Journal of Philosophical Theological Research
Print: 1735-9791
Online: 2538-2500

Consciousness: Human, Artificial, or Both?

Viktor J. Tóth

Doctor of Philosophy (PhD), Fuller Theological Seminary, Pasadena, CA, USA

vtoth@fuller.edu

Abstract

The nature of the "intelligence" of artificial intelligence is a hotly debated notion today. Sensational claims about the successes of AI (including the declaration that synchronicity is already achieved) and warnings against its potential malice (e.g., that once it attains the level of human consciousness it will turn against humanity) are part of our everyday experience. In this article, I argue that the most extreme claims about the present state of AI are not warranted. In the first part of the article, I make the case that human intelligence/consciousness is a unique phenomenon that cannot be matched by machines. Then, in the second part, I turn to what I consider the more valid possibilities and worries about AI. I envision artificial personas (quasi-agents) working in collaboration with human agents on the betterment of the world, and counsel against the possibility of weaponizing AI. I finish the article by reflecting on some of the constructive ways for philosophical engagement in this discussion.

Original Research

Keywords:

human consciousness, artificial intelligence, quasi-agent, embodied cognition, artificial moral agents

Received: 2026/03/24; **Received in revised form:** 2026/05/16; **Accepted:** 2026/06/10; **Published online:** 2026/06/15

Tóth. V. J. . (2026). Consciousness: Human, Artificial, or Both? *Journal of Philosophical Theological Research*, (Special Issue on AI, Consciousness, and Human Agency) 28(3), 155-176. https://doi.org/10.22091/jptr.2026.17198.3657

Copyright © The Author

Introduction

"A cyborg, a robot, and a hybrid walk into a bar . . ." It sounds like the opening line of a joke, but it more or less describes the overall theme of the 2025 FX television series *Alien: Earth* (Hawley, 2025). The story purportedly takes place in 2120. As the first episode starts, a short script appears on a computer screen:

> In the future, the race for immortality will come in three guises:
> Cybernetically enhanced humans: Cyborgs
> Artificially intelligent beings: Synths
> and
> Synthetic beings downloaded with human consciousness: Hybrids
> Which technology prevails will determine what corporation rules the universe. (Hawley, 2025)

For the purpose of this article, I define these three concepts as follows: a cyborg is a human person who was born and developed naturally but receives technological enhancements that provide physical and/or mental abilities exceeding normal human capacities (Option-1[1]); a synth is a robot with a fully synthetic body and artificial intelligence (Option-2); and a hybrid is an entity exhibiting human consciousness (having been born naturally and raised in an ordinary human environment) but whose natural body dies for some reason (e.g., terminal illness), after which their consciousness is downloaded into a synthetic body (Option-3).[2] This paper aims to calrify the difference among these three options from a philosophical point of view and give an explanation of why only Option-1 is a viable prospect at this point of AI development.

Although all the above examples are taken from the world of science

1. Therefore, people who receive medical treatment as a remedy to treat disabilities (e.g., a lost limb) are not cyborgs because they do not gain superior abilities; enhancement aims to restore their natural abelites. Medical science already advanced to the level where neural interfaces can surgically placed into the cortical tissues, meaning that the person's brain is operating an artificial limb in the same way it would move organic limbs (see e.g., Ferris Jabr, "Brain Wave," *The New York Times Magazine*, "The Health Issue," May 15, 2022, 26–31, and Simon Makin, "Brain-Controlled Prosthetic Limbs Take Steps Closer to Reality," *Science News* 206, no. 9 (2024): 22–25. In popular culture both "RoboCop" and the Terminator (a.k.a. Cyberdyne Systems Model 101) are called *cyborgs* (Option-1), but in my definition only RoboCop classifies; the difference being that RoboCop was born as a human being (Alex Murphy) having human consciousness (ultimately regaining his memories), but Model 101 has artificial intelligence operating a synthetic body (even though "covered by" living tissue), thus qualifies as a *synth* (Option-2).

2. Probably the best-known example of this scenario in popular culture is Major Motoko Kusanagi from the *Ghost in the Shell* franchise. Transhumanism embraces this concept as a viable option to enhance human lifespan beyond the biological limitations of a human body.

fiction, they are generally treated as real possibilities in contemporary society. In other words, although they have not yet been realized by science, humanity feels confident that they are not only achievable but will be attained soon enough. Besides the technological complexities, this enterprise raises non-engineering-related questions arising from the fields of ethics, philosophy, theology, psychology, and economics (just to mention a few). In what follows, I will reflect on some of the philosophical aspects of the notion. I will approach the issue focusing on whether Artificial General Intelligence (hereafter, AGI)[1] is a viable option based on the capabilities current AI systems exhibit. The claim that once AGI is reached it will be inevitably followed by technological singularity makes this enterprise significant.[2]

The purpose of this article is to examine these possibilities by using the above-mentioned options as hypothetical alternatives and, in this manner, assess their achievability. First, in Sections 1 and 2, I will consider the cyborg, synth, and hybrid alternatives, using cultural paleoanthropology as a benchmark in the study of human consciousness. In Section 3, I will elaborate on the possibility of AI as an enhancement to human life. I will finish the article by pointing out some of the (in my opinion) most viable avenues for future deliberation.

The Depth and Width of Human Consciousness

In this section, I focus on Option-2 (synths) and Option-3 (hybrids). Although O-2 and O-3 are not identical options, they both operate under the assumption that "human consciousness" (I will use the abbreviation "HC" from now on) can be represented digitally as coded information flow (e.g., as an algorithm) and ultimately "run" as a program.[3] One of the auxiliary premises of this assumption is the widely accepted notion that information is something extra

1. While regular (or narrow) Artificial Intelligence (AI) is a system designed to excel in a particular task (e.g., playing chess or GO, translate between languages, or run a hospital in the most energy efficient way), AGI not only able to transfer skills form one task (or filed) to another, but also solve novel problems via leaning from the history of its interactions and thus reach the cognitive level of humans.

2. The hypothetical point in technology when a human built system transcends human comprehension.

3. I must note here that the nomenclature of the different fields of research dealing with the nature of human consciousness is perplexing; words like *consciousness*, *mind* or *soul*, *intellect*, *intelligence*, *psyche*, *brain*, or *thinking* (as a noun) are used freely and interchangeably which confuses the discussion enormously. In this work I will use the designation "human consciousness" (HC) as constantly as I can.

(or beyond) physical existence – something immaterial. This postulation then feeds into the concept "that conscious awareness can be reduced to the structural residue of information or computational process in the brain" (Frank et al., 2025, p. 15) and ultimately simulated in very sophisticated machines.

I find it somewhat startling that although the questions of what consciousness is and how consciousness works are mired in controversy (I even venture to say that the more it is studied and debated, the murkier the concept becomes), it is almost universally accepted that it is technologically reproducible. The belief that AGI is around the corner (or, according to some, already achieved) is indeed widely accepted and leads to the alarming idea that AI can somehow acquire self-consciousness and determination, and even turn against humans. This technological positivism entirely ignores the so-called hard problem of consciousness (the term was introduced by David Chalmers in 1994, but the inquiry goes back at least as far as the Presocratic philosophers in the West). The hard question concernshow organic material (i.e., animal bodies) can give rise to the phenomenon termed HC (in the idiom artificial intelligence, "artificial" basically stands for the physical part, or "hardware," and "intelligence" for HC as a hardware-based entity). Thus, in order to even entertain O-2 and O-3 as viable options, we need to take a closer look at HC.

While there is only a limited consensus about the nature of consciousness, there are some notions that most scholars agree on. One of the widely accepted opinions is that HC is the outcome of a long evolutionary process. This assumption is mainly based on the principle that if the human body is the outcome of a lengthy evolutionary process, then higher human capacities must be the outcome of the same developmental means.

For a long time, the evolutionary study of HC was instigated in a somewhat obscure way, focusing almost entirely on brain development and technological inventions, leaving other segments of human evolution neglected.[1] This – in my opinion, outdated – approach is based on a cognitivist individualistic view of human nature that endorses the notion of abstract contextualized cognition as the hallmark of rationality. As Cat Bohannon puts it so aptly: "We're terribly impressed with our brain. I'd argue we're in love with it. Which is to say, the human brain is in love with itself. If there's a single, physical trait that most scientists agree delineates humans from the other apes, it's our huge, lumpy, terribly intelligent brain" (Bohannon, 2023, p. 245). Indeed, this

1. For a thorough critique of the "cognitive approach" see Haidle, 2012. Haidle writes: "close correlation between brain size and intelligence still has to be rejected as misleading" (Haidle, 2012, p. 56). Terrence W. Deacon also criticizes "our preoccupation with brain size (Deacon, 1997, Ch. 6).

cognitivist approach implies that reason is the defining characteristic of human nature and rationality is the most important activity for human agency.[1] As an outcome of this one-sided outlook, emotions were treated as a "lower" function, a remnant of our animal ancestry. Thankfully, this attitude toward human intelligence is rapidly changing.

One of the pioneers of creating a different account of HC, one that takes emotions seriously, is Antonio Damasio. In countless publications and so many books, the Portuguese neuroscientist makes the claim that feelings and emotions are as important for the study of HC as our cognitive capacities. His groundbreaking opus, *Descartes' Error*, explains how the cognitive agenda hijacked the narrative about human consciousness from the time of Plato, and then was cemented into modern thinking via Descartes (Damasio, 1995).

The so-called "social brain hypothesis" is one of the alternative attempts to describe HC from a paleoanthropological perspective that puts emotional cognition at the center of hominin evolution. Its advocates claim that what makes humans qualitatively different from other animals is our capacity for complex and diverse social life. In short, human civilization was able to progress from mere sociality (which is found in many other species) to sociability. What distinguishes human sociality is the relative depth, the layers of affiliations and intentions, that we are able to navigate on a day-to-day basis. From this perspective, "intelligence" is not about brain size or the individual's capacity to create and use sophisticated tools, but about the ability to form multilayered, orderly societies. Humans can learn and successfully employ certain social skills that significantly enhance the sociocultural dimensions of the group. Although the "creative spark" (to utilize Agustín Fuentes's watchword in Fuentes, 2017) is indispensable, it is not the brilliant individual who maintains the possibility of human progress, but the stable, multilayered, and transgenerational social structure. In what follows, I will make some deliberations on the notions of transgenerational and multilayered societies in an attempt to relate them to AGI.

Transgenerational memory, the ability to maintain information flow between successive generations, is a must for every organism's survival. It can be described as the memory of the collective. The quality and quantity of the information stored and distributed within social systems is one of the most important marks of the intelligence of a given species. In this sense, the notion includes the ability to create decision-making processes via amplifying the residues of past experiences, and, in a similar vein, to cancel out unconstructive biases across time. However, as the above sentence indicates,

1. For a more detailed argument on this issue, see Lakoff and Johnson, 1999, pp. 3 & 408.

this "memory" is not a kind of stored knowledge in an individual's head (as passive information), but active datasets that routinely re-enter the system from a "higher" (cultural) level. During the history of human evolution, this feedback system amplified humanity's overall freedom of behavior through constant re-evaluation. In essence, this process defines the learning process of the group. This process is utterly different from the operational function of data centers, the facilities used in the maintenance and development of AGI.

Yet there is something even more special about the way living systems learn and remember. In the words of Terrence Deacon, the learning strategy of organisms is "defined by a fundamental incompleteness" (Deacon, 2006, p. 143, emphasis in original). They seem to somehow perceive what is missing or incomplete for their most desirable future (i.e., a target state). Deacon coined the term *teleodynamic* pertaining to this apparent "end-directed tendency" (Deacon, 2012, p. 271); the word comes from the Greek *telos*, which means "end" or "goal." One of the main characteristics of teleodynamic processes is the strong presence of the notion of a self that can benefit or suffer. The notion of self thus underlines the representation, functional organization, and normative characteristics of the organism. Nevertheless, this process requires several selves interacting with each other in a coordinated way. This is why the memory of an organism cannot be depicted as (passive) information, but only as a powerful medium on an always operational (group) level. This continuously active operational dynamic creates non-linear feedback loops that then amplify the initial conditions and generate progress. The dynamic, self-transcendent nature of "memory systems" is a sine qua non of HC – a component that is seemingly missing from AGI development.

The second concept I want to touch upon here in view of the evolution of HC is the notion of a layered society. This concept is closely related to the notion of individual selves (as the locus of intentionality) but extends its scope. The social brain hypothesis distinguishes six possible layers of intentionality. First-order intentionality is the awareness of my own feelings and thoughts. The second layer is my basic awareness that others around me have similar states of mind to mine. It creates the basic ability for the self to "step back" and pose the question of how and why things are as they are. On the third level, we become aware of our own states of mind, which can coexist with, but are different from, the states of mind of other individuals.[1] The fourth layer operates with the supposition that others suppose that we should coordinate

1. It is debated whether this level can exist without full-blown human language not necessarily extended beyond the level of noun-verb statements. Probably the level of Large Language Models, because it uses language without the capability the concept of what language *actually* is.

our actions in relation to an "abstract" rule (e.g., for the benefit of the group).[1] This level is achievable only via abstract representation, which presupposes language with full recursive grammar and phrase structures. A genuinely moral human community emerges only with fifth-order intentionality. On this level, one is able to handle situations in which "I believe that you suppose that I imagine that you want me to believe that" (Dunbar, 2014a, p. 58). It seems that such an abstraction requires special mental devices, such as myths and rituals.[2] Humans use allegorical and metaphorical thinking to operate on this level. What enables HC to designate "general" knowledge – knowledge that is acquired in one area to be used in other areas – is metaphorical thinking.

There is no space here to fully entertain the notion of powerful memory systems such as myths and rituals.[3] But there is no question about the significant role they played (and still play) in creating, sustaining, and changing not only cultural milieus but whole cultures. Rituals and myths are closely connected and were, without a doubt, developed in tandem. Johan Huizinga expressed it eloquently: "In myth and ritual the great instinctive forces of civilized life have their origin: law and order, commerce and profit, craft and art, poetry, wisdom, and science" (Huizinga, 1938/1950, p. 95). On the one hand, myths served (and do so even today) as collective external memory systems (Donald, 1991, p. 382), creating and maintaining the intellectual environment in which explicit consciousness develops (Peterson, 1999, p. 80). On the other hand, ritual "represents the creation of a controlled environment where the variables (i.e., accidents) of ordinary life may be displaced precisely because they are felt to be so overwhelmingly present and powerful" (Bellah, 2011, p. 135). In contemporary societies, shared social rituals and myths function as preservatives of social equilibrium amidst a seemingly corroding

1. This level is claimed to be reached by OpenAI when in July 2026 an artificial intelligent system supposedly broken free of human control. This accident is widely reported by news agencies, but I yet to find a reliable independent scholarly review of the incident. Nevertheless, note that this level is still below HC.

2. N. B.: The general population never operates on the sixth level (only people with special skills do, such as William Shakespeare. Robin Dunbar illustrates this level by explaining how Shakespeare's *Othello* works: "the audience has to *suppose* [level 5] that Iago *intends* [level 4] that Othello *believes* [level 3] that Desdemona *loves* [level 2] Cassio and that Cassio also *loves* [level 1] her (see Dunbar, "Why Only Humans Have Language," p. 440 – emphasis original). Shakespeare operates on the sixth level because he must intend that the audience believes that this is the narrative about Othello.

3. There is a supplementary notion here about the "volume" and "speed" of information processing. AI is much faster than humans when it comes to process *raw* data. However, myth in itself is a memory device that helps to process cultural information collected over countless generations. In this sense, humans are better processing this special kind of (cultural) data. My point is that to use phrases like "information" and "processing" can be misleading when it comes to understanding how *human* intelligence works.

global civilization. As I see it, AGI is inevitably a participant in this cultural milieu since it is based on the cultural setting of its programmers and trainers; however, AGI participates in such notions only as raw information. I see this as a significant difference between HC and AGI.

And the difference does not stop here. HC goes well beyond AGI by the unique way it frames memory. The human way of accumulating memory is not comprised solely of what and how we remember. It is of equal significance what we do *not* remember. Our success is equally based on our ability to ignore a large amount of information – information that we, by some means, determine to be irrelevant. We do it without running entire sets of information against another set of information. We do it so naturally, without even noticing it (for details on this issue, see Frank et al., 2025, pp. 166–172). We prioritize information based on relevance as a context-dependent framing system that "cannot be reduced to heuristic rules and symbolic representations" (Frank et al., 2025, p. 168).

So-called "second-wave" programming uses large amounts of data and attempts to analyze them by finding large-scale patterns and statistical correlations. Once technology achieves this level, AGI will (allegedly) be able to ride the wave of context dependence; however, in the process, it reproduces statistical classifications embedded in the culture of its trainers. I want to be cautious here; I recognize my limited knowledge of computer science. Nonetheless, to move forward with my argument, I propose that HC may display similar characteristics, being based on a special kind of statistical operation. Yet, to highlight what I perceive as the difference between the statistical/probability approach used by second-wave AI computation and HC, I employ here Karl Popper's notion of "propensity":

> Propensities, it is assumed, are not mere possibilities but physical realities. They are as real as forces, or fields of forces. And vice versa: forces are propensities. They are propensities for setting bodies in motion. Forces are propensities to accelerate, and fields of forces are propensities distributed over some region of space and perhaps changing continuously over the region (like distances from some given origin). Fields of forces are fields of propensities. They are real, they exist. (Popper, 1990, p. 20)

Instead of utilizing statistical (probability) calculations, HC operates within propensity fields. To rephrase, while statistical possibilities as mathematical values are successfully incorporated into AI systems, HC operates within propensities – "fields of forces" representing real causal powers. The propensity field is the environment in which the system is embedded, and it has a real impact on the outcome of the system's actions. Consequently, the

system's perspectives are based on values essential to the overall environment (as a metasystem) for its development, human culture being the most important metasystem. I venture to say that AI is twice removed from this level because the non-"value-free" dataset is provided as "valued" information through human intervention. In other words, companies that create AI (and aim to create AGI) must work around this deficiency by constantly providing the necessary "values" (via human agency) to make sure that the operation of the system does not come to a halt.[1] I will come back to these features in the next section.

All in all, I agree with Damasio that it is possible for robots (O-2) to develop consciousness, but their consciousness is probably going to be different from HC (as cited in French, 2026). However, I have no intention of ending on a dismissive note. AI is real and, just like every other powerful invention, can cause both harm and benefit in our lives. In the next section, I will consider an alternative approach to the notion.

Layers and Players: Persona as Quasi-Agent

In a recently published article, Zahra Khazaei proposes an idea that I find remarkably close to my concept of differentiating between the human self and the human persona. Khazaei proposes the concept of a "quasi-agent" (Khazaei, 2026). She defines this concept as "an entity that exhibits stable, adaptive, and often unpredictable patterns of agent-like behavior without possessing ontological subjectivity, intrinsic intentionality, or unified experiential presence" (Khazaei, 2026, p. 9). In what follows, I will link this notion to the idea of a layered society and social games.

HC is not a free-flowing information matrix separated from material existence. It is both embodied (it is revealed by and through human bodies) and embedded (it can be developed and effectively employed only within a particular culture). Most AI narratives seem to apply a mechanistic view of human existence, in which the body (or the brain) is the hardware through which consciousness (or intelligence) operates as the software. But the human body is not a machine. Properly speaking, it is not just a physical entity made up of body parts (e.g., limbs, head, torso, stomach, lungs, brain, etc.) or a composite of different small building blocks (e.g., carbon, hydrogen, calcium, etc.). It is better explained as a system of processes (e.g., circulatory system,

1. Another way to label these necessary "values" is "worldview." As long as humans are involved worldview systems are unavoidable.

digestive system, nervous system, reproductive system, etc.) manifesting as a whole within its environment (Murphy, 2018, pp. 246–1247). These subsystems are better defined by what they do than by their physical properties; they are structural constituents that might overlap with each other and might even serve different purposes over time.

I propose that human identity stems from two intertwining components based on the embedded-embodied condition: the self and the persona. On the one hand, the persona is a conglomerate of bits and pieces in our personal sociocultural toolsets, activated by the demands of playing our roles in different social settings. On the other hand, the self (or "minimal self") is the "constant background feeling of being alive … or the simple self-referential aspect of self-awareness," which "also can be imagined as a hidden, fluctuating affective relief structure; that is, the dimension of combined affective valences, emotions, dispositions, [and] habits" (Horváth, 2024, p. xvii). Thus, a self may be represented by several personas. These personas are rooted in and, ideally, managed by a unified reflective self-identity.

In the following, I use examples from the human ontogenetic process in an attempt to argue why AGI can operate as a persona (i.e., a quasi-agent) or as a placeholder for an individual entity, but not as a "full-fledged" self. Human beings do not enter life with fully developed intelligence; as newborns, we even lag behind other mammalian species. As the late German Lutheran theologian Wolfhart Pannenberg puts it, "In comparison to other higher animals human beings are born a year too soon and still in an unfinished state" (Pannenberg, 1985, p. 38).[2] On the other hand, we enter our sociocultural environment to inhabit readymade personas (i.e., roles to play in life).[3] Here, I am referring to the notion that, because most brain growth occurs after birth, the community functions as a "kind of social uterus" (Haidle & Schlaudt, 2020, p. 167).[4]

---

1. We might also describe subsystems from a biochemical point of view in which we might study the endogenous opioid system, or the neurobiological system, or the skeletomuscular system, or fascia system, as subsystems working in unison for maintaining homeostasis within the body.
2. For a wider perspective on Pannenberg's notion, see Kärkkäinen, 2015, pp. 252–55. Both Pannenberg and Kärkkäinen highlight the fact of how this specially evolved physiological trait makes "enculturation" (a psychological phenomenon) an indissoluble part of human nature.
3. For example, as newborns their *persona* allows them to cry whenever they feel like it, to expose their feces, or even to wake up at the middle of the night and require feeding when their caregiver is fast asleep. The point I am making here is that this *persona* is only a temporal entity; the time must come when it is left behind and the child must go through a metamorphosis to learn to operate different personas, ones which conforms their future circumstances.
4. Kärkkäinen uses the phrase "social womb," attributing it to Adolf Portmann, the great Swiss

There are obvious differences between "training" a human child toward intelligent behavior and "training" an AI model. For one, as the Finnish American theologian Veli-Matti Kärkkäinen points out, "Human development is facilitated rather than hindered by the deficiency of instincts, unlike other animals [and, I insert, AI]; hence, in human life a 'hiatus, a gap, between perceptions and impulses' is productive" (Kärkkäinen, 2015, p. 255) – and the opposite might be said about AI. I find only one similarity in how AI language programs and human beings are trained: both require human oversight.[1] I venture to say that this difference is even more imperative when it comes to AI development. To put it briefly, HC is an embodied and embedded process that emerged through hundreds of millions of years, and each human individual (i.e., self) is trained in their cultural environment based on their particular cultural inheritance. The training of AI, by contrast, is based on datasets provided by humans, mostly focusing on removing "noise" (partially through tokenization[2]), followed by the involvement of human "trainers" in the fine-tuning process (first for preference and then for reasoning, i.e., advanced logic). What I see here is humans (operating via human intelligence) trying to jumpstart AI by providing countless hours of labor by countless people and using a huge amount of energy (both in terms of human manpower and on the machine side).[3] And the product? Something that produces outcomes that, in some special cases, match or even exceed those of humans. Yet, it works only as an enhancement of human intelligence. Therefore, I argue that although AI can be trained by humans to occupy a quasi-agent status in our social matrix, given the current status of development, ontologically speaking, it will never become a genuine self.[4]

---

zoologist (Kärkkäinen, 2015, p. 255).

1. This assumption is based on my Google search on "how deep language programs are trained"; I used an AI generated overview, so it must be accurate!
2. This abstraction is based on the mostly outdated concepts of the structure of language by Noam Chomsky and his followers.
3. One of the worries about LLMs is that they use an enormous amount of energy, because they rely on large industrial data centers. As an example, imagine the GO competition between Lee Se-dol (South Korean former professional Go player of 9 dan) and AlphaGo (a computer program developed by Google DeepMind that mastered Go). While master Lee probably had a regular breakfast before the game (and consumes a normal number of calories per day during his lifetime, which is between 1,600–2,400 calories [= 0.00186kWh–0.002789kWh]), AlphaGo used approximately 170 kW–440 kW = 46,271,510–378,585,086 cal. during the game. It means that the magnitude of energy used by the AI is 170,000 bigger! Surly, this disparity between HC and LLMs alone must rase questions about the potential future of AI systems.
4. The *self* versus *persona* is my nomenclature for Khazaei's distinction between (human) agent and quasi-agent. My concept is explained in detail elsewhere (see Viktor J. Tóth, "The Evolution of the Human Soul: Multidisciplinary Enquiry into the Emergence of the 'Soul'" [PhD diss., Fuller Theological Seminary, 2026], "Chapters 3 and 4;" available at https://doi.org/10.13140/RG.2.2.29498.04802).

I will continue this line of reasoning in an attempt to make my argument clearer. HC, from a sociocultural perspective, is a problem-solving skill of everyday life. It is the accumulation of the kinds of social games we perform as healthy members of society. Most of the time, we do this automatically (i.e., many of our actions do not even rise to conscious awareness). More and more scholars from different fields of anthropology and brain sciences realize what some philosophers (e.g., Ludwig Wittgenstein and Maurice Merleau-Ponty, just to name the two most important) grasped almost a century ago: thinking and decision-making are not processes that take place solely in our brains. The brain, or even the whole body, is only a necessary condition for HC. In other words, HC is not a "thing" but a process – the way humans participate in a shared world.

James Wm. McClendon Jr. and James M. Smith also identify the "social matrix" as an embodied relationship (McClendon & Smith, 1975/1994, pp. 162–174). However, their explanation adds another layer to the concept. They note that a child learns language by learning language-games, in which the context and content of the words used in a given social situation are based on the situation itself (and not on dictionary forms). It is a "way of talk embodying assumptions and ways of reasoning that distinguishes the various forms" (McClendon & Smith, 1975/1994, p. 22). The conventions on which these language-games are based are not even explained to us, and they do not have to be, because "they are beneath the level of explanation, yet quietly determinative of how the language is understood, how the language-game in question is played" (McClendon & Smith, 1975/1994, p. 23). Such practices are validated by, and dependent on, the community. Thus, when Large Language Models (LLMs) are trained by humans, they are linked to this social matrix by their trainers, but without the basic concepts on which the actual language-games are formed. In other words, no matter how confidently an LLM can operate on this level, it is able to do so only because of human intervention and will probably remain in need of human supervision throughout its operation.

I think the best thing about this debate is that it helps us learn more about humanity. Marina Zilbergerts words this notion quite nicely:

> For what AI reveals with striking clarity is not the full nature of human mind but the narrowness of our definition of intelligence. Intelligence, when reduced to calculation, optimization, and symbolic manipulation, may indeed be reproducible in machines. Consciousness, however, is not. While artificial intelligence excels at replicating certain mechanical aspects of reasoning, it falters when confronted with other core capacities of the human mind, such as recognizing boundaries, responding with discernment to novel situations, or expressing (let alone experiencing) the depth and

nuance of human emotion and other qualia. (Zilbergerts, 2026, p. 135)

As a clarifying note, I point out that Zilbergerts uses *intelligence* and *consciousness* differently from my terminology. "Intelligence" for her, being reducible to data processing and optimization, is closer to my depiction of persona. In her jargon, "consciousness" is what stands for the unique way of human existence (I use both *consciousness* and *self* to highlight different aspects of this phenomenon).

In human society, being successful means being able to detect and project the right persona in each situation. Since I argued above that persona is equivalent to the quasi-agent in Khazaei's reasoning, I propose that a well-trained AI can occupy this social space. However, cultures demand that humans acquire several well-developed personas or index-identities[1], which also suggests that, within the framework adopted here, an AI persona will never be able to reach the level of HC (that is, AGI in digital networks). Being one specialized persona trained in one set of etiquette[2] as an accessory for human beings (i.e., Option-1) is plausible; however, reaching the obligatory level of changing etiquette "on the fly" based on a particular social setting (choosing from dozens of possibilities) seems to be unachievable at this point of AI development.

Natural-Born Cyborgs

The notion of enhanced humans in this article refers to people with physical, cognitive, or sensory abilities amplified beyond normal human limits using technology or genetics. The British philosopher Andy Clark made the claim decades ago that *Homo sapiens* is a "natural-born cyborg" (Clark, 2003). Clark's remark, of course, was highlighting the fact that humans evolved not just to adapt to their environment, but also to incorporate elements of their environments into their everyday lives (clothing, tools, and even animals through domestication). Not only that, but we make significant changes to our environment to make elements of it useful to us. We build and maintain well-organized societies far beyond those of every other living organism; we intuitively utilize tools and each other through seamless connections. In a way,

1. I refer to Cole's terminology here because I find it usefully for my argument (for more information on this concept see Cole, 2015).
2. I think "etiquette" here is a better choice than "protocol" since the former refers to a smaller (human) scale of operational setting (e.g., a particular social setting), while the later has an artificial (unnatural) "taste" to it.

HC can be defined by this very special way of living. HC is not a thing; it is a way of life. To use an analogy, it is not the wing that makes a bird able to fly; it is how a bird uses this special feature, adapted to the rest of its body and evolved through millions of years.

The point I make here is that our consciousness (HC) is not something that separates us from nature,[1] nor is it something that separates us from each other. Consciousness links us. Another way of expressing consciousness is living a lifestyle of constant integration. It is the integration of developmental processes within an evolutionary "framework that engages with cultural systems and processes including institutions and recognizes the fluid and entangled interfaces between individuals, groups, and community-level dynamics" (Fuentes, 2016, p. 14). We cooperate with each other and with our environment on a meta-level. In other words, "human nature is actually a meta-nature, represented by the possibility of dynamically assembling a vast gamut of extended cognitive architectures" (Wheeler & Clark, 2008, p. 3563). For us, being conscious means that we are interconnected, which means that all of our actions and capabilities, although experienced in a situated manner, are carried out in interconnection with each other. We coordinate, bond, and accomplish this so proficiently and so naturally that we do not even notice it most of the time. Following this line of thought, I see AI as an organic outcome of the evolution of HC effortlessly integrated into human consciousness.

One of the main tenets of my argument is that we must consider HC as something denoted to humankind and not as an opportunity of the individual. In other words, there is no such thing as individual consciousness; it can only be recognized and studied as a shared experience (i.e., an isolated human body would never experience what a regular human being in her social setting develops into and experiences as consciousness).[2] Thus, I see it as futile to investigate the possibility of downloading a person's consciousness into a machine. In other words, it is impossible to isolate "one" consciousness.[3] However, it is not only feasible to imagine AI becoming part of the global (or

1. The sad fact that humanity exploits nature created a false dichotomy between nature and culture. Yet, harming nature is a suicidal act, and hopefully humanity will realize this mistake soon enough.
2. As a devastative but highly informative source see "Romanian Orphanage Case" (Mackesa et al., 2020).
3. Here, I make a note that one crucial aspect of HC is not even touched upon in this article: the *un*conscious side of human intelligence. Subconscious activity was part of the investigation about human intelligence and agency from the very beginning of human thinking, but since Sigmund Freud and Carl Gustav Jung it was firmly established the into the scientific fields of psychology and psychiatry. No enquiry about HC can be adequate without weighting in the complex world of the unconscious.

local) consciousness matrix, but it is indeed happening as we speak.

Here, again, I weigh in on the notion of a layered society. Such a social construction is built on rules and regulations. Without going into details, I note here that we live our everyday lives following written and unwritten rules.[1] They are cultural artifacts that constrain our lives. Our success is very much based on how well we can adapt to them and utilize them. The current state of the AI debate is mostly about how such regulations can be implemented so that AI will continually be an asset to HC and not an adversary. I very much doubt that simply imposing laws (e.g., the "Three Laws of Robotics") will suffice here.[2] I think it is much more constructive to equip AI systems with "habitual dispositions."[3]

Mark Graves proposes that one potential way to implement the notion of habitual dispositions in AI is to train AI systems toward self-development. Such a method would assist the system in acquiring "habits" as general interpretive decision-making initiatives. Rehearsing natural encounters of the system with human culture and analyzing the possible outcomes (still with human supervision) would eventually build good habits into the system's operational repertoire and move the system toward building a stable (morally good) character. The hope is that eventually the system reaches a point where it can mediate among possible good choices in a way that is close enough to human decision-making. I propose that once this level is reached, the system achieves what Khazaei identifies as a quasi-agent, possessing all the potentials of a persona (without attaining the ontological level of self). Graves identifies "practical wisdom" as a byword for developing habits and character traits from a virtue ethics perspective for the purpose of enabling the system to consistently make morally informed, skillful decisions aligned with our ethical norms (Graves, 2026). Still, I maintain that no matter how useful the habits or practical wisdom such a quasi-agent (artificial persona) might develop, AI capabilities or apparent intelligence do not, by themselves, establish either a genuine self or genuine consciousness.

I propose that the optimal goal is not to replace human intelligence and

1. These rules are codified in civil law systems, inherited as cultural habits, encountered as institutional regulations, etc.
2. The "three Laws of Robotics" are created by the science-fiction writer Isac Asimov and introduced in his short story titled "Ruanround" (Miami, FL Analog Science Fiction and Fact, 1942). The three laws are: (1) a robot may never implement any harm on a human being; (2) a robot must always obey human orders, except if it conflict with the "First Law"; and (3) a robot must protect its own existence except if doing so would conflict with the first two laws.
3. What follows is informed by Graves, 2026, Graves, 2021, and Graves, 2025.

behavior, but to opt for kinds of intelligence that may or may not resemble human performance while being ethically responsible. In other words, the aim is not to copy the execution of an ethically satisfying action, but the outcome of it (i.e., the betterment of human existence). That is to say, AI should not copy human standards but find the best way to augment (support) human endeavors (which is not necessarily through imitation, but, at least sometimes, through the implementation of new possibilities). This objective might benefit from an approach that, instead of focusing on "rationalistic-cognitive" predispositions, treats human intelligence more as a role-playing endeavor.[1] Thus, rather than trying to create human-like agents (agents that we cannot detect as non-human), research should focus on developing quasi-agents (agents that we know are not human but are designed to help humans). Instead of trying to reproduce our intelligence as a narcissistic endeavor, why not simply focus on discovering new ways that can aid human flourishing?

I agree with Graves that achieving perfect rationality is not feasible in complicated environments where computational demands exceed any system's resources (Graves, 2025). However, it is absolutely viable to use AI as a quasi-agent to assist or substitute for humans when computational demand is high. I assess that what I reflected on above as the metaphorical and allegorical dimensions of human comprehension coincides with Graves's "motivational dimensions"; if my assessment is correct, then I also agree with him that this is the threshold that is outside the grasp of (contemporary) AI research. Nevertheless, AI can be trained, for example, to perceive human distress signals (e.g., those of a child or a dying patient in a hospital) and initiate soothing behavior. Such behavior would be indistinguishable from genuine empathy, but still qualitatively different from its human counterpart.[2]

Echoing Graves's analyses, I propose that although current self-monitoring techniques might be able to identify interpretable data within neural network representations and correlate them by identifying activation-space directions compared to moral attributes, these techniques do not yet constitute the reflexive (self-)identity required (Graves, 2021). The most significant restraint in self-monitoring AI systems is that they achieve only partial reflective self-monitoring (and never a full sense of self built upon a narrative self-perception). Graves also highlights that such constitutive identity does not emerge in isolation. The only examples we can work with are those provided by human culture and communal life, and it is very unlikely that such a

1. I do not advocate here for using "game theory" (a highly rationalistic concept) which already plays a crucial role in AI development.

2. In the same wain the fire rescue robot's detecting of higher temperature than its body can endure qualitatively defers from a human firefighter's perception of unendurable heat.

phenomenon can be reproduced in any other kind of setting.

In summary, there is currently no sufficient basis to conclude that AI is as close to gaining HC as some (mostly entrepreneurs from AI development companies) might want us to believe. Yet, there is no question about the seriousness of the enterprise when it comes to the future of human culture. AI can be a powerful instrument both as an aid to human betterment and as a destructive menace. My conclusion, therefore, is that instead of creating false anxiety about AI (or AGI) running rogue against humanity, we should attain a better understanding of how it can be used (and constrained) to aid human advancement. Here, on the one hand, we can promote "roboethics" ("robot ethics") while, on the other hand, endorsing machine ethics to develop AIs with "built-in" morality and thus create artificial moral agents (AMAs) (Boyles, 2024). AMAs are endowed with built-in, top-down ethical systems to elicit moral behavior. The aim here is not to supply AIs with preset ethical principles, but to let them develop moral behavior on their own. It is assumed – and from an evolutionary approach rightly so – that moral behavior would naturally emerge via system-environment interactions with human society (Boyles, 2024, p. 17). From an embodied and embedded perspective, I see it as a framework for a human-endowed simulation of evolutionary processes in a virtual environment.

I suppose that since AI is developed by and within human agencies, which are social environments already furnished with powerful means to deal with the aggregation of power and malicious elements, it would be "difficult for AI designers not to impose their personal convictions in addressing these issues; social choice ethics only mitigates the said impositions" (Boyles, 2024, p. 19). From this approach, it is obvious that for AMAs to function in ways that are not proper according to human ethical standards, they would have to be designed that way. This means that it is human criminal behavior that would make AI systems act against proper behavior. Thus, the "vicious cycle" of crime and punishment circles back to us. It is not an "evil AI" we need to be wary of, but "the evil inclination of the human heart" (Genesis 8:21).

Retrospect and Prospect

AI could be dangerous, like many other state-of-the-art technologies. It threatens the "average Joe" in many ways: it might eliminate jobs,[1] help terrorists create

1. See https://abcnews.com/Business/artificial-intelligence-eliminate-thousands-jobs-new-york-city/story?id=133181521. Although others who claim otherwise (see e.g., Robert

bioweapons,[1] or even drain the beauty out of mathematics.[2] In this article, I argued that one threat is false: AI, at least in its current developmental stage, is not going to gain HC. Nevertheless, AI needs to be taken seriously, and that is why I propose the following from a philosophical approach:

1. As we develop a philosophy of AI, we need to be careful to distinguish it from the philosophy of mind by taking good care of its nomenclature (I consider Khazaei's "The Challenge of Agency" a good example of this; see Khazaei, 2026). Consciousness is a notoriously elusive concept; however, in regard to AI systems, we can use the notion of "quasi-agent" and/or "artificial persona" (AP) as a background for clarifying what "human consciousness" (i.e., "full agency") means. In other words, we need to develop a new conceptual framework for this task.
2. The obvious legal responsibilities are being considered as we speak. I want to add a note of caution here, though: I would not trust big tech companies to spearhead this task by self-regulation. Yet, this is exactly what is happening, and it warrants serious concern.
3. As I argued above, AI is not a transcendent (i.e., non-material) reality. AI is hardware-dependent. And it has already reached the point where the amount of material resources (especially rare-earth elements), human labor, and energy consumption (the electricity cost of current large-scale deep-learning systems is enormous) probably does more harm to the environment (and thus to human lives) than good. In a recently published book, the authors warn: "AI is fundamentally an 'extractive industry,' one that 'depends on exploiting energy and mineral resources from the planet, cheap labor, and data at scale'" (Frank et al., 2025, p. 15).
4. Mental health issues: A recent study on human-computer interaction suggests that "users are no longer viewing AI as a mere tool but are developing meaningful emotional bonds with conversational systems" (Donard & Ribeiro, 2026). I want to add that interaction with some of these entities (i.e., LLMs) becomes part of one's everyday life history, such that any interruption in this relationship might cause distress. At this point, it seems that the information feedback loop between the system

---

Bartholomew, "AI Hysteria: Is the Sky Really Falling?", *Psychology Today* [Posted August 23, 2026], available at *https://www.psychologytoday.com/us/blog/its-catching/202608/ai-hysteria-is-the-sky-really-falling*. Bartholomew argues that it was always the case when new technologies surfaced people feared that they will take away jobs.

1. See https://screendna.org (June 2026).
2. See https://zenodo.org/records/20302944 (June 2026).

and the user has already become an emergent property (i.e., a meaningful event) in the life of the human agent. Looking back, in reminiscence of the history of other "addictive" substances (here I take the liberty of claiming that human-computer interaction has already reached the point at which the exchange itself has become a "substance" in this sense), we seem to be already too late in this matter.

5. I agree with Damasio that it is possible for robots (Option-2) to develop consciousness based on their bodies, but their consciousness is going to be different from HC (as cited in French, 2026). Therefore, I think it is very important to clarify what we mean by "artificial intelligence." Are we talking about reproducing HC, or something else entirely?

▣ Conflict of Interest

The author declares no competing interests.

References

Bellah, R. N. (2011). *Religion in human evolution: From the Paleolithic to the Axial Age*. Harvard University Press.

Bohannon, C. (2023). *Eve: How the female body drove 200 million years of human evolution*. Alfred A. Knopf.

Boyles, R. J. M. (2024). Can't bottom-up artificial moral agents make moral judgements? *Filosofija Sociologija*, 35(1), 14–22. https://doi.org/10.6001/filsoc.2024.35.1.3

Clark, A. (2003). *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*. Oxford University Press.

Cole, J. (2015). Hominin language development: A new method of archaeological assessment. *Biosemiotics*, 8(1), 67–90. https://doi.org/10.1007/s12304-014-9198-8

Damasio, A. (1995). *Descartes' error: Emotion, reason and the human brain*. HarperCollins.

Deacon, T. W. (1997). *The symbolic species: The co-evolution of language and the brain*. W. W. Norton & Company.

Deacon, T. W. (2006). Emergence: The hole at the wheel's hub. In P. Clayton & P. Davies (Eds.), *The re-emergence of emergence: The emergentist hypothesis from science to religion* (pp. 111–150). Oxford University Press.

Deacon, T. W. (2012). *Incomplete nature: How mind emerged from matter*. New York: W. W. Norton & Company.

Donald, M. (1991). *Origins of the modern mind: Three stages in the evolution of culture and cognition*. Harvard University Press.

Donard, V., & Ribeiro, J. C. (2026, April). Technological mourning after AI updates: Mental health and well-being in the GPT-4o/GPT-5 transition. *Academia Mental Health & Well-Being*. https://doi.org/10.20935/MHealthWellB8256

Dunbar, R. I. M. (2014a). 'Mind the gap' or why humans aren't just apes. In R. I. M. Dunbar, C. Gamble, & J. A. J. Gowlett (Eds.), *Lucy to language: The benchmark papers* (pp. 47–60). Oxford University Press.

Dunbar, R. I. M. (2014b). Why only humans have language. In R. I. M. Dunbar, C. Gamble, & J. A. J. Gowlett (Eds.), *Lucy to language: The benchmark papers* (pp. 61–74). Oxford University Press.

Dunbar, R. I. M., Gamble, C., & Gowlett, J. A. J. (Eds.). (2014). *Lucy to language: The benchmark papers*. Oxford University Press.

Frank, A., Gleiser, M., & Thompson, E. (2025). *The blind spot: Why science cannot ignore human experience*. MIT Press.

French, K. (2026, September). Does consciousness require a body? *Nautilus*. https://nautil.us/does-consciousness-require-a-body-1284886/

Fuentes, A. (2016). The extended evolutionary synthesis, ethnography, and the human niche: Toward an integrated anthropology. *Current Anthropology*, 57(S13), S13–S26. https://doi.org/10.1086/685684

Fuentes, A. (2017). *The creative spark: How imagination made humans exceptional*. Dutton.

Graves, M. (2021). Emergent models for moral AI spirituality. *International Journal of Interactive Multimedia and Artificial Intelligence*, 7(1), 7–15. https://doi.org/10.9781/ijimai.2021.08.002

Graves, M. (2025, October 10). *Designing an AI virtuous agent* [Paper presentation]. Pacific Coast Theological Society meeting, San Francisco, CA, United States.

Graves, M. (2026). AI practical wisdom and compassion. *AI and Ethics*, 6(1), Article 39. https://doi.org/10.1007/s43681-025-00877-4

Haidle, M. N. (2012). *How to think tools? A comparison of cognitive aspects in tool behavior of animals and during human evolution* (Vol. 1, Cognitive Perspectives in Tool Behaviour). Tübingen University Press. https://tobias-lib.uni-tuebingen.de/xmlui/handle/10900/49627

Haidle, M. N., & Schlaudt, O. (2020). Where does cumulative culture begin? A plea for a sociologically informed perspective. *Biological Theory*, 15(3), 161–174. https://doi.org/10.1007/s13752-020-00351-w

Hawley, N. (Writer & Director). (2025, August 12). Neverland (Season 1, Episode 1) [TV series episode]. In N. Hawley (Executive Producer), *Alien: Earth*. FX Productions; Hulu.

Horváth, L. (2024). *The affective core self: The role of the unconscious and retroactivity in self-construction*. Springer Nature.

Huizinga, J. (1950). *Homo ludens: A study of the play elements in culture*. Beacon Press. (Original work published 1938)

Kärkkäinen, V.-M. (2015). *Creation and humanity* (Vol. 3, A Constructive Christian Theology for the Pluralistic World). William B. Eerdmans.

Khazaei, Z. (2026). The challenge of agency in the age of artificial intelligence: An ontological perspective. *Journal of Philosophical Theological Research*, 28(1), 7–14. https://doi.org/10.22091/jptr.2026.16299.3583

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. Basic Books.

Mackes, N. K., Golm, D., Sarkar, S., Kumsta, R., Rutter, M., Fairchild, G., Mehta, M. A., & Sonuga-Barke, E. J. S. (2020). Early childhood deprivation is associated with alterations in adult brain structure despite subsequent environmental enrichment. *Proceedings of the National Academy of Sciences*, 117(1), 641–649. https://doi.org/10.1073/pnas.191126411

McClendon, J. W., Jr., & Smith, J. M. (1994). *Convictions: Defusing religious relativism* (Rev. ed.). Trinity Press International. (Original work published 1975)

Murphy, N. (2018). *A philosophy of the Christian religion: For the twenty-first century*.

SPCK.

Peterson, J. B. (1999). *Maps of meaning: The architecture of belief*. Routledge.

Popper, K. R. (1990). *A world of propensities*. Thoemmes.

Wheeler, M., & Clark, A. (2008). Culture, embodiment, and genes: Unraveling the triple helix. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1509), 3563–3575. https://doi.org/10.1098/rstb.2008.0135

Zilbergerts, M. (2026). Algorithms, zombies, and Maimonides' God: A Jewish theology of consciousness for the age of AI. *Tradition: A Journal of Orthodox Jewish Thought*, 58(2),122-149. https://doi.org/10.54469/R9XT5FM2