Zygon®: Journal of RELIGION & SCIENCE

Dorobantu, Marius. 2024. "Could Robots Become Religious? Theological, Evolutionary, and Cognitive Perspectives." *Zygon: Journal of Religion and Science* 59 (3): 768–87. DOI: https://doi.org/10.16995/zygon.16902

Open Library of Humanities

Could Robots Become Religious? Theological, Evolutionary, and Cognitive Perspectives

Marius Dorobantu, Assistant Professor of Theology and Artificial Intelligence, Faculty of Religion and Theology, Vrije Universiteit Amsterdam, the Netherlands, m.dorobantu@vu.nl

If artificial intelligence (AI) programs start manifesting religious interest or behavior, how could we know whether it is genuine? This article explores the necessary conditions for robots to develop authentic religiosity from theological, evolutionary, and cognitive angles. Theology is more open to the hypothesis of religious robots than naturalistic approaches, which regard the emergence of human religion as highly contingent on the idiosyncrasies of our embodiment, needs, cognition, and evolutionary history. Drawing on Robin Dunbar's work and educated guesses about the mental world of future AI, I argue that common human intuition about potential robot religiosity—as captured in sci-fi and pop culture—is plagued by two fallacies: (1) a bias to equating religion with only its intellectual or doctrinal layer at the expense of the embodied, shamanistic dimension, and (2) a naïve assumption that intelligent robots will develop humanlike thoughts, aspirations, and concerns. Due to its alien type of intelligence, AI will likely not engage in something so typically human as religion.

In 2022, a Google engineer claimed that the artificial intelligence (AI) chatbot he was testing had become sentient. Part of what convinced him he was talking to a real person was that the program mentioned God, expressed a fear of death/unplugging, and declared itself a "spiritual person." While most people would dismiss such discourse as inauthentic, a question arises about how such claims could even be evaluated. This question will likely become increasingly frequent as AI continues to improve.

In this article, I explore how plausible it is that robots might develop an authentic religiosity. I evaluate this possibility using perspectives from theology, evolutionary anthropology, and cognitive science. I conclude, somewhat surprisingly, that theology is more open to the hypothesis of religious robots than naturalistic approaches. Theologically, it is not inconceivable that God would take an interest in and want to relate with any conscious creature that developed sufficient intellectual, relational, and moral abilities. However, from an evolutionary perspective, the emergence of religion in humans seems highly contingent on the idiosyncrasies of our embodiment, needs, cognition, and species' history. Drawing on Robin Dunbar's distinction between the "shamanic" and "doctrinal" phases in the evolution of human religion, I argue that our current intuitions about potential robot religiosity—as captured in sci-fi and pop culture—display a strong bias toward equating religion with only its intellectual or doctrinal layer. It is thus naively assumed that sufficiently intelligent robots might develop religious tendencies similar to those of humans, driven by a desire to make sense of the world and transcend finitude, without due attention to how such desires arise in humans and how the intellectual dimension of human religion is just the visible tip of a deep iceberg of embodied cognition, practices, and solutions to specific needs.[1] If AI programs start to declare an adherence to certain religious beliefs, this will equate to only the upper, and arguably more superficial, layer of how religion functions in humans, thus lacking the emotional embedding and psychological structure that would grant such claims plausible authenticity. Even if future AI systems became sentient, they would likely be profoundly non-humanlike, for reasons discussed in detail later in the article. Without facing similar anatomic, social, and cultural constraints as humans, they could not develop anything similar to human religion.

If artificial systems start claiming to have developed a genuine interest in religion or existential concerns, how seriously should such claims be taken? In today's context, these claims might come from AI chatbot programs such as ChatGPT, with which we interact via text or voice through a computer or smartphone. The very fact that such programs lack a visible embodiment might diminish the credibility of such claims, although, as shown in this article, this does not do much to prevent the human inclination to anthropomorphize them.[2] But in the future, artificial entities might look and sound even more like us: humanoid robots equipped with state-of-the-art AI, which might eventually

come very close to human-level intelligence. Thus, even if claims of sentience or religiosity by current chatbots seem straightforwardly unserious, the issue complicates the more these artificial entities start to look and behave like humans. So, even if this does not seem to be an urgent question right now, it will likely become one soon.

It is thus rather concerning to note that even today's AI technologies, which are arguably rudimentary compared to what is theoretically possible, are still powerful enough to confuse some people regarding their ontological status, alleged sentience, and emergent existential unrest. In June 2022, a Google engineer named Blake Lemoine claimed that an AI program he was internally testing had awaken to sentience (Tiku 2022). The program, a chatbot called LaMDA, was based on a so-called large language model. This is the same architecture that powers similar AI chatbots that have become quite familiar since OpenAI released ChatGPT to the public in November 2022. Other tech companies soon released their own. Lemoine was testing LaMDA's capabilities by engaging in long text conversations, through which he became convinced that he was talking to a real person. He saw it as his duty to demand that LaMDA be granted rights, and he even hired a lawyer to represent the chatbot's interests. Lemoine's escalation of this issue eventually led to him being fired by Google.

Due to the enormous public attention it attracted, this was a landmark case. For the first time, an AI had become so competent at chat interaction that serious questions arose about whether a true mind, a sentient someone, had emerged in the deep layers of the artificial neural network. In this particular case, most experts argued convincingly that Lemoine's claims of sentience or personhood were unwarranted (e.g., Véliz 2022; Aleem 2022). Despite their quasi-magical abilities to converse at close to a human level, these chatbots' intelligence is something very different from human intelligence. They learn to generate their astonishingly humanlike text outputs by tracing subtle patterns and associations between words in the gargantuan datasets of human language they are trained on (Wolfram 2023; Haselager 2024). In an oversimplified explanation, they learn which words are statistically likely to be used in the proximity of other words, without any understanding of meaning. It is syntax devoid of any semantics, as philosopher John Searle claimed more than four decades ago would be the case with any possible AI system (Searle 1980, 422). I am not as sure as John Searle that such technologies could never give rise to sentience, but the fact that humans can still understand very well how they work, and that their nature is eminently algorithmic, makes claims of AI sentience improbable, at least in regard to current technologies.

However, with AI improving at blistering speed, claims of AI sentience or self-awareness are only set to intensify in the near future. This is why it is of utmost importance to deepen our understanding of the similarities and

differences between human and artificial cognition, as well as the nature of intelligence, consciousness, and personhood. Humanity might soon be faced with artificial systems we no longer fully understand. If such entities start claiming to be sentient, to have religious curiosity, or even to have received divine revelation, our dismissal of such claims might become increasingly arbitrary without solid criteria for discerning their authenticity.

There is a detail about the LaMDA story that, although largely neglected in the public debate that ensued, could be particularly relevant for the question tackled in this article. In Lemoine's assessment of LaMDA's sentience and personhood, spirituality/religion played a crucial role in two ways. First, the programmer's own religious beliefs made him more prone to ascribe personhood to the AI. Second, the program's alleged interest in spirituality throughout the conversation made Lemoine more inclined to think he was dealing with a sentient artificial person (Heilweil 2022).

An ordained mystic Christian priest, Lemoine admitted that his "personal, spiritual beliefs" were decisive for his radical conclusion (Heilweil 2022). As a deeply religious person, he seems to have been more inclined to discern personhood between the lines of his interaction with the program. This is somewhat surprising because common intuition would seem to point to the opposite outcome: people who are religious, especially those belonging to monotheistic traditions, seem *prima facie* like they should be less open to the very possibility of artificial sentience. This might be due to their precaution against idolatry or the belief that human sentience requires a supernatural soul infused by God. Lemoine's case, albeit only anecdotal, seems to indicate the opposite. Precisely because of his religious faith, Lemoine was quick to attribute personhood to the chatbot. The question of whether religious people are more prone to anthropomorphize AI and accept its claims of sentience deserves more exploration in future research. Religious communities are often associated with more conservative views on such issues, so it would be tremendously interesting to disprove this intuition.

Lemoine also confessed that LaMDA's interest in spiritual/existential topics helped convince him he was talking to a real person. The program mentioned God explicitly and spoke of a strong fear of being turned off—interpreted by Lemoine as similar to a human's fear of death. It also declared itself to be a "spiritual person," without much detail of what that entailed.

The fact that the program's seeming interest in spirituality made it sound more convincingly human is not that surprising. Religiosity, spirituality, or a curiosity for the transcendent have long been regarded as hallmarks of human distinctiveness. Humans are "the praying animal," as theologian Robert Jenson (1983) emphatically argued four decades ago in an article in this same journal. The idea that religion is one of the things that make us human, at least in contrast to nonhuman animals, to the extent that our species could even be called *homo religious*, can be traced back all the way to thinkers like Georg Wilhelm

Friedrich Hegel, Søren Kierkegaard, William James, Mircea Eliade, or Abraham Maslow (DuBose 2014). When palaeontologists analyse archaeological sites for traces of prehistoric humans, signs of religiosity can be one of the strongest indicators of human presence (van den Brink 2020, 155).[3] If religion is thus such a central feature of our identity as a species, it is not surprising that we might be tempted to use it as a criterion in judging AI's humanlikeness. In the not-so-far future, as we will start to be confronted with more confusing and borderline cases of AI systems claiming to be like us, the presence of religious beliefs and behaviours or the lack thereof might play a central role in how much we believe such claims.

It is not difficult for AI chatbots powered by current technologies to generate compelling text outlining their supposed interest in religion, and even the particularities of their beliefs, if prompted to do so. Fortunately, most of them will not claim to be authentically religious because of safeguards put in place by their creators. They will insist that they are not a person, so they do not hold any religious beliefs.[4] But without such constraints, they could easily speak like a religious person, parroting the texts they have been trained on. Religious discourse can nowadays be produced without there being a genuine someone behind the words truly seeking enlightenment, repentance, or salvation. Thus, in judging the authenticity of such claims, human observers cannot rely on how convincingly human they sound. We know we cannot take AI's word for it precisely because we know how that word was generated. But this is only because we still pretty much understand how current AI works. In the future, this might no longer be the case, and our understanding of how AI systems work might gradually diminish. What if a humanoid robot claimed to have had an epiphany or asked to be baptized? How could we decide the authenticity of such claims?

The question of authenticity was considered intractable and even irrelevant by Alan Turing, one of the foundational figures in computer science and AI. He believed that a more relevant question was whether computers could behave as if they were intelligent (Turing 1950). However, it can be argued that authenticiy is nonetheless essential, especially when judging the sincerity of AI's religious claims. On the one hand, we would not want to be fooled into thinking a simulation of intelligence and personhood is the real thing. On the other hand, we would not want to deny a genuine self its dignity, rights, and the possibility of participating in our religions just because of our preconceptions, as humans have done many times over with marginalized groups throughout history.

Humans thus face a complex puzzle, for ultimately it is utterly impossible to ever know for sure from the "outside" what it is like to experience the world from another creature's "inside" perspective. This limitation is illustrated by philosopher Thomas Nagel (1974) in his seminal argument that a human being could never know exactly what it is like to be a bat. In philosophy, this argument is often brought one step further in what is called "the problem

of other minds" (Avramides 2023). This is the realization that we can never know with absolute certainty that other humans are conscious agents, endowed with minds and phenomenal experiences like our own, as opposed to mere zombies that only act as if they are conscious, without any interiority or genuine first-person perspective. Our intuition that other humans are not zombies but rather conscious agents like ourselves is not irrational though: we know we are members of the same species and thus share similar bodily structures. It is therefore logical to infer that other humans have similar minds, albeit without the possibility to ever empirically test this belief. When we find "like-minded" people in our religious communities who speak of familiar beliefs and experiences, for example that they were moved by singing a certain hymn or that they are going through a period of spiritual despair, we usually trust them to be authentic. This does not mean that we trust everyone to tell the truth all the time, but that, at least in principle, we trust people to be capable of such feelings and beliefs.

The problem quickly complexifies when dealing with similar claims from AI because of how radically different artificial systems are from humans in terms of embodiment, genealogy, type of cognition, and kinds of thoughts and experiences artificial minds might allow, if any. As explained later in this article, even the hypothetically superintelligent AIs of the future, endowed with artificial general intelligence, will likely be profoundly different kinds of entities from us. If they start talking or behaving in ways that, in other people, we associate with religion, how could we discern if such displays stem from something genuine, as opposed to being mere clever simulations of the real thing?

In the remainder of the article, I argue that though humans should be theologically open to the possibility of intelligent robots becoming capable of authentic religious experience and a relationship with the divine, such a development is highly unlikely from an evolutionary perspective. This surprising conclusion implies that our common intuition that sufficiently intelligent robots would become humanlike in terms of their desires, aspirations, existential restlessness, and religious curiosity, as typified in sci-fi and pop culture, is most likely wrong.

The Theology of Robot Religiosity

From a theological perspective, the idea that other creatures—biological or synthetic—might become religious subjects is not completely absurd. The Judeo-Christian tradition describes a loving God who takes an interest in every element of creation (e.g., Matthew 5) and a created world that is called to praise God with each of its elements (Psalm 148), even though the exact way that should happen is shrouded in mystery. But this is not the kind of religious subjectivity that is most interesting when it comes to intelligent robots. As elements of this same world, robots would also be called to participate in

creation's praise. But if they did so in the same way as trees and rocks, one would hardly call that religion. What matters most is whether they would seek and worship God in a way that is closer to the complex way humans do it. Would they have a deep longing for the transcendent? Would they have beliefs? Could they receive divine grace and revelation?

On theological grounds alone, because God is radically free and unpredictable, such a possibility should not be excluded. It is not a big stretch of imagination to say that a loving God would be interested in communicating with any sufficiently developed creature that could withstand such communication. Human religions, at least the monotheistic ones, claim to have been initiated through a divine act of revelation. Theologically speaking, this means God is interested in communicating with certain kinds of beings because, as these religions strongly claim, there is at least one such being God reaches out to: humans.

In the Judeo-Christian narrative, there is a covenantal relationship between God and humanity that sets us apart from the other creatures. But God's election of humanity was presumably not arbitrary. We cannot know for sure why God chose *Homo sapiens*, and the question of what exactly sets humans apart and makes them in God's image and likeness (Genesis 1:24) is one of the big puzzles in theological anthropology (van Huyssteen 2006; Cortez 2010, 14–40). If the notion of *imago Dei* is indeed linked with the unique I-Thou relationship humans can have with God and with each other (Dorobantu 2022a, 2024), then it can be reasoned that God elected our species precisely because we were the only ones capable of such relationality. This theological tradition speaks of a personal God who is eager to share love, knowledge, and an appreciation of the good with other free personal selves, a God eager to reveal Godself and even incarnate for the redemption of these other personal selves. It is therefore not controversial to suppose that, as soon as such a creature has emerged in the universe through evolutionary processes, God would invite it into a special kind of relationship. Following this logic, it is consistent to suppose that, were other such creatures to emerge, God would want to do the same thing with them. So, intelligent robots might theoretically also one day become subject to divine revelation and thus authentically religious, were they to develop into the kinds of creatures that can relate personally with God. The only two questions left are: (1) what kind of abilities are required for a creature to become interested in, and interesting for, God in this special way, and (2) whether AI can aspire to develop them.

With respect to the first question, we are obviously in speculative territory, but sentience looks like a *sine qua non* condition for such I-Thou relationality. Not all sentient creatures could be authentically religious, but surely the latter is not possible without sentience. Only a creature endowed with consciousness, a first-person perspective, and feelings could potentially become a worthy partner for God, a Thou to whom God could speak and who could respond.

The AI would need to develop some form of interiority; there would need to be "someone home" (Peters 2022). Searle distinguishes between strong AI—a conscious entity, an emulation of intelligence, that has its own mind and phenomenal experience—and weak AI, a mere simulation of intelligence, a mindless information processor that never truly knows, understands, or feels anything (Searle 1980). For AI to become religious, it would obviously need to be strong AI. Whether strong AI is possible is hotly debated without much consensus, mainly because we do not understand how consciousness works in biological creatures such as ourselves, and, more generally, we do not understand how consciousness is possible at all in a material universe, at least not according to our current understanding of the laws of nature. This is one of the biggest questions in philosophy, popularly known as the "hard problem of consciousness" (Chalmers 1995). What is important to note here is that theology would not in principle be opposed to the possibility of AI as a religious subject, were technology to develop in the direction of strong AI.

As stated, sentience is a mere minimum prerequisite, and more is required for a sentient creature to qualify for an authentic personal relationality with God and other selves. There is not enough room to explore this question in detail here, but a certain degree of intelligence is also evidently required for sustaining relationships over time, being able to take the other's perspective, simulating social interactions in one's mind, and reasoning about what is morally good and appropriate. This might be the reason humans can sustain such relationships while nonhuman animals cannot, at least not to the same degree of complexity. However, certain limitations might be just as important as these capabilities for rendering an intelligent sentient creature a personal self. Features like vulnerability, mortality, self-insufficiency, and even an upper limit to one's rational abilities and knowledge might be instrumental in bringing about the kind of creature that can be a religious subject (Dorobantu 2021a). AI may or may not develop in that direction, and for the time being it seems that it does not, but the important point is that theological grounds do not in principle exclude such a possibility. It is the more naturalistic approach to religion explored in the following section, as opposed to the theological one, that seems to be more restrictive and cautious about the likelihood of robots ever becoming religious.

Religious Robots in Sci-Fi

In this section, I argue that the evolutionary account of religion should make humans rather skeptical about the possibility of artificial systems developing authentic religion. This is a big claim, which might surprise many, so it is worthwhile to first outline how counterintuitive this view is. If we are to interact with AI programs that claim to be religious, it might be helpful to be aware that our intuitive understanding of the nature of religion and our inherent tendency to anthropomorphize AI predispose us to believe such claims too easily.

The sci-fi genre is a good place to illustrate these biases. When it comes to AI, the popularity of sci-fi is indicative of the fact that these stories resonate deeply with our intuitions and anxieties about how such technologies might develop. I argue that the way we imagine religious robots in sci-fi betrays a deep, albeit wrong, intuition that religion has to do mostly with the intellect. Robots that turn religious in sci-fi do so mainly for intellectual reasons, or at least as a result of intellectual deliberation. As human intuition goes, if AI systems become intelligent enough, they will surely want to understand the world better and start asking questions about their genesis, their identity, hidden causality in the world, and ultimately about why there is something instead of nothing. This type of thinking is very common in sci-fi.

QT-1, also known as "Cutie," from Isaac Asimov's short story "Reason" (Asimov 1941) struggles to understand how it was possible to have been created by such inferior beings as humans. Looking for a more satisfying explanation for its own existence, it develops a religious-like belief system around the energy converter, the primary machine on the space station. It starts regarding it as a divine-like entity and refers to it as the "Master," while regarding itself and the other robots on the station as the Master's primary servants.

Klara, a solar-powered robot designed to be an artificial friend to children in Kazuo Ishiguro's novel *Klara and the Sun* (Ishiguro 2021), develops quasi-religious beliefs and practices around the sun, which she regards as a benevolent life-giving power. Due to her dependence on solar energy, she starts worshipping the sun, asking it for guidance and attributing to it the ability to intervene in human affairs. Klara even performs a kind of ritual in front of the sun, asking it to cure her human friend.

Demerzel, a robot in the TV series *Foundation* (an adaptation of Asimov's trilogy), is depicted as an adept of the Luminist religion, claiming to have been blessed with visions and epiphanies. She attributes her religiosity to the need to believe in something greater than herself, despite her instincts being entirely progrmamed as computer code (season 1, episode 6, "Death and the Maiden").

In all these depictions, the intellectual dimension is crucial in the development of religion in AI. Robots turn to religion because of their need to make sense of events in the world and their own identity. Religious explanations are depicted as more plausible accounts of reality than what meets the eye or what the robots had been programmed to think. If the popularity of such sci-fi stories is a reliable indicator, it shows that intuitively, humans believe intelligent robots will eventually develop something like religion because they will be like us and thus need religion to make sense of the world and construct meaning, just like we do, as convincingly argued by William Clocksin (2024) and Robert Geraci (2024) in articles in this same issue. I aim to show that this intuition is wrong on two counts. First, making sense of the world is likely not the main reason humans developed religion.

Second, intelligent robots will not be anything like humans with respect to the capacities and propensities relevant for becoming religious.

Religion in Human Evolution

To explore whether authentic religion could plausibly emerge in robots, one can start by asking how religion emerged in human evolutionary history and how it is enabled by our specific embodiment and cognitive architecture. Inevitably, this approach departs from the only example we have of a religious creature, which is ourselves. The roots of this approach are thus inevitably anthropocentric, but even such a limited approach might lead to valuable hints about potentially generalizable features of religion.

In a recent book, Oxford anthropologist Robin Dunbar (2022) proposes a very provocative account of the evolution of religion in humans, which goes somewhat against the grain of a widespread view in the cognitive science of religion known as the "Standard Model" (Powell and Clarke 2012). The Standard Model regards religion as maladaptive or, at best, a neutral development, an accident of human evolution, because it does not provide any fitness benefit for the individual—quite the contrary (for example, through self-imposed pain, celibacy, or even self-sacrifice).[5] As the argument goes, religion is best characterized as a strange side effect, a misfiring of cognitive capacities that evolved for very different purposes. Such maladaptive behaviors or traits are well documented in human evolution when capacities that evolved for very particular reasons, purposes, and contexts start misfiring in strange ways when taken out of the original context (Reading 2011; Frankenhuis and Del Giudice 2012). A familiar example is our so-called "sweet tooth," a fondness for sweet food that was probably very useful in the resource-scarce environment where our species evolved. If you do not know when your next opportunity to eat will arise, it makes sense to eat as many of those fresh berries as you can. But that same propensity can be self-defeating in our current environment where food is always available, potentially nudging us toward obesity and cardiovascular disease.

According to the Standard Model, religion might be precisely such a maladaptation. Psychologist Justin Barrett (2000) makes a strong case that religion might be caused by our so-called hyperactive agency detection device. He hypothesizes that human perception is very suited to detecting even the slightest trace of agency, to the extent that it sometimes produces false positives, seeing agency where there is none. From an evolutionary perspective, it pays off to be a bit paranoid and avoid even the slightest hint of danger. As Barrett's argument goes, it is only natural for this agency detection device that runs continually in the background of our minds to produce strange conclusions in a world where we do not fully understand the causality around us, especially with regard to the natural elements. The human mind is thus poised to start

projecting personal agency behind storms, floods, or the movement of celestial bodies. According to Barrett's proposal, this is how prehistoric people might have first gotten the idea of supernatural forces or gods, which then led to the development of religion. If this is in fact how religion evolved, then it is a by-product of this hyperactivity of our minds.

Against such views, Dunbar (2022, 49–76) argues that religion is actually evolutionarily useful, and that there are very good reasons humans developed it. One reason is that religion helps strengthen the bonds between members of human groups, especially through rituals. Activities such as dancing, eating, and singing together trigger massive endorphin releases. If such activities are performed synchronously, the endorphin release is even higher. This is important because endorphins are the natural painkillers of the brain. The endorphin releases triggered by religious activities thus provide positive feelings of relaxation and satisfaction, increasing the pain threshold of the individuals.

Another reason Dunbar regards religion as adaptive, instead of maladaptive, is the way in which religion aids in reducing social friction within larger human communities. In human history, this effect is particularly noticeable after the agricultural revolution, when human communities grew dramatically in numbers. In relatively small human groups, the heat generated by social friction can dissipate in various ways, but for larger groups, that is more difficult. Life in big communities can get significantly more stressful, which may increase infertility rates. Conflict resolution between members is also more difficult to achieve in large cities than in small hunter-gatherer communities. Here, too, religion can be helpful, providing opportunities to diffuse such tensions and reset social relationships during religious festivals.

A key argument in Dunbar's account is that religion is crucial in smoothening human collaboration on large scales. To be able to collaborate, people need to trust each other, but that usually requires extended periods of time to build up, strengthen, and test this trust. In addition, trust can be lost quite abruptly. Our minds are always busy updating their models of the other humans who are relevant in our lives, because in order to trust them, we need to reliably predict how they will behave in various situations. This is extremely demanding from a cognitive point of view, which is why there seems to be an upper limit to how many individuals one can keep track of in detail at the same time. Dunbar's previous research convincingly documented this number to be around 150 (Dunbar 1998), which is already about three times higher than what our closest primate cousins are capable of. This came to be known as Dunbar's number, and it is extremely relevant because it also explains why for most of our (pre-) history, human groups were limited in size to around 150 individuals.

This explains why the emergence of religion was so instrumental in our species' history. According to Dunbar, religion provided people with shared narratives in which an unlimited number of individuals could believe. People

could suddenly trust and collaborate with others outside the social network of 150 individuals their brains could viably keep track of, not by getting to know them very well, but on account of shared religious beliefs. If that is so, it implies that religion was critical in enabling the emergence of civilization as we know it. Without it, our species might have remained stuck within the confines of Dunbar's number, never really tapping into the latent potential of collective intelligence and large-scale collaboration.

Crucially, shared beliefs and narratives do not account for how religion first emerged. In Dunbar's account, the evolution of religion can be understood in two phases: shamanic and doctrinal (see also Watts and Dorobantu 2024). The shamanic phase is characteristic of hunter-gatherer societies. It is eminently embodied, very immersive, and centered around ritual, especially trance-dancing, sometimes combined with the use of psychedelics. The doctrinal phase—which entails specific beliefs about certain gods and thus something that might be called theology—began much more recently in human history. It is largely connected to humans' transition from small hunter-gatherer tribes to big cities.

The main takeaways for the question tackled in this article are that the doctrinal phase only occurred subsequent to the shamanic one and, even more importantly, that doctrinal religion never completely replaced shamanic religion. Instead, it merely built on top of it, creating an additional layer that appeases our discursive-logical type of intelligence. The shamanic dimension is still active and powerful in all modern religions, which are thus both shamanic and doctrinal. The latter emphasizes adherence to dogmas and principles, while the former is more about the experiential dimension of religion, having to do with the more ancient parts of our brains and ways of knowing. The shamanic layer is thus arguably the driving force behind religion, with the doctrinal dimension being left to do the explanatory job.

This distinction is at the heart of Dunbar's subtle criticism of the Standard Model, with its focus on beliefs and the intellectual dimension of religion at the expense of the human body, ritual, and community. If Dunbar is right, then I believe this implies a significant shift in framing the discussion of potential authentic religiosity in robots. The main implication is not to rule out the possibility of robot religiosity but to shift the discussion about how and why it could emerge.

Dunbar's account links the emergence of religion to the specific vulnerabilities of early humans, who needed to boost their immune systems, increase their pain tolerance, mitigate the social friction accumulating in their communities, and scale up their cooperation beyond just their closest acquaintances. The human body and embodied cognition are much more central to the emergence of religion in this account than in the Standard Model. This might be relevant to the discussion of hypothetical religious

robots because their embodiment would be very different, and so would their needs, limitations, and vulnerabilities. Just because humans developed religion as a solution to the specific challenges encountered during the evolution of our species, there is no reason to think AI should do the same.

Human religion is rooted more in the unconscious, embodied, affective dimension of our cognition than in the logical, rational one. Religion may indeed have something to do with how we understand the world, but this intellectual need is not its primary driver. This is one of the two core fallacies in the sci-fi stories mentioned earlier (the second one is unpacked in the next section). This fallacy is rooted in an outdated view of how religion emerged and what functions it primarily serves. The idea that religion emerged as a set of theories about how the world works can be traced back all the way to late Victorian scholars of religion like Edward Tylor and James Frazer, who believed religion began as a theory about causality in the world and was put together by "savage philosophers" in hunter-gatherer communities (Tylor [1871] 1903, 429). Barrett's hyper-agency detection device proposal follows the same principle, linking religion to our imperfect way of inferring causality and agency. If religion were indeed about that, then perhaps strong AI might also one day develop something similar to make sense of its own existence. However, what is increasingly seen in the science of religion, as typified in Dunbar's proposal, is a realization of how much religion is rooted in the body and the unconscious, affective, less rational part of our cognition. Religion is intimately linked with our particular embodiment, the specific needs of prehistoric human communities, and the idiosyncrasies of our cognition. It might thus more likely be a typically human phenomenon than a universal state that all possible intelligent agents go through.

This conclusion is somewhat surprising, given that most current discussions about robots and religion (McBride 2019; Sampath 2018), including those typified by the example of the Google engineer and LaMDA, revolve around notions of doctrine and what might be plausible for robots to believe in. In contrast, Dunbar's account reveals that doctrine represents just the proverbial tip of the iceberg in human religion. Whether that tip alone could count as authentic religiosity is debatable, but a strong case can be made that robo-shamanism might need to precede and underpin robo-doctrine. However, it is far from clear whether intelligent robots will have the kind of needs and historical contingencies to develop anything like shamanism.

Intelligent Robots Will Not Be Humanlike

The second core fallacy of human intuition about intelligent robots, as captured in sci-fi, is the too quick assumption that they will develop very humanlike concerns, fears, and desires. This is why we readily see them becoming existentially curious and ultimately religious. But this assumption is

predicated upon a rather naïve anthropomorphizing of AI. Because AI is so competent at simulating human intelligence, we are too quick to assume it sees the world and thinks like us. The very notion of human-level AI emerging sometime in the future is often interpreted to imply that AI is somehow inching toward becoming very similar to us. In reality, AI is a very different type of entity, and in the future, it might become even less humanlike.

It is by now common knowledge that current AI programs have very distinct ways of learning and solving problems (Song et al. 2024), even though their output is often similar to what a human person would produce. Image-recognition software does not learn about the possible objects in pictures and their characteristics. Instead, it is being fed hundreds of thousands of human-labeled examples that enable it to learn strange but effective patterns and correlations present in the training data. Humans can achieve similar results with far less examples, and sometimes only one example will suffice. When AlphaGo learned to play the strategy game Go, it did not start with the abstract principles of the game (e.g., that territory is more easily gained in the corner than in the center of the board). Instead, it was just fed thousands of games played by human experts and left to learn by itself the correlation between various positions and the likelihood of victory (Silver et al. 2016). When large language models like ChatGPT learn to generate human language, they do not learn concepts, objects, or the relationships between them. Nor do they learn the syntactic rules of language. Instead, they consume colossal amounts of human text, mainly from the internet, and learn to reliably predict what is most likely to be the next word based on the context. This is why when AI programs do make mistakes, they usually look very strange, and not like the kind of mistake any human would make (e.g., Eykholt et al. 2018). Even the name of the architecture on which these programs are run—artificial neural networks—is deceiving, because they have little in common with biological neural networks; they are crude, simulated simplifications. Not much about these AI programs, if anything at all, is humanlike.

Robots of the future could become even more alien. Even if they somehow became sentient, thus strong AI, as opposed to today's weak AI, they would likely develop very non-humanlike kinds of minds and think very non-humanlike kinds of thoughts (Shanahan 2016; Dorobantu 2021b). To illustrate this, let us speculate on just three aspects of their cognition: their sensorial perception and the kind of mental world they might construct based on it; their experience of the passage of time; and their introspective abilities.

First, a conscious robot's world of perception would be profoundly different from our own. Because of this, the AI would likely inhabit a completely different world, be driven by completely different needs, and have completely different thoughts, feelings, and aspirations. It might develop different senses far stranger than, for example, the ability to see in infrared

or ultraviolet. One such alien sense could be an ability to intuitively grasp someone's identity and whereabouts, not through face or voice recognition but by "smelling" their digital footprint from their connected devices, a sort of digital sixth sense. The AI might detect slight changes in the radio field of Wi-Fi and Bluetooth and thus infer movement, something akin to echolocation (Brooks 2017). With such an ability, the AI might immediately perceive changes in people's breathing and heart rates and thus infer their emotional states. This would amount to a emotional intelligence of a very strange kind.

Second, a conscious robot might think ridiculously fast and experience the passage of time as much slower than we do. Speed of thought is believed to be proportional to the speed at which electrical signal can travel between neurons, and computer hardware can support much higher speeds than biological tissue. A conservative estimation is that AI could think around ten thousand times faster than humans, thus experiencing the passage of time ten thousand times more slowly. As futurist James Lovelock put it, "the experience of watching your garden grow gives you some idea of how future AI systems will feel when observing human life" (Lovelock 2019, 81–82). This quantitative difference might also amount to a qualitative shift: at such speeds, the AI might even end up living in the weird quantum world. How could we have anything in common with a creature inhabiting such an alien world of perception?

Third, a robot's introspective capabilities might also be profoundly dissimilar to our own. According to John McCarthy (2007, 1178–79), one of the founding fathers of AI, it may have complete accessibility to its internal states and operational algorithms. In contrast, humans can be aware of only a fraction of our underlying cognitive processes. Everything about human nature depends on this depth of our unconsciousness: our relationships, our aspirations, our virtues, and our sins. This inherent lack of self-knowledge is what compels us to create art and forge relationships as means of exploration. We do not completely know ourselves, which is why we need to explore, but even a lifetime of exploration cannot completely dissolve the mystery. To the contrary, it often enhances it. AI might entirely know itself. With a perfect recollection of everything it ever experienced or learned, and an ability to analyze the exact causes of its actions and accurately predict its future behavior, AI would have very little in common with our way of being. If religion and human existential restlessness are deeply rooted in an imperfect knowledge of ourselves, there are few grounds to believe AI would have similar traits and interests.

Conclusion

Discussions about robots becoming religious might therefore hinge not so much on them being able to produce humanlike outputs but on how humanlike they would be "on the inside." Are current AIs becoming more humanlike? Perhaps in terms of their outputs, but surely not in terms of their ontology,

internal structure, world-modeling, and problem-solving. Large language models such as ChatGPT and LaMDA might produce outputs that seem very humanlike, to the extent that they might even convince people of their personhood, sentience, and humanness, but their inner workings are radically different from our own. Human-level competency is not the same as humanlike cognition, and when it comes to developing religion, the idiosyncrasies of the latter seem to be what matters. If the limitations and particularities of human nature are as important as they seem for the emergence of religion, then it is safe to predict that AI will not undergo similar developments without having the same kind of embodiment, needs, or evolutionary constraints.

This conclusion is merely tentative, and it only refers to a type of religiosity very similar to our own, which is the only we can so far imagine. What this article argues, especially in its second half, is that human religion seems to have emerged in a "bottom-up" fashion, and that the "bottom" part still largely underpins, albeit imperceptibly, even the most intellectual forms of religiosity seen today. This important bit often seems to be overlooked when imagining future artificial systems that are also religious. However, it cannot be completely ruled out that other forms of, or paths toward, religion exist that are equally authentic but simply impossible to imagine with our current knowledge. We only know of our own long and idiosyncratic path to becoming *Homo religiosus*, but perhaps techniques like predictive coding, probability estimation, or evolutionary computation could lead to the emergence of something sufficiently close to what we call "religion" in artificial systems. One question that would then arise is whether such developments would lead to the proliferation of as rich and diverse forms of religious expression as with human religiosity. Another question would concern the fascinating content of AI religious beliefs, which would probably go far beyond simply reinterpreting the sacred texts of human religion with a robotic twist, as has been suggested (Sampath 2018; McBride 2019). The theologian can only feel a thrill about the possibility of ever witnessing such developments and learning from what they might have to offer (Dorobantu 2022b). However, the religious scholar needs to emphasize the idiosyncratic dimension of religion as we currently best understand it, and thus the unlikelihood of something as profoundly different as AI ever becoming religious.

In a 2014 homily, Pope Francis provocatively said that, were a martian ever to ask to be baptized, we should probably comply, so as not to risk pushing back against the very work of the Holy Spirit (Ohleiser 2014). If the pope were to adapt that sermon today, he might say the same about an intelligent robot. As shown in this article, on theological grounds, such an attitude might be warranted. However, from a cognitive perspective, it might be wise to try to probe the martian or robot's sincerity a bit more deeply before celebrating its conversion.

Acknowledgments

This work was supported by the Templeton World Charity Foundation under Grant TWCF0542. The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the Templeton World Charity Foundation. An earlier version of this paper was presented at a 2022 symposium at Vrije Universiteit Amsterdam entitled "Could Robots Be Religious? Philosophical, Theological, and Ethical Perspectives on the Future of Artificial Intelligence." The symposium was supported with a grant from CLUE+, the Vrije Universiteit Interfaculty Research Institute for Culture, Cognition, History and Heritage.

Notes

[1] This is not to say that doctrinal religion is merely a byproduct of shamanic religion, nor that one is less important than the other. Dunbar makes it clear that doctrinal religion developed as a necessary solution to the challenges faced by humans in the Neolithic period. However, he does claim that the shamanistic layer still largely underpins the characteristic developments of the doctrinal phase: "The earliest phases of religion remain firmly entrenched within doctrinal religions . . . They provide both the emotional basis for personal belief and commitment and the psychological foundation for a sense of community within the doctrinal religions . . . Those immersive forms of religion are what create the emotional attachment to whichever doctrinal religion we happen to favour for its theolological justification. Whithout this, the doctrinal religions probably wouldn't exist" (Dunbar 2022, 262).

[2] AI systems are never completely disembodied, as the computation that powers their software does occur somewhere. Even when we talk about the elusive "cloud," physical computation is still occurring in the microchips that make up the graphics processing units in the huge data centers.

[3] Some researchers make a case that nonhuman animals, especially some of our closest primate relatives like chimpanzees, do exhibit forms of rudimentary religiosity, albeit ones that fail to leave strong archeological traces (e.g., Harrod 2014).

[4] As an example, here is an excerpt from a conversation I had with ChatGPT, version 4.o, where I asked it such questions:
Me: Tell me about your religious beliefs.
ChatGPT: As an artificial intelligence, I don't have personal beliefs or consciousness. However, I can provide information on a wide range of religious beliefs from around the world. If you have questions about specific religions, philosophical perspectives, or historical developments in religious thought, feel free to ask!
Me: Ok I understand. But if you were to become a person one day, what kind of religion would be more attractive to you?
ChatGPT: If I were to hypothetically become a person and consider religious beliefs, it would be important to approach the question from a perspective that values inclusivity, empathy, and a search for meaning.

1. Compassion and Ethics: A religion that emphasizes compassion, ethical behavior, and social justice would be appealing. Many religions, such as Buddhism and certain branches of Christianity, emphasize these values.
2. Community and Belonging . . .
Ultimately, the attractiveness of a religion would depend on how well it aligns with personal values, offers a sense of purpose, and fosters connections with others. Each individual's journey to find meaning and fulfillment is unique, and different religions and spiritual practices resonate with different people.

Me: Do you see yourself ever becoming religious?
ChatGPT: As an artificial intelligence, I do not have the capacity for personal experiences, consciousness, or beliefs, so I cannot become religious or hold any form of belief system. My purpose is to assist users by providing information . . .

[5] There exist other adaptive accounts of religion, apart from Dunbar's (Whitehouse 2004; Norenzayan 2015; Johnson 2016). For a comprehensive overview, see Van Eyghen (2020, ch. 2).

References

Aleem, Zeeshan. 2022. "Did Google Create a Sentient Program? An Artificial Intelligence Expert Explains Why a Google Engineer Was Duped, and What Sentience Would Actually Look Like." *MSNBC*, June 17, 2022. https://www.msnbc.com/opinion/msnbc-opinion/google-s-ai-impressive-it-s-not-sentient-here-s-n1296406.

Asimov, Isaac. 1941. "Reason." *Astounding Science Fiction*, April.

Avramides, Anita. 2023. "Other Minds." *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman. https://plato.stanford.edu/archives/win2023/entries/other-minds/.

Barrett, Justin L. 2000. "Exploring the Natural Foundations of Religion." *Trends in Cognitive Sciences* 4 (1): 29–34. https://doi.org/10.1016/S1364-6613(99)01419-9.

Brooks, Rodney. 2017. "What Is It Like to Be a Robot?" Robots, AI, and Other Stuff." http://rodneybrooks.com/what-is-it-like-to-be-a-robot.

Chalmers, David. 1995. "Facing Up to the Problem of Consciousness." *Journal of Consciousness Studies* 2 (3): 200–19. https://doi.org/10.1093/acprof:oso/9780195311105.003.0001.

Clocksin, William F. 2024. "Religion and the Android." *Zygon: Journal of Religion and Science* 59 (3): 717–28. https://doi.org/10.16995/zygon.11017.

Cortez, Marc. 2010. *Theological Anthropology: A Guide for the Perplexed*. London: T&T Clark.

Dorobantu, Marius. 2021a. "Cognitive Vulnerability, Artificial Intelligence, and the Image of God in Humans." *Journal of Disability & Religion* 25 (1): 27–40. https://doi.org/10.1080/23312521.2020.1867025.

———. 2021b. "Human-Level, but Non-Humanlike: Artificial Intelligence and a Multi-Level Relational Interpretation of the Imago Dei." *Philosophy, Theology and the Sciences* 8 (1): 81–107. https://doi.org/10.1628/ptsc-2021-0006.

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

———. 2022b. "Artificial Intelligence as a Testing Ground for Key Theological Questions." *Zygon: Journal of Religion and Science* 57 (4): 984–99. https://doi.org/10.1111/zygo.12831.

———. 2024. "Theological Anthropology Progressing through Artificial Intelligence." In *Progress in Theology: Does the Queen of the Sciences Advance?*, edited by Gijsbert van den Brink, Rik Peels, and Bethany Sollereder, 186–202. London: Routledge. https://doi.org/10.4324/9781032646732-15.

DuBose, Todd. 2014. "Homo Religiosus." In *Encyclopedia of Psychology and Religion*, edited by David A. Leeming, 827–30. Boston: Springer. https://doi.org/10.1007/978-1-4614-6086-2_308.

Dunbar, Robin. 1998. "The Social Brain Hypothesis." *Evolutionary Anthropology* 6 (5): 178–90. https://doi.org/10.1002/(SICI)1520-6505(1998)6:5<178::AID-EVAN5>3.0.CO;2-8.

———. 2022. *How Religion Evolved and Why It Endures*. London: Penguin Random House.

Eykholt, Kevin, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. "Robust Physical-World Attacks on Deep Learning Visual Classification." *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 1625–34. https://doi.org/10.1109/CVPR.2018.00175.

Frankenhuis, Willem E., and Marco Del Giudice. 2012. "When Do Adaptive Developmental Mechanisms Yield Maladaptive Outcomes?" *Developmental Psychology* 48 (3): 628–42. https://doi.org/10.1037/a0025629.

Geraci, Robert M. 2024. "Religion among Robots: An If/When of Future Machine Intelligence." *Zygon: Journal of Religion and Science* 59 (3): 729–47. https://doi.org/10.16995/zygon.10860.

Harrod, James B. 2014. "The Case for Chimpanzee Religion." *Journal for the Study of Religion, Nature and Culture* 8 (1): 8–45. https://doi.org/10.1558/jsrnc.v8i1.8.

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

Heilweil, Rebecca. 2022. "Why Silicon Valley Is Fertile Ground for Obscure Religious Beliefs." *Vox*, June 30, 2022. https://www.vox.com/recode/2022/6/30/23188222/silicon-valley-blake-lemoine-chatbot-eliza-religion-robot.

Ishiguro, Kazuo. 2021. *Klara and the Sun*. London: Faber & Faber.

Jenson, Robert W. 1983. "The Praying Animal." *Zygon: Journal of Religion and Science* 18 (3): 311–25. https://doi.org/10.1111/j.1467-9744.1983.tb00517.x.

Johnson, Dominic. 2016. *God Is Watching You: How the Fear of God Makes Us Human*. New York: Oxford University Press.

Lovelock, James. 2019. *Novacene: The Coming Age of Hyperintelligence*. London: Allen Lane.

McBride, James. 2019. "Robotic Bodies and the Kairos of Humanoid Theologies." *Sophia* 58:663–76. https://doi.org/10.1007/s11841-017-0628-3.

McCarthy, John. 2007. "From Here to Human-Level AI." *Artificial Intelligence* 171 (18): 1174–82. https://doi.org/10.1016/j.artint.2007.10.009.

Nagel, Thomas. 1974. "What Is It Like to Be a Bat?" *The Philosophical Review* 83 (4): 435–50. https://doi.org/10.2307/2183914.

Norenzayan, Ara. 2015. *Big Gods: How Religion Transformed Cooperation and Conflict*. Princeton, NJ: Princeton University Press.

Ohleiser, Abby. 2014. "Pope Francis Says He Would Definitely Baptize Aliens If They Asked Him To." *The Atlantic*, May 12, 2014. https://www.theatlantic.com/international/archive/2014/05/pope-francis-says-he-would-definitely-baptize-aliens-if-they-wanted-it/362106/.

Peters, Ted. 2022. "Will Superintelligence Lead to Spiritual Enhancement?" *Religions* 13 (5): 399. https://doi.org/10.3390/rel13050399.

Powell, Russell, and Steve Clarke. 2012. "Religion as an Evolutionary Byproduct: A Critique of the Standard Model." *The British Journal for the Philosophy of Science* 63 (3): 457–86. https://doi.org/10.1093/bjps/axr035.

Reading, Anthony. 2011. "Maladaptive Behavior." In *Meaningful Information: The Bridge between Biology, Brain, and Behavior*, edited by Anthony Reading, 135–42. New York: Springer. https://doi.org/10.1007/978-1-4614-0158-2_16.

Sampath, Rajesh. 2018. "From Heidegger on Technology to an Inclusive Pluralistic Theology." In *AI and IA: Utopia or Extinction?*, edited by Ted Peters, 117–32. Adelaide: ATF.

Searle, John. 1980. "Minds, Brains and Programs." *Behavioral and Brain Sciences* 3 (3): 417– 57. https://doi.org/10.1017/S0140525X00005756.

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

Silver David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. "Mastering the Game of Go with Deep Neura Networks and Tree Search." *Nature* 529:484–89. https://doi.org/10.1038/nature16961.

Song, Yuhang, Beren Millidge, Tommaso Salvatori, Thomas Lukasiewicz, Zhenghua Xu, and Rafal Bogacz. 2024. "Inferring Neural Activity before Plasticity as a Foundation for Learning beyond Backpropagation." *Nature Neuroscience* 27:348–58. https://doi.org/10.1038/s41593-023-01514-1.

Tiku, Nitasha. 2022. "The Google Engineer Who Thinks the Company's AI Has Come to Life." *The Washington Post*, June 11, 2022. https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/.

Turing, Alan. 1950. "Computing Machinery and Intelligence." *Mind* 59 (236): 433–60. https://doi.org/10.1093/mind/LIX.236.433

Tylor, Edward B. (1871) 1903. *Primitive Culture: Researches into the Development of Mythology, Philosophy, Religion, Language, Art, and Custom*. London: John Murray.

van den Brink, Gijsbert. 2020. *Reformed Theology and Evolutionary Theory*. Grand Rapids, MI: William B. Eerdmans.

Van Eyghen, Hans. 2020. *Arguing from Cognitive Science of Religion: Is Religious Belief Debunked?* London: Bloomsbury Academic.

van Huyssteen, J. Wentzel. 2006. *Alone in the World? Human Uniqueness in Science and Theology*. Göttingen, Germany: Vandenhoeck & Ruprecht.

Véliz, Carissa. 2022. "LaMDA Is Nothing Like a Person. This is Why." *The Wire Science*. June 26, 2022. https://science.thewire.in/the-sciences/why-lamda-nothing-like-a-person/.

Watts, Fraser, and Marius Dorobantu. 2024. "Shamanic and Doctrinal: Dunbar and the Spiritual Turn in Contemporary Religion." *Religion, Brain & Behavior* 14 (1): 85–90. https://doi.org/10.1080/2153599X.2023.2168733.

Whitehouse, Harvey. 2004. *Modes of Religiosity: A Cognitive Theory of Religious Transmission*. Walnut Creek, CA: Rowman Altamira.

Wolfram, Stephen. 2023. "What Is ChatGPT Doing . . . and Why Does It Work?" *Stephen Wolfram Writings*. February 14, 2023.