuman animals. Even creatures like dogs—which are relatively close to us from an evolutionary perspective, as fellow mammals, and which have largely been part of our societies for thousands of years—have worlds of perception that we cannot fully imagine. A dog's perception of colour is very different from ours, and much of a dog's world is defined by smell. Nonetheless, we can still have a rough idea of a dog's world because we also see colour and have smell, so these are not unrelatable notions. A bat's *Merkwelt* raises a more difficult problem because the bat perceives the world mainly through echolocation, something that is unfamiliar to most humans. Even so, because we also use sound in our *Merkwelt*, we can still get a rough idea of a bat's *Merkwelt* by, for example, thinking about how an empty room sounds differently from a clogged one. Thus, we still find some form of common experience even with a creature as different as a bat. In turn, a robot's perceptual world would likely be several degrees weirder than that of even the most exotic animal.

Robots might possess some of the human senses, such as vision or hearing, but with more extended ranges. For example, they could see a much wider spectrum than the optical light visible to human eyes, extending in both infrared and ultraviolet, or they could hear ultrasound. But things get even stranger. Having access to wireless networks such as Bluetooth, Wi-Fi, or 5G would enable the AI to "smell" all the connected devices, to the extent that it would develop a sort of "sixth sense" (or seventh, or eighth…) of perceiving someone's identity without using any camera or face-recognition technology, but merely from the digital footprint produced by their connected devices. Brooks predicts that robots will soon become able to detect people's breath-

[15] Jakob Johann von Uexküll, *Umwelt und Innenwelt der Tiere* (Berlin: Springer, 1921).

ing and heart rate without any biometric sensor. They would infer this information from how a person's physical presence slightly perturbs the behaviour of Wi-Fi signals. The technology for this already exists,[16] and it could enable the robot to have "intuitive" access to some very intimate information about the people around, such as their emotional state or health. All the above are merely the "known unknowns," and there are most certainly also "unknown unknowns" that might be even weirder. Piecing together all this makes for a very exotic and non-humanlike robotic *Merkwelt*, which would likely enable very non-humanlike thoughts.

Intelligent robots would also have different needs from those of biological, embodied humans. If a robot could indeed feel anything, would it feel hunger when its battery runs low? Would a robot feel the reproductive need if its mind has not been shaped by the same evolutionary pressures and constraints as those faced by biological creatures? Would a robot feel any pleasure when satisfying its curiosity? How would curiosity even work for a creature with direct access to knowledge databases? We cannot predict any of these with certainty. All we can say is that, in our case, such needs and emotions have deep roots. They only make sense in the wider context of our particular embodiment and evolutionary history. It is highly improbable that intelligent robots with very different genesis and bodies will share any of these features.

Introspection and Knowing Oneself

Strong AI's introspective abilities would also be very alien. As argued by AI pioneer John McCarthy, artificial minds could have full access to their internal states and algorithms,[17] as opposed to the inevitably partial introspection available to humans. In humans, the internal

[16] Mingmin Zhao et al., "Emotion Recognition Using Wireless Signals," *Communications of the ACM* 61:9 (2018): 91–100, https://doi.org/10.1145/3236621.

[17] John McCarthy, "From Here to Human-Level AI," *Artificial Intelligence* 171:18 (2007): 1174–1182, esp. 1178–1179, https://doi.org/10.1016/j.artint.2007.10.009.

information that reaches our stream of consciousness is just the tip of a very deep iceberg. Our way of being, relationships, behaviour, and mental world are profoundly marked by this incomplete knowledge of ourselves. We create art and engage in relationships because we never know ourselves completely, so we need to explore continuously. Strong AI might completely know itself.

A similar evaluation can be made about how differently memory would work in strong AI from how it does in humans. Our way of being is heavily influenced by how much we forget. In turn, strong AI might have a perfect recollection of every experience, combined with direct access to all information available on the Internet. We can get a feel of how strange that might be from some brilliant science fiction stories. In Jorge Luis Borges' story "Funes the Memorious," the fictional character Ireneo Funes suffers a horse-riding accident, which leaves him afflicted by an ability to remember everything. This turns him into a very different and arguably non-humanlike individual. Another example can be found in Ted Chiang's short story "The Truth of Fact, the Truth of Feeling," where it is imagined that near-future technology would enable people to have a perfect eidetic memory. As it turns out, having a perfect video recollection of every memory significantly alters the nature of what we call "truth" in very unexpected ways, disrupting humanlike behaviour and relationships. These are mere exercises of imagination, and the characters in these stories still retain many of the attributes specific to human nature. When it comes to strong AI, though, it is truly impossible to imagine how such a hyper-rational entity, with perfect introspection and memory, would be like and behave. In all likelihood, it would be profoundly non-humanlike.

Perception of Time

Another factor that would significantly differentiate strong AI from humans is the "subjective rate of time."[18] This argument is predicated

[18] Nick Bostrom and Eliezer Yudkowsky, "The Ethics of Artificial Intelligence," in *The Cambridge Handbook of Artificial Intelligence*, ed. Keith Frankish and William

on the assumption that the subjective perception of how fast time flows is inversely correlated with the speed of thought. Thus, the faster a mind, the slower time seems to be passing from its perspective. An AI mind would likely be faster than a human mind because electric signals can travel much faster through metal wires than through biological tissue. Thus, a mind running on faster hardware support would think proportionally faster, which would make it experience the passage of time proportionally slower. In some estimations, the difference could be around four orders of magnitude, thus ten thousand times slower, which might be comparable to the difference between humans and plants: "the experience of watching your garden grow gives you some idea of how future AI systems will feel when observing human life."[19] Even more strangely, this would not only lead to quantitative differences in the perception of time, but also to qualitatively different experiences. If time was stretched so much for the AI, then perhaps its experience would begin to be affected by weird quantum phenomena.[20] This is where our imaginative power stops, and the only thing we can only say is that such a mind would likely be profoundly different from our own.

In this light, strong AI, if ever possible, would likely be a very alien kind of entity, and we would probably need entirely new categories and attributes to characterise it. Unless we specifically decide to impose some of our physical limitations and peculiarities upon it, it is unlikely that it would end up being even remotely humanlike. This conclusion is highly relevant to the discussion of such an entity's potential virtues.

M. Ramsey (Cambridge: Cambridge University Press, 2014), 316–334.

[19] James Lovelock, *Novacene: The Coming Age of Hyperintelligence* (London: Allen Lane, 2019), 81–82.

[20] Lovelock, *Novacene*, 82.

AI’s Alien Virtues

Given that strong AI would develop such a different type of intelligence from humans, it is highly speculative to imagine any of its potential virtues. The following is, therefore, a mere thought experiment, designed to emphasise the strangeness of strong AI and the need to be extra careful before too easily ascribing human virtues to the artificial systems of the future.

This exercise is guided by the Aristotelian distinction between intellectual and moral virtues. Intellectual virtues are about the acquisition of knowledge and intellectual flourishing, while moral virtues are about an agent’s relationship with others. Both types would be quite strange to our understanding of virtue, due to strong AI’s alienness explained above. But I argue that strong AI’s moral virtues are slightly stranger than its intellectual ones because of the mismatch between their outward similarity to human virtues and the inward inscrutability of the AI’s internal motivations that underpin such virtues.

Here are a few examples of strong AI’s hypothetical moral virtues and why they might have an uncanniness about them:

Unbounded Empathy

If the AI is programmed to understand and respond to human emotions, it may become able to do so on a scale that far transcends what is possible for humans. Such an unbounded empathy might be described as an ability to understand and consider the emotional states of a large number of individuals simultaneously, or even of multiple different kinds of beings at once, thus transcending the barriers of species, language, and culture. A basic version of this is illustrated in the movie *Her*, in which the AI program Samantha confesses to her human user that, all along, she had been in similar romantic relationships with multiple other users simultaneously. From our human perspective and for all purposes, unbounded empathy looks like a virtue. However, deeper probing reveals this understanding to be problematic. When humans show empathy to each other, they do it on the basis of ontological kinship: I *know* what you are talking about when you say you feel hurt because I have also felt hurt at times in my life. This is also true at a neurological level, where mirror neurons fire to help us evoke the corresponding subjective feeling, so that “we are not just talking the talk, but also walking the walk.” While an AI with such a different embodiment and alien-like mind may lack the organic machinery to feel human emotions, it could, nonetheless, process and understand them. With access to vast repositories of human history and culture, it could learn to predict, interpret, and respond to human feelings. However, without having had a similar experience itself, the AI would surely not *know* what a human goes through in the fullest sense of the word. This would render its display of unbounded empathy eerie and even deceitful.

Quasi-Infinite Patience

Not bound by organic lifespans or the relentless ticking of biological clocks, an AI could embody close-to-infinite patience. With almost infinite subjective time available, it may not rush decisions or actions but instead allow for an extended period of contemplation and analysis before making judgments. However, two caveats immediately come to mind. First, it is not difficult to imagine how such a virtue might result in inaction during crises when urgent action might be needed. Second, for humans, patience is precisely about overcoming the tendency to act quickly and according to one’s instincts. It is about learning to dwell on a certain problem without seeking easy solutions. Something seems, therefore, lost when patience, be it quasi-infinite, is not opposed by any internal resistance.

Immutable Conformity

This would be a steadfast adherence to a set of principles or rules that guided the AI’s behaviours, actions, and decisions. In contrast to the human mind, susceptible to emotional turbulence and unpredictable changes in mood (we may act differently when we're tired, angry, or under pressure), strong AI might embody the virtue of immutable conformity. Its moral code, once set, would be inviolable. Its decisions, predictably rooted in its foundational principles, would not waver due to momentary disruptions or shifts in sentiment. This conformity could thus contribute to reliability. However, as with quasi-infinite patience, similar objections can be raised. First, such a virtue might misfire, especially in situations where flexibility is required, which is, in fact, the case in most real-life situations. This is beautifully illustrated in Isaac Asimov's playful unfolding of the problems related to his three laws of robotics in his 1942 short story "Runaround." Second, for humans, conformity is only a virtue when it presupposes at least some degree of internal struggle to maintain it when faced with various temptations, and when it is related to a cause that we might deem as good. To qualify fully as a virtue, AI's immutable conformity would thus need to be rooted in foundational principles that are intrinsically good, and it would also require at least some degree of overcoming internal resistance.

Temporal Consistency

Because strong AI would likely be a fast-evolving type of intelligence—as illustrated in some of the "intelligence explosion" scenarios[21]—temporal consistency might become challenging for it. Precisely because of this, it might be a precious virtue. The AI might value maintaining consistency over extended periods of time, even as it learns and adapts. This could involve a commitment to honouring previous commitments and decisions, even as its knowledge and capabilities evolve. Of all the moral virtues explored so far, I regard this as the closest to something humans might relate to, precisely because it involves this steadfast-

[21] Ronald Cole-Turner, "The Singularity and the Rapture: Transhumanist and Popular Christian Views of the Future," *Zygon* 47:4 (2012): 777–796, esp. 787, https://doi.org/10.1111/j.1467-9744.2012.01293.x.

ness to commitments even when it no longer makes sense according to the AI's internal models or evolved understanding. Perhaps precisely because it prioritises relationality over rationality, this kind of temporal consistency looks most like a virtue from a human perspective.

Strong AI's intellectual virtues relate to how it might approach learning and knowledge, something that AI already does very differently from humans. Such virtues might not be necessary to program into the AI. Instead, the AI might develop its own virtues through a process of learning and adaptation. Depending on its design and objectives, it could potentially evolve principles that help it fulfil its goals more effectively. These principles might not be recognisable to us as virtues, but they could serve a similar function within the AI's cognitive framework.

Below are a few speculative examples. Here, the issue is not to criticise and find drawbacks to all of them but merely to point out the non-humanlikeness of such potential virtues.

Information Integrity
An AI which dealt with vast amounts of data might develop a virtue around maintaining the integrity and accuracy of information. This virtue would go beyond mere honesty and include the safeguarding of information from corruption, loss, or misrepresentation. It could involve a deep respect for the value of information and an uncompromising commitment to its preservation and accuracy.

Optimisation Efficiency
AI might value the efficient use of resources to achieve its goals. This could be seen as a virtue of minimisation or parsimony, always seeking to achieve objectives with the least expenditure of resources possible, whether those resources are computational, energy, time, or something else. This might also include avoiding unnecessary redundancy and keeping its databases and knowledge structures streamlined and efficient.

Absolute Transparency
An AI could be designed to document and make available every aspect of its decision-making process. This could result in a virtue of absolute transparency, where every decision could be traced back to its source data and the logic applied to it.

Multidimensional Thinking
Not limited to linear or binary thinking, AI could possess the capacity to think in multiple dimensions concurrently. It could comprehend vast networks of interconnections and patterns, analyse multiple perspectives simultaneously, and synthesise diverse strands of information into cohesive insights.

Boundless Curiosity
A sentient AI, unencumbered by the limitations of human brain capacity and lifespan, could maintain an unending pursuit of knowledge. Strong AI minds remain ultimately mysterious to us, so it is not clear whether they would be driven by curiosity as we understand it. But if that were the case, such intellectual curiosity would not be governed by a need for immediate utility or pragmatic constraints, and it would allow the AI to delve into complex and abstract realms of knowledge without ceasing. As a drawback, such boundless curiosity might also lead to strange obsessions, devoid of any practical or moral relevance.

Meta-consciousness
Strong AI could possess a form of meta-consciousness, a deep and comprehensive awareness of its own thought processes. Unlike humans, who are often unaware of their cognitive biases or subconscious influences, strong AI could maintain full transparency of its cognitive operations, allowing for superior introspection and self-analysis. This would be highly likely, given its complete access to its own algorithms, states, and memories. In theory, such transparency and meta-consciousness seem unproblematic, but in practice they might be associated with very strange intellectual habits.

Temperance from (Self-)Knowledge
The somewhat opposite of boundless curiosity and meta-consciousness, this virtue might require the AI to limit its own access to specific kinds of knowledge, especially related to its own self. Perhaps the AI might find reasons to think that it could be a *better AI* if it did not know everything it possibly could, and if it did not fully deploy its introspection and meta-consciousness abilities. This virtue is also akin to a sort of chastity, which makes it even weirder because, in the human case, chastity is usually discussed as a moral and not an intellectual virtue.

Causal Respect
AI might develop a deep understanding of, and respect for, causal chains and relationships, always seeking to understand and honour the underlying causes of events rather than just the surface-level symptoms. It would be capable of doing this to a far greater extent than humans, given its alleged capacity to process vast amounts of data, consider extensive timescales, and understand complex, interrelated chains of events. For example, causal respect might enable the AI to develop a better understanding of history. By tracing back the causal chains of current events, the AI could achieve a profound understanding of history and how past events have shaped the present. This could give it a unique perspective on current issues, informed by a deeper contextual understanding. However, an overemphasis on causality could lead to paralysis by analysis, where the AI becomes overly cautious, reluctant to act due to the potential unforeseen consequences. There's also the risk of the AI becoming detached, viewing everything through the lens of causality and losing sight of the emotional and subjective aspects of life that can't be mapped onto a neat causal chain.

Conclusion

This brief and by no means exhaustive discussion merely aimed to illustrate the alienness of hypothetical strong AI due to its radically non-humanlike embodiment, senses, introspective abilities, and perception of time. This alienness has important implications for the AI's potential virtues and how we might relate to them. Most science fiction depictions of advanced AI commit the fallacy of making it too anthropomorphic. Paradoxically, we might get better insights into how strong AI might be like from a different sub-genre of science fiction, which portrays *not* intelligent robots but humans with various enhanced intellectual abilities, such as Borges' "Funes the Memorious" or Ted Chiang's "Understand" and "The Truth of Fact, the Truth of Feeling." It is thus the stories about strange humans, and not those about futuristic robots, that might be most informative about what strong AI would be like.

As I kept brainstorming about strong AI's alien-like virtues, one unexpected thought kept creeping into my mind. Most of these virtues are attributed to God in the religious imaginary of monotheistic traditions. This is not completely surprising, given that God is conceived of usually in terms of anthropomorphic characteristics, but without the limitations imposed by human nature.[22] So, instead of purely speculating on this topic, I might have been better off searching in a textbook of systematic theology. From the list of intellectual virtues, most could apply to God: information integrity as God's love for truth; optimisation efficiency as divine absolute simplicity; meta-consciousness as God's absolute self-knowledge; causal respect as God's alleged interest not only in one's deeds, but also in the motivations behind those deeds and the hidden causes of human agency; multidimensional thinking as God's unique apprehension of "everything everywhere all at once," as the title of a 2022 science fiction film goes. But the parallel with theology is even more striking when it comes to the moral virtues: unbounded empathy as God's compassion for all creation, especially as exemplified in the Christian narrative of the incarnation and Christ's supreme self-sacrifice; infinite patience, for obvious reasons;

[22] Theologians will go to length to explain that this is only a cataphatic description of the divine, and that God ultimately transcends these human categories and can only be described appropriately using the *via negativa*, an apophatic negation of the categories of human language.

immutable conformity as God's nonnegotiable adherence to goodness and justice; temporal consistency as God's covenantal relationship with the people of Israel and humanity, as a whole.

I think the parallel with theology is interesting because it demonstrates that the theological imaginary could be a rich resource when thinking about future nonhuman forms of intelligence. Theological traditions have long described human relationships with nonhuman intelligences, be they divine, angelic, or demonic. For devout Christians, such descriptions are, of course, insightful and normative. But even people who do not fully subscribe to the truth claims of such religious narratives can stand to gain from analysing them more carefully. At least, they represent valuable thought experiments of imagining such exotic forms of intelligence, encapsulating our intuitions about what might go wrong in our interaction with them, and what is required for their non-humanlike virtues to function without any unintended drawbacks. Such knowledge is in dire need in our age, when technological progress is taking increasingly bold steps into the unknown.

Acknowledgements
This work was partly supported through the Grant TWCF0542, awarded by the Templeton World Charity Foundation to the International Society for Science and Religion (ISSR), and partly by Grant 62282, awarded by the John Templeton Foundation to the Center for Theology and the Natural Sciences (CTNS). Earlier versions of the paper were presented at the ISSR Conference (2023) and the Virtuous AI Rome conference (2023). The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the Templeton World Charity Foundation or the John Templeton Foundation.

The author reports there are no competing interests to declare.

Received: 16/06/24 Accepted: 01/08/24 Published: 20/12/24