Scientia*et*Fides 14(2)/2026 ISSN 2300-7648 (print) / ISSN 2353-5636 (online)

Received: June 25, 2025. Accepted: Novembre 20, 2025 DOI: http://dx.doi.org/10.12775/SetF.2026.017

STANISŁAW SZELĄG
University of Warsaw
szelag.st@gmail.com
ORCID: 0009-0007-6127-3909

The Quest for Genuine Artificial Intelligence: What Are We Looking For?

Abstract. The rise of ChatGPT and various other publicly available AI chatbots has brought the discussion regarding AI to the forefront of academic and public discourse. Yet when it comes to the most important question: what would make AI genuinely intelligent, or human-like, it seems that philosophers by and large are satisfied with the answers given by in the past century by Searle and Turing, among others. The essay surveys some philosophical visions of what (artificial) intelligence and consciousness are, prominent in the past century. This work contends that rapid advancements in AI demand that those in the field of philosophy of mind take a stance with all the seriousness that goes together with expecting real answers. The paper attempts to provide new, original answers to the question of what would make AI 'genuinely intelligent' or 'human-like', and analyse which features of the human mind are necessary to conclude that their owner is "genuinely intelligent". Consequently, the nature and importance of an open phenomenology are highlighted, and the paper advocates for the acceptance of a form of realism about the world, coupled with a Putnam-style externalism. The combination of these frameworks aims to provide one with a list of conditions that allow one to verify the emergence of a genuine artificial mind. The proposed criteria also clarify when claims about AI's moral status – and theological concerns about human distinctiveness (e.g., imago Dei) – are philosophically warranted.

Keywords: AI, philosophy of mind, phenomenology, realism, imago Dei.

Contribution. This paper offers a novel set of necessary conditions for "genuine" AI that moves beyond the Turing–Searle stalemate: (i) world-embedment that secures an appropriate causal connection for reference (Putnam-style externalism), and (ii) self-referential feedback whereby a system tracks how its outputs reshape future inputs. Framed in a connectionist picture of internal representation, these conditions yield a more testable and ethically/theologically relevant criterion for when AI could count as an agent rather than a mere symbol manipulator. This, in turn, has novel and profound implications for AI ethics and the theological idea of imago Dei.

Use of AI. AI tools and technologies not used in this work.

1. Defining the question

Given the recent (whenever I refer to the 'recent' surge in interest, developments, or to the "current" discussion, debate etc., I mean the aftermath of the launch of ChatGPT, so November 2022) surge of public interest in the topic of Artificial Intelligence (AI), sparked by the release of ChatGPT in 2022 and other Large Language Models (LLMs) soon after, it seems only natural that those already in the field of AI research should study the rumoured revolution extensively.

Indeed, the academic interest in the topic of AI seems considerable; even if the exact extent to which the recent developments influence the discussion remains to be seen. Likewise, to conduct an extensive survey of the current state of the discussion might be premature. One can nonetheless gauge the productivity of the debate not by examining the answers given, but by specifying what kinds of questions need to underpin it, and verifying whether these are being posed or not.

In the case of the current discussion about AI, two root questions are:

(1) Is the AI system in question genuinely intelligent?

and

(2) What makes, or would make such a system genuinely intelligent?

Without answering at least (1), the relevance of asking questions about AI ethics, mechanisation, and the existential risk AI could pose to humanity, just to name a few, is itself put in question.

This is also a religion-and-ethics problem, not only a technical or metaphysical one. If "genuine intelligence" is a precondition for moral agency or moral considerability, then clarifying what it would take for an artificial system to qualify bears directly on questions of responsibility, dignity, and – to the extent that theology is concerned – especially debates that invoke the imago Dei.

To address these issues, one needs to determine, depending on one's approach to philosophical questions about consciousness, either

(3) what the mental structure or phenomenology of AI is, or

(4) how AI works, what its mechanism does; a mechanistic and functional analysis.

These differing approaches align with Daniel Dennett's "levels of description". The particularly useful outcome of Dennett's work on the topic of consciousness is that one can, in circumstances such as those of the present essay, abstract from the differences between questions (3) and (4), and consider them, for specific intents and purposes, interchangeably.

For more, see (Dennett 1978), and (Dennett 2007).

This essay uses the word 'phenomenology' in a strict sense of an analysis of the structure of experience. In that it follows Shaun Gallagher and Dan Zahavi in denying that phenomenology is merely an introspective enterprise; see (Gallagher, Zahavi 2013:21–23).

The necessity to ask (3) and (4) if one decides to answer (1) and (2) is clear. Having an understanding of how an AI system works, allows one to superimpose it on the essential features of the human mind and obtain a clear result as to how close the AI system researched is to being 'human-like'

Naturally, the debate over the extent to which an artificial system can be intelligent will have profound implications for the theological idea of imago Dei. Although more than one formulation of the idea exists, the one rooted in human intelligence and together with it the human distinctiveness from the rest of the world, is one of the central ones. Although there is not enough space in this paper to do full justice to the theological implications of its conclusions, some will be pointed to in the conclusion.

2. The Turing-Searle Spectrum

To give a satisfactory answer about what would make AI genuinely intelligent, one has to first survey the prevalent attitudes and intuitions. The spectrum of attitudes towards AI equips one with the necessary terminological and conceptual tools to tackle the problem. Key in the discussion is what can be called the Turing-Searle paradigm, which represents varying intuitions about AI with the Turing Test on one end of the spectrum and Searle's Chinese Room thought experiment on the other, with other attitudes falling in between. While philosophers of mind and AI researchers have since introduced various models to conceptualise intelligence, the paradigm's goal is not to replace or reduce their insights; I argue that Turing's and Searle's stances serve as a good illustration of the two extreme attitudes that persist in the discussion about AI.

Upon describing the 'imitation game', Turing writes:

> We may now consider again the point raised at the end of §3. It was suggested tentatively that the question, "Can machines think?" should be replaced by "Are there imaginable digital computers which would do well in the imitation game?" If we wish we can make this superficially more general and ask "Are there discrete state machines which would do well?" But in view of the universality property we see that either of these questions is equivalent to this, "Let us fix our attention on one particular digital computer *C*. Is it true that by modifying this computer to have an adequate storage, suitably increasing its speed of action, and providing it with an appropriate programme, *C* can be made to play satisfactorily the part of *A* in the imitation game, the part of *B* being taken by a man? (Turing 1950, 442)

The question posed above became subsequently generalised into what is now popularly known as the Turing Test, where the facts concerning the machine's 'thinking' depend on how convincingly it can act like a human agent. To say that the Turing Test 'failed' on all fronts by being too liberal in granting intelligence (which is the most popular criticism of the Turing Test) is to misread Turing's original paper. Passing the Test was never supposed to define *what it means to think*. Skepticism toward AI being

(possibly) genuinely intelligent arises not from the Test's failure but from its perceived uselessness.

Given the level of conversational proficiency of the currently available chatbots and the fact that this proficiency has to do with statistical predictions based on training data (as opposed to human-like understanding) (Gupta et al. 2023)(Roumeliotis et al. 2023), the natural tendency is to reject the Turing Test as in any way meaningful. This is not to argue that the proficiency of currently-available chatbots is in any way 'satisfactory'; no criteria to be satisfied have been given. My argument is that the assertion of the genuineness of a chatbots' intelligence should not be based on its ability to use language. This is because there is a (in Turing's terms) 'machine' that would most likely do well in the imitation game, as envisioned by its author, and yet we know for a fact that its operating structure is nothing like what one might expect from a genuinely 'thinking' actor.

Should the Turing Test, thus understood, be dismissed as uninformative, along with any claims that systems like ChatGPT possess 'thinking' abilities? I think yes. The intuition of rejecting the putative 'thinking' status of ChatGPT and other LLMs is not misplaced.

Intuitively it does not only matter *what* a person, or for that matter an AI chatbot, is saying, but *why* they are saying it, although perhaps more fittingly, *how* they are saying it. Moreover, criteria for genuine intelligence that could be satisfied by mere statistical prediction would rightly be seen as not stringent enough

The correctness of the popular intuition notwithstanding, it points to the opposite end of the spectrum; the claim that AI can never become genuinely intelligent or human-like (the strong formulation) or that what is needed for AI to be genuinely intelligent or human-like is some kind of an intentional 'secret sauce', the recipe for which is not to be found in AI, understood as a machine-based property.

Searle's point is to reject the functionalist claim whereby a computer program can genuinely think. Searle does not deny, however, that a 'machine' (a system in which a program operates) could exhibit intentionality; he does, however, seem to be convinced that such

a machine would have to be a biological one; '(…) strong AI has had little to tell us about thinking, since it has nothing to tell us about machines. By its own definition, it is about programs, and programs are not machines. Whatever else intentionality is, it is a biological phenomenon, and it is as likely to be as causally dependent on the specific biochemistry of its origins as lactation, photosynthesis, or any other biological phenomena.' (Searle 1980:424)

Such is, in its most influential aspects, the stance of John Searle in his Chinese Room thought experiment. What follows is a summary description of Searle's thought experiment (Searle 1980):

> The heart of the argument is Searle imagining himself following a symbol--processing program written in English (which is what Turing called "a paper machine"). The English speaker (Searle) sitting in the room follows English instructions for manipulating Chinese symbols, whereas a computer "follows" (in some sense) a program written in a computing language. The human produces the appearance of understanding Chinese by following the symbol manipulating instructions, but does not thereby come to understand Chinese. Since a computer just does what the human does – manipulate symbols on the basis of their syntax alone – no computer, merely by following a program, comes to genuinely understand Chinese. (Cole 2023).

This intuition is correct; one-dimensional symbol manipulation alone does not constitute genuine understanding. Where Searle and other proponents of the Chinese Room argument err is in taking their skepticism too far, failing to offer a realistic explanation of how AI could achieve genuine intelligence.

The force of Searle's argument is grounded mostly in the intuition that a speaker in the Chinese Room does not understand Chinese. However, it is also this intuition that affects the assessment of the ability of neural networks to gain genuine understanding; it seems that it rules out symbol manipulation as potentially constitutive of understanding, perhaps in more complex forms, from the outset. And that is not a move without consequences, some of them unwanted. More importantly, however, it leads to a circular argument, as I am going to show.

The more contemporary analog of the Chinese Room thought experiment is the 'black box picture' (Blazek et al. 2021) which, in turn, informs our attitudes towards AI systems and their ability to genuinely understand data. This is not to dispute the usefulness or appropriateness of the black box picture which is supposed to illustrate the multiplicity and inaccessibility of the hidden layers present in neural networks. Rather it too becomes a tempting incentive to abandon the search for answers and occupy one of three positions; an agnostic one, where 'we just cannot know', a skeptical one, illustrated by the anti-Turing and Chinese-Room spirit of critique of the feasibility of genuine understanding in AI, or an overly-enthusiastic one where the currently-available LLMs already are intelligent.

All three attitudes, together with the root issue, the philosophers' misuse of the black box picture, are misguided. Exactly why that is will become evident once we consider the methods with which a functional, or phenomenological, analysis of what is happening inside the black box can be given. Before that, however, it is worth mentioning that the misguided understanding of the black box picture may also be due to the mistaken likening of a one-dimensional symbol-processing unit (the Chinese Room) to a neural network (the black box). Such a formulation would be a mistake, as a Chinese Room would be more akin to (a part of) an individual layer of a neural network than to a neural network in its entirety.

3. Understanding how AI thinks

Consider **Figure 1** and **Figure 2**. **Figure 1** is an illustration of a folk-psychological way in which one may think about an AI system predicated upon fulfilling a specific goal bestowed upon it. For the sake of the example, we can interpret a in this diagram as the function we want the system to fulfill, like recognising if there is an apple in a picture. Once such a system is trained, we think of a as its disposition to identifying apples in pictures if there indeed are any. b, then, is the 'hidden layer'

of all the mechanisms the system has developed to provide the kind of output that was reinforced when it was given feedback.

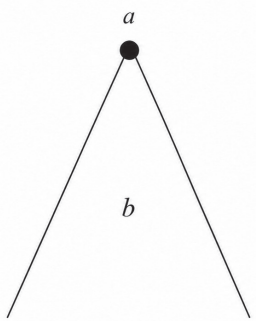


Figure 1

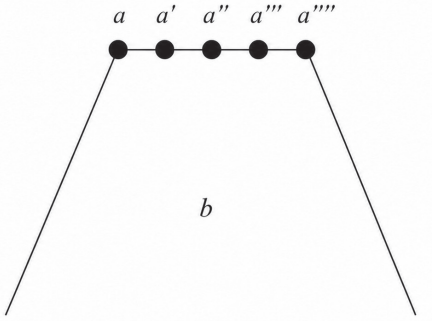


Figure 2

But this is not an appropriate diagram. It assumes that the system is a subject, who in some way has grasped its goal (which is to *a*), and how exactly that has happened (*b*) is obscured in the black box. In other words, **Figure 1** suggests a system that is both intentional (in that its perception has a property of aboutness to it) and teleological (being oriented toward achieving a specific goal). This tendency to overattribute understanding is precisely what Searle and proponents of the Chinese Room argument criticize in overly enthusiastic supporters of Turing's view. And justifiably so. This is not how AI systems, neural networks, or any other pattern-recognition mechanism work.

An apple-recognition system's disposition is distinct from our description of it. Unless we are satisfied with freely (and lavishly) ascribing intentionality to machines, such a system has no relationship with the *purpose* of perceiving apples, and, in terms of the Chinese Room argument, no actual comprehension of what an apple is. As **Figure 2** shows, for the apple-recognising system to recognise an apple is not to have a 'motivation' to *a*, which then gets 'translated into' the given purpose like: *to identify apples in pictures if there are any*, but to *a*, to *a′*, to *a″*, to *a‴*, and to *a⁗*, which do not 'translate' into anything in the folk-psychological sense.

The appropriateness of this model is supported empirically by Templeton et al. (2024) who were able to identify features that activate when an input is given to an AI system, like the mentions of the Golden Gate Bridge, in English and other languages, as well as its being featured in pictures. Subsequently, the researchers were able to measure the activation strength at each token (text and images alike) and create a map of the 'mind' of the AI system, which illustrates which tokens it recognises form what clusters of information, and in what relation to one another.

The above analysis also takes a lot from connectionism, which does have some empirical support when it comes to how human minds may work and even more so when neural networks are considered (Ramsey et al. 1990) (Templeton et al. 2024).

But the point is not to merely elevate connectionism. The fact of the complex (not in the sense of a deepened one, but rather one that is distributed across data-points) nature of a system's understanding (not yet referring to 'genuine' understanding) of what it is created to do is indisputable. This has profound implications for its outputs.

For a long time, AI struggled with generating realistic depictions of human hands (Matthias 2023) (Chayka 2023). This seems to prove the non-intentionalisty of these systems. As UCL's Peter Bentley says: A.I. tools 'have learned that hands have elements such as fingers, nails, palms. But they have no understanding of what a hand really is.' (Chayka 2023). This is because the AI system is an amalgamation of data points that are capable of replicating something *like* the genuine thing, but, as is visible through the cracks, possesses no 'genuine' understanding of what it depicts.

To have genuine understanding, it needs to understand these objects as the paradigmatic (biological) mind does, in their multidimensionality.

4. Lost in (AI's) thought – introducing world-embedment

At this point in our considerations, our connectionist commitments seem to have taken us further from, not closer to, an answer. Out of one experientially internal, clearly delineated motivational nodule, one gets

a whole net of data points, each of them without an expressible purpose. Part of the problem facing the researcher at this point is the question of emergentism: How out of a net of non-singular data points, do we get something that humans undoubtedly have, a purpose they can state, or an object that functions as content for their intentionality?

The reason for invoking connectionism is not just that it is a more realistic picture of the internal states of a neural network, but that it is only this realistic picture that can take one from the non-object-focused base to a purpose, and the only way to present how it can be done, is through world-embedment.

World-embedment presents a neural network with its own outputs, which it then can analyse as inputs and, if it is processing the data in a dynamic-enough manner, is able to see the data as relevant for its next stages of operation, so to take into account the consequences of its actions already when making decisions. In simple terms, this means introducing into AI's information-transforming processes the same kind of simulatory power that humans possess; one that is able to take the agent's (potential) actions as input.

These decisions, then, which in reality are a set of signals in a net of non-singular data points will have to be conceptualised in an increasingly more useful way, as singular actions. In that, then, an AI system would get closer to understanding what it is doing in forming meta-perceptual teleological clusters, and through that understand also the world that surrounds it. In other words, by simulating its actions and their consequences to itself, a system has to conceptualise them in a higher-level way than entirely anonymous input.

Why has that not happened yet (at the time of writing this paper)? Organic systems, like human minds, have a significant advantage due to their natural sensory connection to the real world. Not many AI systems currently available possess such a broad spectrum of world-accessing (or world-embedding) tools. But why do these tools matter in the first place?

To fully understand the idea, one has to appreciate the externalist implications of world embedment, and why it is an essential feature of any system that is to pass the Chinese-Room-style or Putnam-style objections unscathed.

What is key in Putnam's *Brains in a Vat* essay (1999). is the role of the causal interaction between a mind (also 'a potential mind', like an AI system) and the object in the world that is depicted by it. Putnam writes:

> If apples, trees, steeples and fields had not existed, then, presumably, the programmers would not have produced that same program. Although the machine does not perceive apples, fields, or steeples, its creator—designers did. There is some causal connection between the machine and the real world apples, etc., via the perceptual experience and knowledge of the creator—designers. But such a weak connection can hardly suffice for reference. Not only is it logically possible, though fantastically improbable, that the same machine could have existed even if apples, fields, and steeples had not existed; more importantly, the machine is utterly insensitive to the continued existence of apples, fields, steeples, etc. (Putnam 1999, 11)

Putnam is right. His 'causal constraint on reference' principle states that: "A term refers to an object only if there is an appropriate causal connection between that term and the object" (Hickey 2024).

The key phrase being, of course, 'appropriate causal connection'. How should this requirement be understood, and how to justify what range of modes of being causally connected we find acceptable? It is the key idea of the present paper that connecting the already-existing neural network(s) to a whole range of sensors (including, but not limited to auditory, visual, and tactile ones, and one day perhaps smell and taste), can create clusters of data points which then resemble the human perception of the world and, further down the road, produce genuine understanding.

5. The Robot Reply

This possibility is addressed by Searle, when, after proposing the Chinese Room thought experiment, he discusses verious potential objections to the experiment that he forsees. 'The Robot Reply' stands out in particular:

'Suppose we put a computer inside a robot, and this computer would not just take in formal symbols as input and give out formal symbols as

output, but rather would actually operate the robot in such a way that the robot does something very much like perceiving, walking, moving about, hammering nails, eating, drinking – anything you like. The robot would, for example, have a television camera attached to it that enabled it to "see," it would have arms and legs that enabled it to "act," and all of this would be controlled by its computer "brain." Such a robot would, [the reply goes], have genuine understanding and other mental states.' (Searle 1980, 420) Searle disagrees with the above reply by mentioning that:

> I am the robot's homunculus, but unlike the traditional homunculus, I don't know what's going on. I don't understand anything except the rules for symbol manipulation. Now in this case I want to say that the robot has no intentional states at all; it is simply moving about as a result of its electrical wiring and its program. And furthermore, by instantiating the program I have no intentional states of the relevant type. All I do is follow formal instructions about manipulating formal symbols. (Searle 1980, 420)

Searle's response gives a counterintuitively stringent set of requirements for the data transformation processes of 'the robot'. The requirement that *inside the brain* there should be some data-processing being that is self-aware, or intentional, or knows what its function is, would constitute a circular explanation. If one postulates that for a system to have the ability to understand, it has to have in it a thing that understands, such an explanation explains nothing. To reiterate what **Figure 2** or the architecture of neural networks attests to: it is simply not the case that one of the data points, or one of the threads of data transformation is on its own in any way aware of what it is doing. In a neural network, it is simply impossible to match individual data points with whole thoughts, actions, or purposes.

The Robot Reply presents a compelling vision of how individual data-transforming processes together constitute a bundle that can create a unified mental picture of an object; but only together, not individually. There is nothing counterintuitive about a single data-transforming process operating on the 'Chinese Room basis', without any awareness or understanding of what it is doing. Genuine understanding comes only after all the data-transforming processes are grouped, forming a cluster.

6. Genuine Intelligence

To recapitulate, a necessary condition outlined by this paper for a system to be 'genuinely intelligent' is that it has to perceive the world and the objects in it in a multimodal fashion and has to be sufficiently aware of the consequences of its outputs for its future inputs. It can be the above-mentioned sensory (visual, auditory, tactile etc.) perceptions, it can be intelligent data transformations and grouping all these together, it can also be the recognition of behavioural patterns of other systems, and the ability to predict and explain their actions. And indeed, there is a sense in which intelligence is a feature of incremental levels.

The necessary feature of multimodal output is key, however, insofar as it can give rise to a possibly narrow necessary feature that would make a system 'genuinely intelligent'; the ability of a system to recognise itself as one of the objects which can constitute its input.

Some may be ready to grant an intelligence status to information-transforming computers, like a system from the Robot Reply, or some currently available AI systems.

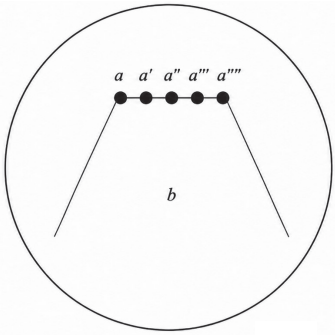


Figure 3

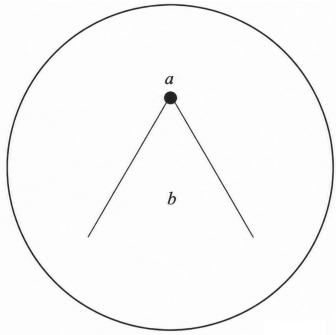


Figure 4

The view of this paper, however, is that for a system to be something more than a data-transforming machine, it has to possess some capacity to recognise what its own actions are, and how those might change its environment. The next section will discuss exactly why this feature is essential.

Firstly, the criteria for what counts as 'understanding' of an action as 'one's own' action has to be specified. Following a standard definition of 'understanding', the word denotes a process of perceiving the meaning, significance, explanation, or cause of an event. Importantly, this essay does not attempt to describe the process of understanding in full. What we are concerned with is the necessary (not sufficient) features of 'self-referential' understanding, and any other knowledge states. **Figure 3** illustrates a modified version of **Figure 2**, with the 'hidden' layer, adjusted in the training process (*b*), so the information-processing structure takes also itself into account, together with what *a*, *a′*, *a″*, *a‴*, and *a⁗* truly stand for. In such a process of recognition, a system can get to its own understanding of itself, somewhat similar to **Figure 1** – depicted in **Figure 4**.

In other words, understanding of a system's activity by that system as the activity of that system can only be achieved through a dynamic tracking of how an output produced by such a system influences its future inputs. This is to ensure that the system (even though initially the system is not, of course "aware" of this output being produced by itself) in question possesses an ability to perceive of an appropriate type; one which allows it to track the future implications of its present actions. We have remarked above that for a system to create a conceptual cluster meaningfully similar to ours, it has to have access to a possibly broad range of multimodality. The essential feature we are discussing now has to do with the system's conceptualisation of itself by perceiving its own world-embedment. This requires access to a full spectrum of sensory data, dynamic tracking of input-output influences, and repeated re-training to refine the system's understanding.

7. Answering the questions of self-reference

Taking stock of what has been said, two questions can now be answered. The first one is: 'Why is it so important for an AI system to "perceive itself", and what does that mean?', and 'Why are we highlighting the connectionist character of the "internal understanding" inside a neural network?'.

Starting with the latter, to point out that how a neural network 'perceives' what 'it is doing' is more complex than what folk psychology might suggest is necessary to dispel the aura of mystery surrounding such a question. In discussing how artificial intelligence (but also other systems that one may want to call 'intelligent') works, it is paramount to steer clear of a category mistake. In other words, one should not look for one teleological nodule labeled with the exact name of the function visible to the outside observers. Folk-psychological descriptions of actions differ fundamentally from an in-depth analysis of a system's structure.

In other words, we should not be looking for one neatly packaged box labeled, for instance, 'seeing' inside of a system.

This raises the former question: Why is it so important for an AI system to 'perceive itself', and what does that entail? Firstly, it means the system 'perceiving' itself in the world. Such self-referential loops not only avoid the question-begging 'premature identification' (if we put it like this, the claim can be misunderstood as begging the question about cognition; 'how does a system know that *its* internal processes are its?'); a system does not need to perceive its information-processing transformations as abstracted, only to recognise the correlation between *whatever it is that it will output* and a consistent change in the input it is getting.

Searle contends that one can imagine a system that transforms information, and yet is not conscious, because the hypothetical agent inside the system (the equivalent of a data-transforming process) has no understanding of what it is doing. I have criticised this view, highlighting the fact that it hardly seems problematic that some part of the process of understanding will not itself possess understanding (otherwise the only available model of consciousness would be a circular one, where to have an understanding of a process, one has to have an understanding of a process). I suggest, however, adopting a more favourable view of Searle's thesis. The point is this: genuine intelligence[7] (and with it, understanding)

[7] I am here adopting the most stringent requirement where in order to be genuinely intelligent the system has to be conscious. It is, I think, debatable whether that is the case, but, as I will show, even if one adopts such a requirement, genuine intelligence in AI systems is possible..

requires a genuine connection with the real world. This is because, according to Searle, there has to be more to consciousness than data-transformation. We are here also informed by Putnam's requirement of an appropriate causal connection. The 'genuine connection' that I propose requires perceiving oneself as part of the world that can act upon it and engaging with data in a way that allows outputs to influence future inputs dynamically. Seeing oneself as part of the world, then, emerges from the dynamic self-referential loops already mentioned.

A genuinely intelligent AI system would generate data, analyse real input, identify correlations between the two, and recognise the dynamic fashion in which its output influences future input.

A key aspect of this perspective is that both input and output data must be dynamically tracked and multimodal, as previously emphasized. A system that can dynamically foresee what symbols are going to be fed as its input when it produces a given output would be a competent player in the symbol game, but not yet an agent with the ability to refute all of Searle's objections. Incorporating multiple types of input and output feedback helps address this issue.

This is precisely what moving from **Figure 3** to **Figure 4** illustrates; becoming aware of what one's motivations are, or learning about their precise nature. It is by doing (outputting data), and correlating that doing with input (the preceding output being reflected by the world) that one (a system) can gain an adequate conceptualisation of what its motivations are.

Conclusion

This paper has examined the conditions under which an artificial intelligence system could be considered genuinely intelligent. Beginning with an analysis of the paradigmatic philosophical positions on AI of Turing and Searle, I have argued that although simple behavioral imitation or syntactic symbol manipulation is not sufficient to establish genuine intelligence, Searle's attitude is overly skeptical and the demands it makes seem to introduce circularity. Instead, I have suggested that for

AI to be genuinely intelligent, it should not just deal with information dynamically but also entertain self-referential thinking-apprehending itself as an agent in the world whose actions affect its environment.

One of the main problems of AI research, as I have shown, is the misguided focus on intentionality, especially as championed by Searle. It seems to require an AI researcher to look for a clearly identifiable, localised structure of intentional states within AI systems. I have argued that this approach is fundamentally flawed because intelligence is not located in any of the individual data-points but rather in the dynamic interaction of the sum of the data-points and the environment a system is in.

To support my argument, I proposed the concept of world-embedment, drawing on Putnam's externalism and connectionist networks, to show that genuine intelligence must involve the capacity for integrating multimodal information and changing dynamically based on its interactions with the world. Furthermore, I discussed how neural networks process information and argued that AI systems must develop a feedback mechanism whereby their outputs shape their future inputs, allowing for a form of self-awareness akin to emergent properties observed in cognitive systems.

Last but not least, there are two main ways in which my conclusions can enter into a theological debate about the idea of imago Dei.

Firstly, insofar as this paper described how genuine intelligence requires world-embeddedment, the thought here mirrors that of Charles Darwin:

> The high standard of our intellectual powers and moral disposition is the greatest difficulty which presents itself … But every one who admits the principle of evolution, must see that the mental powers of higher animals … are capable of advancement. (Darwin 1898, XXI)

The Darwinian case for eschewing the intelligence-as-distinctive-property model of imago Dei is, therefore, that because intelligence is an evolutionarily-emergent property, it is not something that is in a significant way distinct from the rest of the natural world.

The evolutionary principle, in turn, describes intelligence as a quality emergent "in response" to the external environment, safeguarding survival for the increasingly intelligent representatives of a species. The Darwinian relation between intelligence and the world mirrors the principle of world-embeddement put forward by this paper. The implications for imago Dei are, therefore, the same; any organism (or system) which is "world-embedded" enough to produce self-recognition of a system in the world counts as a genuinely intelligent one; it is thus not only humans who can potentially fit the intelligent requirement of this specific model of imago Dei.

References

Blazek, Paul J., and Milo M. Lin. 2021. "Explainable Neural Networks That Simulate Reasoning." *Nature Computational Science* 1 (9): 607–18. https://doi.org/10.1038/s43588-021-00132-w.

Chayka, Kyle. 2023. "The Uncanny Failures of A.I.-Generated Hands." *The New Yorker*, March 10. Accessed May 21, 2024. https://www.newyorker.com/culture/rabbit-holes/the-uncanny-failures-of-ai-generated -hands.

Cole, David. 2023. "The Chinese Room Argument." In *The Stanford Encyclopedia of Philosophy*, Summer 2023 ed., edited by Edward N. Zalta and Uri Nodelman. Accessed May 21, 2024. https://plato.stanford.edu/archives/sum2023/entries/chinese-room/.

Cruz, Helen, and Yves Maeseneer. 2014. "The Imago Dei: Evolutionary and Theological Perspectives." *Zygon* 49 (1): 95–100.

Darwin, Charles. 1871. *The Descent of Man, and Selection in Relation to Sex*. Vol. 1. London: John Murray. Project Gutenberg. Accessed May 21, 2024. https://www.gutenberg.org/ebooks/34967.

Dennett, Daniel C. 1978. *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge, MA: Bradford Books.

Dennett, Daniel C. 2007. *Consciousness Explained*. Boston: Back Bay Books.

Fergusson, David. 2013. "Humans Created According to the Imago Dei: An Alternative Proposal." *Zygon* 48 (2): 439–53.

Gallagher, Shaun, and Dan Zahavi. 2013. *The Phenomenological Mind*. London: Routledge.

Gupta, Bulbul, Tabish Mufti, Shahab Saquib Sohail, and Dag Ø. Madsen. 2023. "ChatGPT: A Brief Narrative Review." *Cogent Business & Management* 10(3). https://doi.org/10.1080/23311975.2023.2275851.

Hall, Brian K. 2012. "Evolutionary Developmental Biology (Evo-Devo): Past, Present, and Future." *Evolution: Education and Outreach* 5: 184–93. https://doi.org/10.1007/s12052-012-0418-x.

Hickey, Lance. 2024. "Brain in a Vat Argument." *Internet Encyclopedia of Philosophy*, May 26. Accessed May 26, 2024. https://iep.utm.edu/brain-in-a-vat-argument/.

Jackson, Frank. 1982. "Epiphenomenal Qualia." *The Philosophical Quarterly* 32 (127): 127–36. https://doi.org/10.2307/2960077.

Koide-Majima, Naoko, Shinji Nishimoto, and Kei Majima. 2024. "Mental Image Reconstruction from Human Brain Activity: Neural Decoding of Mental Imagery via Deep Neural Network-Based Bayesian Estimation." *Neural Networks* 170: 349–63. https://doi.org/10.1016/j.neunet.2023.11.024.

Matthias, Meg. 2023. "Why Does AI Art Screw Up Hands and Fingers?" *Encyclopedia Britannica*, August 25. Accessed May 21, 2024. https://www.britannica.com/topic/Why-does-AI-art-screw-up-hands-and-fingers-2230 501.

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

Ozcelik, Furkan, Bhavin Choksi, Milad Mozafari, Leila Reddy, and Rufin VanRullen. 2022. "Reconstruction of Perceived Images from fMRI Patterns and Semantic Brain Exploration Using Instance-Conditioned GANs." In *2022 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Putnam, Hilary. 1994. "Sense, Nonsense, and the Senses: An Inquiry into the Powers of the Human Mind." *The Journal of Philosophy* 91 (9): 445–517. https://doi.org/10.2307/2940978.

Putnam, Hilary. 1999. "Brains in a Vat." In *Knowledge: Readings in Contemporary Epistemology*, edited by Sven Bernecker and Fred I. Dretske, 1–21. Oxford: Oxford University Press.

Ramsey, William, Stephen Stich, and Joseph Garon. 1990. "Connectionism, Eliminativism and the Future of Folk Psychology." *Philosophical Perspectives* 4: 499–533. https://doi.org/10.2307/2214202.

Roumeliotis, Konstantinos I., and Nikolaos D. Tselikas. 2023. "ChatGPT Training Process." *Encyclopedia.pub*, May 29. Accessed May 17, 2024. https://encyclopedia.pub/entry/44957.

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

Templeton, Adly, et al. 2024. "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet." *Transformer Circuits* (thread). Accessed

May 21, 2024. https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Trzęsicki, Kazimierz, Stanisław Krajewski, and Jan Woleński, eds. 2012. *Papers on Logic and Rationality: Festschrift in Honour of Andrzej Grzegorczyk*. Białystok: University of Białystok.

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

Visala, Aku. 2014. "Imago Dei, Dualism, and Evolution: A Philosophical Defense of the Structural Image of God." *Zygon* 49 (1): 101–20.