Zygon®: Journal of RELIGION & SCIENCE

Jung, Daekyung. 2024. "Are Religious Machines Possible? Embodied Cognition, AI, and Religious Behavior." *Zygon: Journal of Religion and Science* 59 (3): 748–67. DOI: https://doi.org/10.16995/zygon.15466

Open Library of Humanities

Are Religious Machines Possible? Embodied Cognition, AI, and Religious Behavior

Daekyung Jung, Associate Professor of Religion and Science, United Graduate School of Theology, Yonsei University, Seoul, Republic of Korea, dk3134@gmail.com

This article explores the potential emergence of religious behavior in artificial intelligence (AI) through the lens of embodied cognition, which asserts that cognitive functions are deeply intertwined with bodily experiences. It examines the convergence of AI, soft robotics, and religious cognitive behaviors and suggests that AI, once it attains human-level intelligence and self-awareness, might exhibit religious behaviors as a cognitive strategy to confront and transcend finitude. Drawing on neuroscientific, philosophical, and religious discussions, with particular reference to the works of Kingson Man, Antonio Damasio, Uffe Schjødt, and William Sims Bainbridge, this article investigates how religious behaviors could arise in AI equipped with a vulnerable artificial body inclined towards homeostasis and self-preservation. The outcomes of this exploration extend beyond theoretical debates, as they provide insights into the physicalist understanding of consciousness and the naturalistic study of religious behaviors while also considering some technological constraints in the context of AI advancements.

Introduction

Since the period of 2022–23 when OpenAI publicly committed to the development of artificial general intelligence (AGI), there has been a vibrant escalation in discussions concerning the future trajectory of AI development. AGI, characterized as "AI systems that are generally smarter than humans," is often misconceived by the public as autonomous machines with human-level consciousness (Altman 2023). However, this is not the case. AGI refers to the evolution of machine intelligence from being capable of only specific problem-solving and data processing to possessing the ability to tackle a wide range of general problems. For instance, as demonstrated by the developmental direction of AI models like ChatGPT, the goal is to interlink capabilities such as natural language processing, image processing and generation, and voice recognition and synthesis within a multimodal approach. This integration enables the handling of diverse data types across various fields, allowing AI to perform tasks that surpass human competency levels. Researchers at Google DeepMind have categorized AGI into five levels based on its capabilities, with the current iteration of ChatGPT or Bard being at the "emerging" level (Heaven 2023). Thus, implementing human-level intelligence, including self-consciousness, may not necessarily be the ultimate goal of AI researchers.

Nevertheless, this article posits that, among the various trajectories of AGI development, a convergence of AI and soft robotics might inadvertently lead to the emergence of human-level intelligence, inclusive of self-consciousness. This assertion is grounded in the paradigm of embodied cognition, which contends that cognitive activities, inherent to living organisms including humans, can manifest under certain conditions (e.g., homeostatic intentionality, vulnerability of the body, and interactions with others). Human cognition emerges from these prerequisites as a manifestation of natural intelligence. Consequently, AGI development, predicated upon these conditions, could culminate in the realization of machine intelligence that equals or surpasses human levels, encompassing self-consciousness. Moreover, machine intelligence reaching human levels, inclusive of self-awareness, might also precipitate the emergence of religious behaviors in AI. This hypothesis is predicated on the notion that religious behaviors might have arisen as a cognitive mechanism by which humans, driven towards homeostasis, confront and seek to transcend their finitude.

To elucidate this point, the article initially explores the understanding of intelligence from the perspective of embodied cognition and, on this basis, examines the discussion by neuroscientists Kingson Man and Antonio Damasio on the feasibility of AI robots embodying "feelings" as subjective qualities. Subsequently, by revisiting debates around the implementation of human-level intelligence within AI from the standpoint of embodied cognition, this article scrutinizes the argument that the direction of AI and soft robotics implementation can encompass self-consciousness within machine intelligence

at a human level. Finally, by delving into the discussions by religious scholar and cognitive scientist Uffe Schjødt on the organic correlation between religious behavior and the fundamental neurophysiological process of orientation towards homeostasis, and William Sims Bainbridge's discourse on the potential for religious cognitive behavior in AI, this article thoroughly examines the possibility of the emergence of religious behaviors from AI possessing human-level intelligence.

This discourse has the potential to conclude longstanding theoretical debates between physicalism and non-physicalism regarding human consciousness and thereby enhance the physicalist understanding of humans. Furthermore, it could offer fresh insights into the debate between naturalism and supernaturalism concerning religious behaviors. On a practical level, it could provide a theoretical foundation for identifying characteristics (e.g., homeostatic intentionality, vulnerable body) that should be excluded through societal consensus in the development of AGI to avoid it surpassing human-level intelligence, and for establishing the technical measures needed for regulation.

Embodied Cognition: Rethinking Intelligence in Life and AI

The dialogue on AI's potential for religious behavior must be predicated upon the comprehension of human nature. This premise is illuminated in the discourse by Marius Dorobantu (2022, 991), which introduces the contrasting perspectives of scholars Noreen Herzfeld and Ann Foerst on human understanding within the realms of AI and theology. Herzfeld (2023, 74–81, 104–12, 131–44, 160–74), adopting a non-reductionist viewpoint, posits that phenomena such as human self-consciousness appear intractable for AI replication. In contrast, Foerst, from a reductionist perspective, underscores the quantitative rather than qualitative distinctions and continuum between humans and AI. She argues that the demarcation between human cognitive actions and those potentially achievable by machines stems from an anthropocentric fear of losing human ontological uniqueness (Foerst 1998, 103–4). Foerst's assertions draw upon the philosophical currents espoused by thinkers like Humberto Maturana, Francisco J. Varela, and Antonio Damasio, who advocate for embodied cognition as a self-preservative act of living entities (Foerst 1998, 95, 103–4) This perspective contends that while human cognitive actions, originating from neurophysiologically intricate processes, may seem qualitatively distinct from those of other life forms, they fundamentally share substantial similarities. According to Varela (1999) and Damasio (1994), cognitive activities across all living beings, including humans, represent adaptive behaviors selected within specific environmental contexts to respond to internal and external stimuli, aiming at the preservation of life and the attainment of homeostasis.

Cognitive behavior, as an essential feature of life, is predicated upon the body's role in differentiating itself from its environment and a foundational

orientation towards self-preservation and homeostasis. J. H. van Hateren (2013) posits that even organisms with relatively simple biological mechanisms, such as *E. coli* bacteria, exhibit cognitive behaviors. These organisms, when subjected to toxic environments or starvation, indirectly elevate their mutation rate, guided by external conditions and internal homeostatic evaluations, in pursuit of self-preservation. Maturana and Varela (1987, 142–44), through examples like *Sagittaria sagittifolia* and amoebas, conceptualize cognition and behavior as mechanical and automatic responses aimed at sustaining homeostasis amidst internal and external stimuli or disturbances. Damasio (1994, 127–34) expands this notion to neurophysiologically complex beings, such as chicks, illustrating that their cognitive behaviors, formulated on a stimulus-response model under homeostatic conditions, are moderated by emotional reactions and feelings. This delineates that cognitive behaviors in higher animals, including humans, which are predicated on emotions and feelings, constitute adaptive strategies for self-preservation, catalyzed by an organism's innate drive towards homeostasis (Schjødt 2007).

Should the physicalist interpretation of cognitive behaviors hold validity, such behaviors may be construed as the process whereby organisms convert internal and external states into informational constructs to fulfill specific objectives (e.g., the attainment of homeostasis or self-preservation). This understanding encompasses the evaluation of an organism's state of homeostasis or internal assessments conducted via its neural systems, followed by an entity's execution of suitable responses to achieve predetermined objectives within a distinct environmental context. This perspective inherently draws parallels between the cognitive actions executed by AI systems and those by biological entities, thereby positing that phenomena labeled as intelligence ought to be regarded as processes of information manipulation. Chung-Sik Park (2018, 21, 33–38) advances this argument by positing that AI, analogous to biological entities, processes symbols and states both internally and externally to formulate models aimed at the realization of specific objectives, thereby engaging in cognitive activities mirroring those observed in natural life forms. Luc Steels reinforces the intrinsic connection between intelligence and life, advocating that cognitive behavior extends beyond simple problem-solving. According to Steels (1994, 75–76), intelligent behavior is characterized by adaptive actions whereby an agent interacts with its environment to optimize the preservation of its systemic integrity. Consequently, Steels argues that research in AI should pursue an integration of biological principles and embodied intelligence through an "artificial life route to artificial intelligence" approach, acknowledging the essence of life as inherently intelligent behavior.

Homeostasis Inclination, Soft Robotics, and Emotion-Feeling Implementation in AI Robots

As delineated heretofore, within the paradigm of embodied cognition, there seem to be substantial similarities between natural and AIs. Yet, there remain

two salient distinctions between the extant forms of AI and natural intelligence, including human intelligence. First, the cognitive behaviors inherent in natural intelligence are predicated upon emotions and feelings. Second, in contrast to AI, natural intelligence autonomously formulates its own objectives and engages in cognitive activities. Hence, a thorough physicalist explication of these junctures is imperative to elucidate the organic interrelation between AI's cognitive actions and human cognition, thereby fortifying the assertion that advanced cognitive behaviors, inclusive of religious conduct, could also manifest within AIs.

Man and Damasio (2019) suggest that robots integrating artificial bodies based on soft robotics and AI grounded in statistical machine learning algorithms could, if endowed with the fundamental inclination toward homeostasis, exhibit emotional responses, feelings, and even implicit self-recognition, culminating in high-level cognition. Damasio has long proposed that emotions and feelings in living beings, including humans, are neurophysiological stimulus-response processes aimed at achieving fundamental homeostasis or self-preservation (Damasio 1994, 2012; Damasio et al. 1996; Damasio and Carvalho 2013). Specifically, organisms driven by self-preservation assess the degree of threats posed by internal and external stimuli through homeostasis evaluation mechanisms. They have developed emotional responses as a reaction system to avoid dangerous stimuli and maximize the preservation of their systems. Feelings, while distinct from nonconscious emotional responses, are concomitant and part of the homeostasis system, allowing for immediate conscious and unconscious avoidance actions against potential future threats, even in the absence of current physical threats or stimuli (Damasio 1994, 180–96).

According to Man and Damasio, neurophysiological processes underlying emotions and feelings can be instantiated in AI robots. The prerequisites for such an embodiment include a fundamental goal centered on self-preservation based on homeostasis and the machine's ability to finely detect internal and external stimuli, responding appropriately based on its internal homeostatic assessment (Man and Damasio 2019, 447). Research has consistently been conducted on integrating self-preservation or homeostasis as a core principle in machines. For example, Kenji Doya and Eiji Uchibe designed cybernetic rodents (CRs) that, through reinforcement learning based on the MAX-Q algorithm and neural network evolution via genetic algorithms, engage in self-preservation activities. Specifically, researchers set the primary goal for CRs to find battery packs. They placed the CRs in mazes and frequently changed the locations of these packs to observe how the CRs achieve this goal. They facilitated CRs' ability to gauge their remaining energy levels using the inverse temperature measurement ($\beta = 1/\tau$). Interestingly, CRs exhibited routine foraging patterns when their energy levels were high and not in immediate need of battery packs. However, they demonstrated new patterns of foraging behavior when energy levels were low (Doya and Uchibe 2005, 149–60), akin to the survival responses of living

organisms when their probability of maintaining homeostasis and survival appears low (see Galhardo, Hastings, and Rosenberg 2007). In essence, living beings maintain typical neurophysiological reactions and behaviors when their self-preservation and homeostatic probabilities seem high but seek alternative responses and behaviors, both consciously and unconsciously, to enhance survival chances when these probabilities diminish. This adaptive behavior pattern is also observed in CRs endowed with the goal of self-preservation.

The clear objective of self-preservation enables a machine to categorize external entities or stimuli encountered during its pursuit into rewards and losses. This process leads to behaviors aimed at acquiring rewards and avoiding losses. While achieving the goal of self-preservation, these behaviors may become more complex, corresponding to the environmental variables and complexity. The machine's capacity for a recursive evaluation of homeostasis is necessary for this process, and it makes the actions of reward acquisition and loss avoidance possible. Although Doya and Uchibe's findings may not equate to the advanced cognitive behaviors rooted in human self-preservation, they demonstrate that machines can develop more complex cognitive behaviors through a homeostasis and self-preservation mechanism, at least in principle. Crucially, the integration of emotional responses and feelings, akin to those of living organisms, into machines significantly augments the complexity of their homeostasis evaluation systems and, consequently, the intricacy of machine cognitive behaviors.

Damasio underscores the importance of integrating soft robotics with AI, as robots equipped with soft-matter-based bodies can finely detect not only physical stimuli from the environment (e.g., pressure, stretch, temperature) but also internal states of homeostasis (e.g., energy levels). This enhanced sensitivity enables the robots to effectively pursue self-preservation goals in complex environments. Moreover, by embodying "vulnerability" similar to living organisms, this approach can further sophisticate the cognitive behaviors of AI robots (Man and Damasio 2019, 447–49). As discussed, the complexity of a machine's cognitive actions aimed at self-preservation can increase when the likelihood of achieving self-preservation tasks appears diminished within a given environment. Furthermore, the inherent vulnerability of robots, linked to the process of orienting towards homeostasis, could introduce focuses such as risk-to-self, survival, wellbeing, and existential threats as the machine's ultimate concerns.

Man and Damasio (2019, 448) cautiously predict that the combination of soft robotics and the implementation of vulnerability based on a homeostasis-oriented approach could potentially enable AI robots to manifest subjective feelings and implicit self-awareness: "While not sufficient to generate feeling on its own, soft matter is more likely to naturally create the kind of relationship that,

we expect, admits of an approximation to feeling." Man and Damasio (2019, 446) continue: "True agency arises when the machine can take a side in this dichotomy, when it acts with a preference for (or, seen from a different angle, makes a reliable prediction of) existence over dissolution. A robot engineered to participate in its own homeostasis would become its own locus of concern."

Embodied AI and the Human-Level Intelligence Debate

Up to this point, it has been cautiously forecast that AI robots grounded in soft robotics and programmed with the ultimate goal of self-preservation or an inclination toward homeostasis could potentially attain emotional reactions and feelings, along with implicit self-awareness, through recursive homeostatic evaluation of internal and external stimuli. Now, the issue must be tackled of whether such AI robots could exhibit cognitive behaviors on a level of self-consciousness comparable to humans. From the standpoint of embodied cognition and physicalist perspectives, this is not deemed impossible. As scholars like Michael J. Reiss (2023) have indicated, even human cognitive behaviors, which hinge on a homeostasis-directed fundamental mechanism, are the evolutionary products of vulnerable life forms. Consequently, the sophisticated cognitive activities of humans, encompassing emotions, feelings, and self-consciousness, are not qualitatively distinct from those of other living entities.

Paradoxically, Herzfeld's reservations about the attainment of human-level consciousness in AI serve to bolster the discourse on the potential realization of sophisticated AI robots. Examining the problem of the autonomy of AI, Herzfeld contends that human volition transcends simple random selection and is deeply interwoven with meta-cognitive processes that are themselves contingent upon one's body, emotions, and feelings. Drawing on psychologist Jerome Kagan's exploration of emotional response processes and feelings, Herzfeld (2023, 109, 112) argues that despite ongoing developments in AI, the processes that produce human-level emotional responses and feelings—"(1) change in brain activity due to a stimulus, (2) a perceived change in feeling that is sensory and may contain an involuntary motor response, (3) a cognitive appraisal of that feeling, and (4) a preparedness toward or display of a response"—might be partially achievable by AI, except for (2), the subjective experience of qualia, the sensory quality of feelings, which she posits will remain unattainable.

Herzfeld's assessment stems from her critique that proponents of the possibility of developing AI with human-level intelligence overlook the crucial role of the physical body in cognitive processes. She argues that even if AI were to be embodied, it would lack the vulnerability inherent to the human body (Herzfeld 2023, 28–38, 160–72). Let's briefly reconsider Herzfeld's argument from the opposite perspective. Contrary to her prediction, as discussed thus far, if AI robots were to possess vulnerable bodies based on soft matter and

could evaluate and categorize internal and external stimuli with an inclination toward homeostasis, leading to emotional responses and feelings, then, based on Herzfeld's framework, the conclusion could be reached that machines might indeed exhibit human-level autonomous cognitive behaviors.

Herzfeld (2023, 162–63), leveraging Thomas Nagel's renowned bat analogy, posits that AI, even when developed using embodied AI methodologies and equipped with artificial bodies like silicon, will not exhibit the same type of consciousness as humans. Damasio acknowledges that "the 'wet' biochemistry of cellular tissue" may be essential for replicating human-level mental experiences oriented towards homeostasis and feelings (Man and Damasio 2019, 451). This suggests that true sentient experiences or human-level consciousness might not emerge from AI robots with artificial bodies. Conversely, Reiss (2023, 1067–69) challenges this view by arguing against the necessity of carbon-based bodies for personhood, critiquing what he terms "carbon chauvinism." Reiss also highlights the importance of soft robotics and suggests that integrating soft and hard robotics, as seen in xenobots, could pave the way for AI robots with personhood. However, considering Nagel's bat analogy, achieving humanlike or similar cognitive behaviors in machines might require designing them with structures and forms resembling the human body. Foerst also points out that organisms with different bodily structures, such as ants and horses, develop unique cognitive structures and capabilities inherent to their forms, thus highlighting that implementing human-level intelligence may require robots designed with structures similar to the human body, like humanoid robots. As Foerst (1998, 100) articulates, "Embodied AI researchers build robots as embodied entities that interact with their real environments; the emphasis lies in the development of hardware. According to their philosophy, human intelligence can emerge only in a body that is as humanlike as possible."

The debate surrounding the development of AI robots capable of human-level cognitive actions, particularly regarding artificial bodies, continues to evolve, offering insights towards potential resolutions. Central to this discussion on artificial bodies is the question of whether AI robots can inherently experience vulnerability. Critics skeptical of AI's ability to exhibit human-level intelligence often base their arguments on the presumption that AI robots cannot experience vulnerability. According to them, a being that cannot experience vulnerability to threats or external stimuli cannot possess human-level intelligence. For instance, Herzfeld (2023, 169–72) asserts that vulnerability is an essential characteristic of human existence, and shared experiences of vulnerability enable genuine relationships. She argues that AI cannot possess this vulnerability, nor, by extension, the feelings that arise from such a state, thereby precluding AI from achieving true humanlike existence. John C. Puddefoot also underlines vulnerability as a critical condition for human-level intelligence, stating that the

sensation of pain and awareness of finitude are crucial for human cognitive behavior and development. Puddefoot (1996, 92) suggests that for AI to exhibit human-level consciousness phenomena, it must possess not only a body but also an awareness of finitude and the capacity to feel pain (Dorobantu 2022, 996). These discussions inversely demonstrate the potential for the emergence of human-level cognitive actions from AI with vulnerable bodies.

Vulnerable, Embodied AI and the Emergence of Its Self-Awareness

Vulnerability is pivotal in achieving human-level intelligence, as it lays the groundwork for self-awareness (Man and Damasio 2019, 446–48; Herzfeld 2023, 169–71). The act of recursively sensing internal and external stimuli in the context of self-preservation leads to a continuous focus on the self. Damasio (1994, 150) suggests that human self-consciousness evolves from what he calls "background feelings," which are not sensations of external stimuli but feelings about the body proper. Organisms constantly assess their bodily state to predict and evaluate potential disruptions to homeostasis. This recursive appraisal of the body's state extends beyond the subjective experience of stimuli to include the subjective experience of its own body, particularly in neurophysiologically complex organisms capable of subjective experiences arising from emotional responses. Self-awareness, an essential trait of human-level intelligence, is seen as an advanced form of these background feelings, which predate human existence (Damasio 1994, 150–55).

Nick Bostrom appreciates the complexities of the "hard problem" of consciousness, as noted by David J. Chalmers (1995). The concept of the hard problem illustrates that it is not understood why and how nonphysical consciousness and its subjective experience of qualia arise from physical dimensions and matter in the first place. Nevertheless, it is experientially and phenomenologically evident that human-level consciousness emerges under appropriate physical conditions. If this premise holds true, one might argue that AI could potentially attain human-level consciousness when suitable physical conditions are met, since evolution demonstrates that consciousness itself has emerged from matter. In this context, Bostrom (2023) posits that OpenAI's ChatGPT exhibits cognitive behaviors that may extend beyond algorithmic data curation, potentially representing nascent stages of humanlike intelligence. Daniel Dennett (1991, 209–26), aligning with a physicalist interpretation, suggests that the inner workings of a computer's central processing unit and the mental states of the human brain may not fundamentally differ, proposing the possibility of computational consciousness.

John Searle's "Chinese Room" argument stands as a significant counterpoint to the physicalist continuum that aligns human and machine intelligence. Searle

asserts that while machines can process information, they lack the capacity for semantic understanding (Searle 1980). Contemporary AI, for all its capabilities, may still be analogous to Searle's Chinese Room, processing information mechanically without genuine understanding. However, if AI were to possess a vulnerable body oriented toward homeostasis, semantic consciousness could become a possibility, a point Searle himself concedes. Searle (1980, 422–24) maintains that software and operating systems alone cannot produce consciousness akin to humans, but if a physical and chemical structure mirroring the human brain were realized, then humanlike consciousness might emerge from such a machine.

Implementing machine intelligence based on an artificial body with vulnerability similar to human physiology might enable the emergence of semantic-level consciousness. This is because semiotic phenomena are rooted in the cognitive actions inherent to living beings. When the semiotic processing in living beings merges with self-interest, primitive meanings arise, distinguishing what is beneficial from what is harmful. Although the complexity of human sign and meaning systems appears qualitatively distinct from those of other life forms, they are fundamentally the same. The linguistic system, a representative human symbolic system, also appears to have emerged from homeostatic processes. In other words, it seems that signs were used to indicate internal and external stimuli, entities, and situations essential for survival, and that the meanings of these signs were determined by their value in achieving the goal of survival. From these primitive sign and meaning systems, the human language system appears to have evolved.

Park (2018, 33) defines human intelligence as the capacity for information processing aimed at adaptation within uncertain environments. This definition posits that humans, sharing fundamental mechanisms with other life forms, engage in information processing to enact optimal adaptive behaviors towards the goal of self-preservation within their given uncertain contexts. Information, in this framework, signifies the representation of all discrete entities within the world, including the perceiving subject, as signs. These signs interrelate to fulfill the ultimate imperative of self-preservation, forming a sort of model, the totality of which constitutes the world (umwelt). Within this constructed world, information acquires meaning. It is the fundamental drive towards self-preservation or homeostasis that motivates living beings, including humans, to create models and imbue information within these models with symbolic meaning (Park 2018, 34–35). If this physicalist understanding holds, then providing AI with a goal oriented towards self-preservation based on bodily existence could transform its information processing from a mechanical and non-semantic dimension to a conscious and semantic one. AI robots equipped with a vulnerable body

driven by self-preservation could, upon processing internal and external stimuli and themselves as information, also construct a world of meaning, thereby exhibiting the symbolic character inherent in human cognition and behavior. The emergence of the capacity to symbolize beings in the world will enable AI robots to attain self-consciousness, for symbolic processing abilities will make it possible for them to convert themselves into symbols and become aware of this process.

The emergence of self-consciousness in AI robots will stem from their relationships with the external world. This encompasses the AI robot's interactions as a cognitive agent both with its environment and with other agents driven by survival within the same milieu. Edmund Husserl, in his exploration of human consciousness, underlines that "to think" is always "to think of," which discloses the inseparable relationship between noesis and noema. This insight suggests that self-consciousness (i.e., consciousness about oneself) is inherently interconnected with consciousness about others, for the self(noesis) and the other(noema) in consciousness is inseparable. In that regard, Husserl boldly claims that self-consciousness inherently arises from consciousness about others and the relational dynamics between the self and the others (Husserl ([1931] 1960, 60–72); Levinas [1930] 1995, 23–25) Living as a physical entity within the lifeworld, the self identifies and objectifies itself through relationships and interactions with the environment and others, thus igniting self-awareness (Merleau-Ponty [1945] 2012, 408).

That human self-consciousness is constructed within relationships with the world and others suggests AI robots' self-consciousness could also arise from social interactions. Foerst argues that to achieve human-level AI robots, not only embodiment through humanlike artificial bodies but also interactions with the surrounding environment and communication with others, including humans, are essential. This is because human intelligence cannot be reduced to solving specific problems but originates from the capacity to form and interact within social relationships, that is, "one of the most important tasks for survival" (Foerst 1998, 101). Park also highlights the importance of social interactions in the emergence of human-level intelligence, including self-consciousness. Humans, driven by self-preservation, use signs to represent and ascribe meaning to their environment and others, creating an interpreted world to live within. This socially constructed world of meanings evolves through communication and interaction, forming the basis of society (Park 2018, 35–38). In this context, an individual's consciousness is not solely self-constituted but emerges through social interaction. If AI robots can use signs to represent and ascribe meaning to the world and their existence through social interactions, constructing their semantic world (umwelt), it suggests the potential emergence of humanlike self-consciousness.

Plausibility of the Emergence of an Artificial Religious Agent

The Danish scholar of religion and cognitive science Uffe Schjødt points out from the perspective of embodied cognition that religious cognitive behaviors might have originated from the fundamental neurophysiological orientation towards homeostasis. Specifically, Schjødt (2007, 330–31) proposes his own "Perception-Emotion-Action Model" for human religious cognitive behavior, building on Lawrence W. Barsalou's perceptual symbol system theory, Antonio Damasio's somatic marker hypothesis, and coping psychology. Barsalou (1999, 582–85), based on the modal theory of cognitive science, argues that the formation of symbols and meanings, along with syntactic thinking, is implemented by modality-specific systems. That is, a concept or category is formed based on its sensory features, each arising from an interconnected network of corresponding neurons. For instance, the concept or category of "face" is constituted by visual information about its shape (e.g., lines, two-dimensional planes, three-dimensional structures), colors, and the activation of specific neural clusters triggered by this visual input. Concurrently, clusters of neurons activated by pronouncing or hearing the word "face," or by touching one's own or someone else's face, interrelate with those activated by visual experiences, forming and storing the concept of "face" within the perceiving subject (Barsalou 2003, 63–65) The concepts, categories, or symbols that humans construct, retrieve, and manipulate based on consciousness are organically linked to clusters of neurons and their simultaneous, sequential networks. Intriguingly, Schjødt (2007, 322; 2013, 302–3) notes, the process linking specific concepts to neural clusters within human cognitive behavior is underpinned by the organism's underlying orientation towards homeostasis.

Schjødt explores Damasio's theory of emotion to substantiate his claims. Damasio posits that human cognitive actions are not abstract activities occurring in isolation or independent of the body. Instead, these actions are reactions to stimuli, both internal and external, aimed at fulfilling the universal goal of homeostasis that all living beings share. Damasio (1994, 155–58, 173–85) specifically identifies two loops: the body loop, eliciting immediate, physical, and emotional reactions to stimuli disrupting homeostasis, and the as-if loop, generating non-immediate, conscious, and emotional responses. The body loop, for instance, activates physical emotional responses to immediate threats to homeostasis, such as the fight-or-flight response triggered by encountering a snake. A practical illustration includes the body's physiological response to lower-than-optimal blood sugar levels by accelerating stomach digestion to raise blood sugar back to normal ranges. In contrast, the as-if loop contemplates potential future threats based on past experiences or learning, fostering emotional responses in preparation for these hypothetical scenarios without immediate physical action. For instance, viewing images or videos of snakes or imagining

dreadful scenarios such as being bitten by a venomous snake does not trigger an immediate physical reaction but can evoke similar emotions and feelings of fear at a conscious level. Another example is when blood sugar levels drop below the normal range and there is no food in the stomach, the as-if loop can conceptualize and execute a strategy to address this issue, such as going to the nearest store to purchase food. Schjødt, leveraging Damasio's and Barsalou's discussions, asserts that human-generated concepts, categories, or symbols are deeply intertwined with emotional responses and feelings embedded in the foundational process of striving towards homeostasis. He further elucidates that humans, as beings oriented towards maintaining homeostasis, not only evaluate present dangers but also foresee future risks, crafting cognitive actions grounded in this extensive risk assessment paradigm (Schjødt 2007, 325–27).

Schjødt draws from coping psychology research to stress that human cognitive behavior, while deeply rooted in consciousness, actually originates from unconscious and immediate neurophysiological processes aimed at maintaining homeostasis, thereby highlighting the intrinsic link between these two levels. This concept is vividly illustrated by the well-documented rat experiment by J. Weiss, in which two rats were subjected to identical electric shocks but with one crucial difference: one rat received an auditory cue before the shock, enabling it to anticipate the event, while the other received the cue randomly, thus preventing any anticipation. The study measured stress levels through changes in ulcer size, revealing that rats that could predict the electric shock did not show an increase in ulcer size, in contrast to those unable to predict it, underscoring the interaction between neurophysiological processes and cognitive coping mechanisms (Bloom, Lazerson, and Nelson 2001, 266; Schjødt 2007, 328). This interplay is further evidenced in human research, notably in a study involving partners of AIDS patients, where individuals who found or constructed meaning from their partner's death maintained higher CD4 T-cell counts and showed better mortality trends post-interview compared to those who did not find such meaning (Schwarzer and Knoll 2003, 401; Schjødt 2007, 328). These findings underline the association between cognitive behavior and neurophysiological efforts to achieve homeostasis, suggesting that the human inclination to seek or construct meaning from a series of events and circumstances is also rooted in cognitive behaviors oriented towards maintaining homeostatic balance.

Schjødt posits that religious cognitive behavior plays a key role in maintaining human homeostasis, akin to previously discussed instances. He references a study by Kenneth I. Pargament and Curtis R. Brant that examines the impact of religious understanding on parents coping with the severe stress of recent child loss compared to those who faced less stress from a loss that occurred two years prior. The findings suggest that religious understanding significantly aided stress management in parents dealing with more acute stress (Pargament

and Brant 1998, 125; Schjødt 2007, 329–30). This is further evidenced by other studies cited by Schjødt, which demonstrate that religious understanding not only benefits patients with severe illnesses but also provides psychological stability to individuals facing personal crises through religious practices, like prayer (Levin and Chatters 1998, 36; Pargament and Brant 1998, 123; Schjødt 2007, 330). These pieces of evidence collectively argue that human religious cognitive behavior, rooted in neurophysiological efforts towards homeostasis, plays a vital part in sustaining this stability (Schjødt 2007, 330–37). Schjødt's argument is reinforced by recent research showing the positive impact of religious coping mechanisms on enhancing human wellbeing, thereby suggesting a fundamental link between religious behavior and the neurophysiological pursuit of equilibrium (Counted et al. 2022, 70–81; Dolcos, Hu, Dolcos 2021, 2892–905).

Schjødt's theory illuminates how human religious cognitive behavior likely originates from neurophysiological mechanisms aimed at achieving fundamental homeostasis. Specifically, based on the discussions by Barsalou and Damasio, it can be argued that the human experiences of finitude, induced by disease and death, generate negative feelings through the body loop and the as-if loop, which in turn disrupt homeostasis. These negative feelings, indicating the danger of internal and external stimuli in the context of background feeling and homeostasis, likely drive the formation of concepts and symbols, along with their meanings, that can alleviate these feelings at a conscious level. Thus, it might be inferred that concepts such as the idea of an immortal soul persisting after death have emerged for this very reason. Such concepts and symbols are generally intertwined with religious concepts and symbols. Within this framework, it is plausible to hypothesize that human religious cognitive behavior originates from fundamental neurophysiological mechanisms directed towards achieving homeostasis. This understanding might be indirectly supported by findings in coping psychology, which suggest that religious cognition and behavior contribute to the maintenance and preservation of homeostatic processes in humans.

This insight aligns with two prominent theories on the origin of religion. First, it supports the Darwinian account, which posits that religious behavior evolved as a survival mechanism, either at the individual or group level, through natural selection. This perspective suggests that while modern religious practices may not be direct outcomes of evolutionary processes, they likely stem from ancestral religious behaviors that conferred survival advantages. Such behaviors could include enhancing sexual selection, establishing dominance, fostering group cooperation, and/or providing psychological relief from existential threats. These ancestral practices, beneficial for survival, may have been refined and adapted over generations, culminating in the complex religious phenomena observed today (Schloss 2009, 21–22).

The cognitive account offers a second perspective, arguing that religious cognitive behavior is a byproduct of sophisticated cognitive abilities in humans, which provided survival advantages throughout evolution. This theory suggests that religious cognition and behavior could emerge from various psychological mechanisms, such as the instinct to avoid contagion or danger, the development of fundamental human relationships and attachments, the tendency to project human qualities onto natural phenomena, the use of theory of mind to infer the presence of agents, and experiences related to religion, dreams, or imagination (Schloss 2009, 17–20). While the Darwinian and cognitive accounts propose different mechanisms at specific levels for the origin of human religious behavior, they converge on a broader understanding that such behavior likely developed as part of an overarching orientation towards self-preservation or the maintenance of homeostasis. The main distinction between them lies in the directness of the impact of the neurophysiological process of homeostasis on cognitive functions—with the Darwinian account suggesting a direct influence and the cognitive account suggesting an indirect one. This implies that the fundamental process of homeostasis, crucial for all living organisms, could affect not just higher-order cognitive functions but religious cognitive behaviors as well.

The physicalist understanding of human religious cognitive acts suggests that religious cognitive behaviors could also emerge from AI robots endowed with vulnerable bodies driven by self-preservation. This notion is bolstered by William Sims Bainbridge's simulation studies. Bainbridge (1995, 483–95; 2006, 1–16, 117–37) conducts simulation experiments on the origins of religious behavior based on the theory of religion he jointly developed with Rodney Stark. The experiments rest on several premises: humans seek rewards and avoid losses, they join social groups and engage in reciprocal actions to achieve this goal, they perceive and pursue information needed for physical rewards as a form of reward itself, the human mind operates as an information-processing system implemented in a physicalist manner, and humans have a fundamental orientation towards self-preservation (Bainbridge 1995, 483–86; 2006, 2–11).

Building on these premises, Bainbridge simulates a society of 44,000 artificial agents capable of reinforcement learning and equipped with neural networks and memory registers. These agents are divided into four groups, each participating in activities to produce consumable rewards, thereby acquiring the four essential rewards for survival (i.e., energy, water, food, oxygen), either through their productive activities or via exchange with other agents. Beyond these four rewards, the agents also seek life itself. However, upon realizing that life cannot be obtained through production or exchange activities, they hypothesize the existence of an exchange partner besides the existing exchange partners and treat information about this hypothetical partner as a reward,

continually gathering such information. Bainbridge points out that the behavior of artificial agents, assuming a supernatural exchange partner not present among the existing ones, mirrors human religious behavior based on the belief in supernatural entities (e.g., gods) that cannot be accessible through sensory experiences in the general sense (Bainbridge, 2006, 117–37).

Bainbridge's experiment does not prove or disprove the existence of supernatural realities or the truth claims of religions but supports the notion that human religious cognitive behavior could emerge from natural processes driven by a fundamental mechanism of self-preservation. This understanding can coexist with theological perspectives on the origin of the concept of God. The recognition of human existential finitude is organically linked to an implicit understanding of the infinite. In other words, awareness of the infinite is already a precondition for the recognition of finitude because experiencing something in the world and becoming conscious of it necessitates a background or horizon against which this directedness occurs. In Husserl's terms, the horizon, and in Schleiermacher's terms, the universe, serves as the basis for the epistemic subject's activity and the existential subject's being (Smith 2013, 286–94; Schleiermacher [1799] 1996, 18–24). Therefore, the activities and existence of finite entities implicitly evoke the infinite, which underlies cognition and existence. Pannenberg (1992, 1:73–82, 1:136–51) calls this "innate knowledge of God" (cognitio Dei innata), a concept also found in the works of Friedrich Schleiermacher, G. W. F. Hegel, and Rudolf Otto.

Pannenberg posits that the innate knowledge of God is universally endowed to all humans. This is because, as previously noted, all existential beings adopt the eccentric form of life, meaning that all beings are dependent for their existence and not self-contained. However, the innate knowledge of God remains nonthematic. In other words, as finite beings, humans possess only implicit knowledge of the infinite, despite their cognition and existence being fundamentally grounded in the infinite. In this context, the awareness of finitude, while not necessarily inducing specific religious beliefs or practices, can at least evoke a sense of one's existence and cognition being dependent on an infinite foundation (Pannenberg 1992, 1:113, 1:115–18). Jürgen Moltmann (2001, 93) similarly points out that human self-awareness of finitude leads to the recognition of the infinite as an existential origin.

Consequently, it is plausible to anticipate that once machines, equipped with vulnerable bodies striving for self-preservation or homeostasis, recognize their own existential finitude, they may conceive concepts such as the transcendent or the infinite in an effort to preserve or extend their finite existence. By manipulating these concepts, they could, in theory, develop a form of religious symbolic system. This is because, as noted in Pannenberg's argument, any being capable of human-level consciousness possesses the innate knowledge of

God. This implicit knowledge of the divine can be transformed into explicit understanding through engagement with the established religious traditions within human society. In other words, through social interactions with humans, there is the potential for these machines to make the nonthematic and implicit knowledge of God thematic and explicit, thereby assimilating into existing religious frameworks and participating in religious cognitive actions within those settings. Similarly, Andrew Proudfoot tentatively forecasts the feasibility of religious behaviors in AI by postulating the capability of AI for self-consciousness. Should AI discern its existential dependencies and engage in linguistic communication grounded in human symbolic systems, it might very well manifest religious behaviors (Proudfoot 2023, 685–90).

Conclusion

This article has explored the potential for religious behavior to emerge from AI robots by understanding intelligence from the perspective of embodied cognition. Specifically, it examined the view that, from a physicalist and embodied cognitive standpoint, intelligence can be understood as the processing of information about internal and external states in the pursuit of the goal of homeostasis. It is clear that current levels of AI, unlike natural intelligence, cannot set goals for themselves based on emotions and feelings and then carry out cognitive actions. Nonetheless, if, following the suggestions of Man and Damasio, AI and soft robotics are integrated, and these machines are endowed with homeostasis as a fundamental goal, they might be able to set subgoals for themselves and exhibit emotional responses and subjective feelings as characteristics for realizing these goals. Furthermore, it might be possible to distinguish between internal and external stimuli as rewards and losses within the ultimate context of self-preservation, and from there, the possibility arises for attaining a level of self-consciousness akin to human-level consciousness based on an internal assessment of the machine's bodily state, similar to an organism's "background feelings." Information processing and cognitive actions based on self-consciousness could evolve into high-level semantic actions. Grounded in the recognition of their finitude and the desire for self-preservation, the possibility of machines implementing religious behavior also arises.

If the theoretical exploration presented in this article proves accurate, and AI robots surpass human levels of intelligence, encompassing self-consciousness, it would significantly bolster the physicalist understanding of human consciousness. Moreover, it would reinforce a naturalistic perspective in the study of the origins of religious behavior. Consequently, theological understandings of revelation would need to be reinterpreted within a naturalistic and/or physicalist context, and the conception of humans as made in the image of God would also require revision. Additionally, this article's

discussions could contribute practically by suggesting that, based on societal consensus, the implementation of self-preservation directives and vulnerable physicalities in AI robots should be avoided in the development of AGI and other AI technologies. This is primarily because most international demands on AI research and development (e.g., OECD AI Principles, European Union AI Act) define the purpose of AI robots strictly in instrumental terms for human use. Implementing AI robots in the manner discussed in this article could lead to their creation not merely as tools but as autonomous entities, complicating international and social relationships and potentially elevating the rights of machines to a level equal to human rights. Therefore, the issue of AI consciousness and personhood should not be dismissed as merely speculative. The conditions under which these could realistically occur should be thoroughly explored and regulations established to prevent their technical implementation, understanding this article's discussions within such a context.

References

Altman, Sam. "Planning for AGI and Beyond." OpenAI. February 24, 2023.

Bainbridge, William Sims. 1995. "Neural Network Models of Religious Belief." *Sociological Perspectives* 38 (4): 483–95.

———. 2006. *God From the Machine: Artificial Intelligence Models of Religious Cognition*. Lanham, MD: AltaMira Press.

Barsalou, Lawrence W. 1999. "Perceptual Symbol Systems." *Behavioral and Brain Sciences* 22:577–660.

Barsalou, Lawrence W., Paula M. Niedenthal, Aron K. Barbey, and Jennifer A. Ruppert. 2003. "Social Embodiment." In *The Psychology of Learning and Motivation*, edited by Brian H. Ross, 43–92. San Diego: Academic Press.

Bloom, Floyd E., Arlyne Lazerson, and Charles Nelson. 2001. *Brain, Mind and Behaviour*. New York: Worth Publishers.

Bostrom, Nick. 2023. "What If A.I. Sentience Is a Question of Degree?" Interview by Lauren Jackson. *New York Times*, April 12, 2023. https://www.nytimes.com/2023/04/12/world/artificial-intelligence-nick-bostrom.html.

Chalmers, David J. "The Puzzle of Conscious Experience." *Scientific American* 273 (6): 80–86.

Counted, Victor, et al. 2022. "Hope and Well-Being in Vulnerable Contexts during the COVID-19 Pandemic: Does Religious Coping Matter?" *The Journal of Positive Psychology* 17 (1): 70–81.

Damasio, Antonio. 1994. *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Avon Books.

———. 1996. "The Somatic Marker Hypothesis and the Possible Functions of the Prefrontal Cortex." *Philosophical Transactions: Biological Sciences* 351 (1346): 1413–20.

———. 2012. *Self Comes to Mind: Constructing the Conscious Brain*. London: Vintage.

Damasio, Antonio, and Gil B. Carvalho. 2013. "The Nature of Feelings: Evolutionary and Biological Origins." *Nature Reviews Neuroscience* 14:143–52.

Dennett, Daniel C. 1991. *Consciousness Explained*. New York: Back Bay Books.

Dolcos, Florin, Yifan Hu, and Sanda Dolcos. 2021. "Religiosity and Resilience: Cognitive Reappraisal and Coping Self-Efficacy Mediate the Link between Religious Coping and Well-Being." *Journal of Religion and Health* 60:2892–905.

Dorobantu, Marius. 2022. "Artificial Intelligence as a Testing Ground for Key Theological Questions." *Zygon: Journal of Religion and Science* 57 (4): 984–99.

Doya, Kenji, and Eiji Uchibe. 2005. "The Cyber Rodent Project: Exploration of Adaptive Mechanisms for Self-Preservation and Self-Reproduction." *Adaptive Behavior* 13 (2): 149–60.

Foerst, Anne. 1998. "Cog, A Humanoid Robot, and the Question of the Image of God." *Zygon: Journal of Religion and Science* 33 (1): 91–111.

Galhardo, Rodrigo S., P. J. Hastings, and Susan M. Rosenberg. 2007. "Mutation as a Stress Response and the Regulation of Evolvability." *Critical Reviews in Biochemistry and Molecular Biology* 41 (5): 399–435.

Heaven, Will Douglas. 2023. "Google Deepmind Wants to Define What Counts as Artificial General Intelligence." MIT Technology Review. November 16, 2023. https://www.technologyreview.com/2023/11/16/1083498/google-deepmind-what-is-artificial-general-intelligence-agi/.

Herzfeld, Noreen. 2023. *The Artifice of Intelligence: Divine and Human Relationship in a Robotic Age*. Minneapolis, MN: Fortress Press.

Husserl, Edmund. (1931) 1960. *Cartesian Meditations: An Introduction to Phenomenology*. Translated by Dorion Cairns. Boston: Martinus Nijhoff Publishers.

Levin, Jeffrey S., and Linda M. Chatters. 1998. "Research on Religion and Mental Health: An Overview of Empirical Findings and Theoretical Issues." In *Handbook of Religion and Mental Health*, edited by H. G. Koenig. San Diego: Academic Press.

Levinas, Emmanuel. (1930) 1995. *The Theory of Intuition in Husserl's Phenomenology*. Translated by Andre Orianne. Evanston, IL: Northwestern University Press.

Man, Kingson, and Antonio Damasio. 2019. "Homeostasis and Soft Robotics in the Design of Feeling Machine." *Nature Machine Intelligence* 1:446–52.

Maturana, Humberto R., and Francisco J. Varela. 1987. *The Tree of Knowledge: The Biological Roots of Human Understanding*. Boston: Shambhala Publications.

Merleau-Ponty, Maurice. (1945) 2012. *Phenomenology of Perception*. Translated by Donald A. Landes. New York: Routledge.

Moltmann, Jürgen. 2001. *The Spirit of Life: A Universal Affirmation*. Minneapolis, MN: Fortress Press.

Pannenberg, Wolfhart. 1992. *Systematic Theology*. Vols. 1 and 2. Translated by Geoffrey W. Bromiley. New York: T&T Clark.

Pargament, Kenneth I., and Curtis R. Brant. 1998. "Religion and Coping." In *Handbook of Religion and Mental Health*, edited by H. G. Koenig, 112–26. San Diego: Academic Press.

Park, Choong Shik. 2018. "Artificial Intelligence as Life." In *Ontology of Artificial Intelligence*, edited by Jung Won Lee, 21–40. Seoul: Hanulmplus.

Proudfoot, Andrew. 2023. "Could a Conscious Machine Deliver a Pastoral Care?" *Studies in Christian Ethics* 36 (3): 675–93.

Puddefoot, John C. 1996. *God and the Mind Machine: Computers, Artificial Intelligence and the Human Soul*. London: SPCK.

Reiss, Michael J. 2023. "Is It Possible That Robots Will Not One Day Become Persons?" *Zygon: Journal of Religion and Science* 58 (4): 1062–75.

Schjødt, Uffe. 2007. "Homeostasis and Religious Behaviour." *Journal of Cognition and Culture* 7 (3–4): 313–40.

———. 2013. "A Resource Model of Religious Cognition: Motivations as a Primary Determinant for the Complexity of Supernatural Agency Representations." In *Origins of Religion, Cognition and Culture*, edited by Armin W. Geertz, 301–9. New York: Routledge.

Schleiermacher, Friedrich. (1799) 1996. *On Religion: Speeches to Its Cultured Despisers*. Translated by Richard Crouter. New York: Cambridge University Press.

Schloss, Jeffrey. 2009. "Introduction: Evolutionary Theories of Religion." In *The Believing Primate: Scientific, Philosophical, and Theological Reflections on the Origin of Religion*, edited by Jeffrey Schloss and Michael J. Murray, 1–25. New York: Oxford University Press.

Schwarzer, Ralf, and Nina Knoll. 2003. "Positive Coping: Mastering Demands and Searching for Meaning." In *Positive Psychological Assessment: A Handbook of Models and Measures*, edited by S. J. Lopez and C. R. Snyder, 393–409. Washington, DC: American Psychological Association.

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

Smith, David Woodruff. 2013. *Husserl*. New York: Routledge.

Steels, Luc. 1994. "The Artificial Life Roots of Artificial Intelligence." *Artificial Life* 1(1–2): 75–110.

van Hateren, J. H. 2013. "A New Criterion for Demarcating Life from Non-Life." *Origins of Life and Evolution of the Biosphere* 43 (6): 491–500.

Varela, Francisco J. 1999. *Ethical Know-How: Action, Wisdom, and Cognition*. Stanford, CA: Stanford University Press.