religions

MDPI

Article

Powers of the Soul Beyond AI

Angus John Louis Menuge

Department of Humanities and Social Science, Concordia University Wisconsin, Mequon, WI 53097, USA; angus.menuge@cuw.edu

Abstract

Could Large Language Models (LLMs) exhibit rational characteristics traditionally attributed to the human soul? I argue that five features of human rationality will likely remain beyond LLMs and other adaptive physical systems. *Insight into truth*: using billions of pages of text, a LLM may harvest a sound rule of inference. However the LLM has no insight into *why* the rule is true. *Meta-insight*: both humans and machines can follow instructions that constitute an infinite loop. Yet humans can, but machines cannot, recognize that they are in an infinite loop. *Free will*: once humans realize they are trapped in a loop, they can exercise free will to break out of the loop. By contrast, when a machine is trapped in an infinite loop, an external intervention is required to end the task. *Access to necessary conceptual relations*: LLMs are inductive learners and cannot justify universal necessary truths. By contrast, a human being can, via insight, see that a conceptual relation is necessarily true. *Non-combinatorial creativity*: LLMs can recombine the products of human creativity in amazing ways. But unlike humans, they cannot use universal concepts to find a possible item that is not derived from items already instantiated in the world.

Keywords: AI; the soul; consciousness; rationality; free will; creativity

1. Introduction

Anthropic mechanism claims that all aspects of a human being can be explained in mechanical terms. Opponents typically appeal to powers of the mind (I will argue, of the soul), such as consciousness, intentionality, and reasoning, that resist mechanical analysis. Yet some defenders of anthropic mechanism think that recent developments in Artificial Intelligence (AI) are beginning to overcome these objections, because current AI systems seem to exhibit basic forms of the capacities that had been thought unique to the human mind. Large Language Models (LLMs) in particular appear to have made impressive gains in this area. These include: the spontaneous acquisition of the rules of grammar, a prodigious capacity for automatic language translation and computer code generation, the rapid resolution of protein folding problems, and the ability to optimize logistical and industrial processes, to provide accurate medical diagnoses based on exhaustive analysis of extant scientific literature, and to produce (mostly) comprehensible texts and appealing music and art.

However, as impressive as these achievements are, I will argue that they are not a good reason to think that the human mind (and hence the human being as a whole) is *merely* a machine. In what follows, I will first briefly unpack the thesis of anthropic mechanism in its contemporary form and outline the developments in LLMs that may seem to provide its thesis about the mind with convincing support (Section 2). Then I will argue that five characteristics of the human soul are beyond current LLMs and that these limitations seem

Academic Editor: John A. Bloom

Received: 4 December 2025
Revised: 12 December 2025
Accepted: 15 December 2025
Published: 22 December 2025

to apply to *any* purely physical adaptive system, including brains and AI more generally (Section 3). We will see that there are good reasons to think that only an immaterial entity could manifest these powers.

2. Anthropic Mechanism

The English political philosopher and materialist Thomas Hobbes (1588–1679) argued that our mental states are (or entirely depend on) motions in the brain, and that reasoning is only "reckoning" (Hobbes [1651] 2008, p. 23) or computation, making an immaterial soul redundant. Julien Onfray de la Mettrie (1709–1751) concurred, claiming that the "soul" reduces to an "enlightened machine" (de la Mettrie 1943, p. 128). But the thesis of anthropic mechanism about the human mind remained both vague and implausible without a sufficiently clear and general definition of "machine". It is obvious that humans are not identical to many particular machines, like clocks or sewing machines, because these devices have a dedicated function, yet humans are capable of an indefinitely wide range of tasks.

However, Alan Turing (1912–1954) developed a mathematically precise account of a general-purpose problem-solver. His idealized "Turing Machine" captured the essential character of all programmable digital computers, and this allowed anthropic mechanism to be more sharply stated as the computational theory of mind (Rescorla 2024). According to this theory, a human being is an organic realization of a Turing machine, and this implies that the powers of the human mind reduce to those of a suitably programmed digital computer.

Today, enthusiasts for this view believe in Strong Artificial General Intelligence (Strong AGI). Unlike *Weak AI*, which claims only that computers can simulate mental activities, *Strong AI* asserts that "the appropriately programmed computer *is* a mind. . .[and that] computers given the right programs can be literally said to *understand* and have other cognitive states" (Searle 1980, p. 417). And unlike the early examples of AI, which were inflexible, domain-specific, and highly dependent on having rules and targets programmed in (e.g., game-playing programs, theorem provers, and expert systems), proponents of AGI claim that recent advances in AI are highly versatile and can discover new patterns, including rules that were not even implicitly present in their programming. These systems, it is claimed, demonstrate "a capacity to learn. . .the ability to deal effectively with uncertainty. . . [and a] faculty for extracting useful concepts. . .for. . .logical and intuitive reasoning" (Bostrom 2014, p. 23).

Enthusiasts for Strong AGI, like Ray Kurzweil, claim that at some point in this century, artificial systems will be able to adapt their problem-solving strategies to any domain and think in the same, fully general way as humans. This will culminate in a "singularity", when artificial systems will meet or exceed the standards of human intelligence in all areas (Kurzweil 2024). Some evidence of this was offered two years ago, when fourteen Microsoft researchers argued that the LLM, GPT-4 was already exhibiting basic forms of the rational characteristics traditionally attributed to the human soul. Their systematic study claims to demonstrate that:

> GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4's performance is strikingly close to human-level performance.
>
> (Bubeck et al. 2023, p. 1)

As evidence, the investigators note that GPT-4 can combine rules from multiple domains. For example, it generated a Platonic dialogue that applies Plato's critique of

rhetoric and sophistry to autoregressive language models (of which GPT-4 is one), pointing out the dangers of AI-generated deepfakes and other misinformation for manipulating public opinion. It also provided a version of Euclid's proof of the infinity of prime numbers in the form of a Shakespearean play!

This ability to synthesize multiple genres while providing a coherent response to a non-trivial question is impressive, and it is unquestionably an advance on earlier, domain-specific AI. If all one is interested in is behavior, it leaves little doubt that GPT-4, its successors, and rival LLMs can generate *output* which unaided human beings would require intelligence to produce. However, the deeper question is whether such systems can really be credited with human *powers* of reasoning. Are the operations of these LLMs signs of human-like cognition? I will now argue that there are five features of a rational soul which are beyond current LLMs and will likely remain beyond not only LLMs, but all adaptive physical systems. These are: (1) insight into truth, (2) meta-insight, (3) free will, (4) access to necessary conceptual relations, and (5) non-combinatorial creativity.

3. Powers of the Soul Beyond LLMs

Elsewhere, I have argued for the existence of an immaterial soul (Menuge 2004, 2009, 2016, 2018). But the arguments I give here do not depend on my own view of the nature of the soul. My own sympathies lie with a version of Augustinian substance dualism, according to which body and soul are coadapted to exist as an integrated body-soul union in which the whole soul is present in every part of the body, and I hold that the soul is required to explain not only the powers of consciousness, intentionality, and free will, but also a human's organismal identity as one living thing at and over time. But many others follow Aristotle and Aquinas in maintaining some version of hylomorphic dualism, according to which a rational soul is not a strict substance, but is rather the form of a person's body. And as Goetz and Taliaferro (2011) show in their historical survey of views of the soul, there are many other options as well. Since the arguments I give here merely point to the fact that there is *something* immaterial beyond the mechanical operation of the human brain, they do not (by themselves) offer a decisive reason for preferring one of these options.

3.1. Insight into Truth

There is no doubt that LLMs can find truths. Trained on billions of pages of carefully curated data, ChatGPT 4 (and higher) and other LLMS can generate a lot of accurate information about history, geography, physical science, psychology, and more. Impressively, by detecting patterns in this data, they can derive correct rules that were not explicitly programmed into the system. While it is common in the AI field to call this "machine learning," if by "learning" we mean the acquisition of *knowledge*, we will see that there are serious grounds for doubting that this is literally occurring. ChatGPT was designed as a conversational chatbot and trained on enough data, it finds probable, and mostly sensible completions for prompts, by analyzing them into a series of tokens (Wolfram 2023). As a trivial example, it can tell that 'mat' is a more probable completion than 'aardvark' for the prompt, 'The cat sat on the. . .' But what programmers did not expect, was that ChatGPT would also learn the basic rules of English grammar, so that it can, not only generate grammatically correct sentences, but also detect grammatical errors and suggest remedies. Does that mean that ChatGPT has gained *insight* into truths about English grammar?

But I will not linger on this example, because much of grammar is conventional, and so one might wonder if its rules rise to the level of interesting truths. What we need is a case where a LLM finds a statement or rule that is clearly and substantively true, so that we can consider whether there is reason to think that the LLM has insight into its truth. A

good test case is provided by basic rules of logic, like *modus ponens*. This rule tells us that if both a conditional, if P then Q, and its antecedent P are true, then the consequent Q is true. It is easy to show that, expressed as the single statement 'if ((if P then Q) and P) then Q', this rule is a tautology, and hence necessarily true. There is no doubt that, trained on enough examples of the correct use of *modus ponens*, the LLM can find this truth, and may even be able to check whether other reasoning complies with it. But is this a reason to think that the LLM has acquired *insight* into the truth of *modus ponens*?

The answer to this question clearly depends on what one means by 'insight'. As I will understand the term, to have insight into the truth of a statement is to understand the reason *why* it is true. If I see a terrible traffic accident involving cars on fire, I certainly encounter a truth, but may be unclear why it is true. Later, if I learn that a gasoline truck jackknifed, and was ruptured in a collision with another vehicle that caused the gasoline to ignite, I will have insight into the truth I had encountered. Likewise, any beginning logic student may be shown the rule *modus ponens*, and even acquire some ability to use it, without immediately having insight into its truth. This insight may only come later, with the application of truth-tables for validity testing, or more rigorously, via a soundness proof which shows that *modus ponens* is truth-preserving.

While the LLM can certainly find *modus ponens*, there are three reasons to think that it does not have insight into the truth of the rule. First, all that the LLM can discover is responses to prompts that are probable given its training data (plus any reinforcement or prompt engineering designed to steer it in the right direction). But this means that if the LLM were trained on carefully curated examples of fallacious reasoning, it could just as easily have found an unsound rule, such as affirming the consequent: 'if ((if P then Q) and Q) then P.' It follows that even if the LLM gets the right answer when given good training data, it is only right by accident, since using exactly the same method, it could easily have found a wrong answer.

A famous example from epistemology is the broken clock. It may happen that one looks at the clock at just one of the two times a day that it is correct. But that does not mean one has knowledge of the time, because the clock is not a reliable indicator of the time. Similarly, even though the LLM may find a correct rule of logic, the process it uses to acquire that rule is not a reliable one. This is important because insight is a form of *knowledge* (analogous to recognition or, under the right conditions, perception), and one cannot know something if one is right by accident. This would be like guessing that the answer to a multiple-choice test was "C" and discovering that, by good fortune, it was. The process that derived the right answer was not reliably connected to its truth and did not involve any insight into its truth.

This leads to a second point. When one really has insight into the truth of a statement, one's belief stands firm in the face of contrary proposals. Plato noticed that knowledge is resistant to change because, unlike mere opinion, it is anchored in a *logos*, the reason why the statement believed is true:

> True opinions are a fine thing and do all sorts of good so long as they stay in their place, but they will not stay long. They run away from a man's mind; so they are not worth much until you tether [or anchor] them by working out the reason.
>
> (Plato 1985, 97e–98a, p. 381)

Thus opinions about a football team's chances of winning may shift from week to week, but it is hard to imagine anything that could budge our confidence in A = A or 2 + 2 = 4.

But here there is a clear asymmetry between the LLM and (at least some) human beings. If a trained logic student has followed the soundness proof for propositional logic, including *modus ponens* as one of the cases, she will know that this rule is true on all

valuations and can never lead us from true premises to a false conclusion. Once insight into this truth has been achieved, not even months of exposure to fallacious reasoning will have the slightest tendency to make the student abandon the rule. However, even if the LLM has converged on *modus ponens*, all we have to do is to retrain it on a barrage of fallacious reasoning, and it will quickly abandon the rule and may settle on unsound alternatives. This is a convincing reason to think that the LLM never had insight into the truth of *modus ponens* in the first place: though it found the rule, it was not anchored by any understanding of why the rule is true.

Third, this lack of insight traces back to the fact that there is no reason to think that the LLM has any understanding of the responses it makes. For example, when ChatGPT generates a response, it does so one token a time, based on an analysis of the probability of that token, given its training data. This analysis operates purely at the level of syntax (tokens are evaluated as series of characters) and at no point requires the system to know either the sense or the reference of any of the terms it manipulates. In order to estimate that 'mat' is a more probable completion than 'aardvark' for the prompt, 'The cat sat on the. . .' the system does not need to know what 'cat' and 'mat' refer to, or what it means to sit on something.

Despite its vastly greater computing power, in this respect, ChatGPT is no advance on the much earlier AI systems criticized by Searle (1980) in his famous Chinese Room argument. Searle pointed out that he can simulate the intelligent behavior of a Chinese speaker by matching the syntax of Chinese questions and answers, but with no understanding of what the symbols mean. He argued that extant AI performs the same kind of pattern matching of symbols without understanding their sense and reference, because the states of these systems have no intentionality: they are not really *about* something beyond themselves in the way that a human thought can be *about* items or relations in the world. Similarly, the LLM can generate statements using the tokens 'cat' and 'mat' without these terms meaning anything to the LLM. But without intentionality, there can be no understanding: the LLM is processing uninterpreted patterns, and even its rules for doing so are rules that computer scientists understand, but the LLM does not. Thus, there is no reason to think that the LLM understands the meaning of its output, even if it is meaningful to human beings. But without understanding, there can be no insight into the truth of a statement. So even if 'The cat sat on the mat' is true, the LLM cannot have insight into its truth.

If even statements about unproblematically physical items, like cats and mats, pose challenges for the LLM's insight, matters are obviously much worse for the truths of logic and mathematics. Since the LLM is instantiated in a purely physical system, just how can it access abstract objects like *modus ponens* and other logical relations, numbers, axioms, etc., ? Abstract objects do not exist in space and time and therefore cannot interact with physical objects like computers. Even if the LLM could acquire some basic intentionality by causal interaction with the physical world, this would not explain insight into truths about abstract relations.

3.2. Meta-Insight

Both computers and human beings can follow instructions that constitute an infinite loop. But is there a difference in the way computers and human beings handle infinite loops? I will argue that there is. Consider a few examples.

Within a computer program, there are numerous ways to create infinite loops. One may simply code:

Step 1: call Step 1

Or one may code:

Step 1: call Step 2
Step 2: call Step 1

Or a method may recursively call itself but without reducing a parameter to a base case that halts the process. In general, any code that begins with a false loop termination condition and precludes the satisfaction of that condition will generate an infinite loop.

Humans also can be presented with infinite loops. A mischievous colleague of mine included the two following entries in the index of one of his books: 'Infinite loop: see vicious circle', and 'Vicious circle: see infinite loop'. But it seems that the human response to encountering an infinite loop is very different from a computer's. Normally, an infinitely looping program cannot stop itself: it can only be terminated via an external intervention, either by the user (e.g., using the task manager to end the process) or by the operating system (if it monitors and limits the allocation of system resources and times out repeating processes). But those humans who have read my mischievous colleague's index did not require an external intervention to exit the infinite loop—which is a good thing, as they may otherwise have starved to death in his office!

A plausible explanation of the different response to infinite loops is that human beings have what I call *meta-insight*. While insight permits one to understand why a statement is true, meta-insight moves up a level and tells an individual what acceptance of one or more statements entails. In this case, the reader of my colleague's index has the meta-insight *that* the entries constitute an infinite loop and sees that if she continues to follow their direction, she will be stuck in an endless process. This meta-insight leads *her* to abort the process. This does not seem analogous to the external intervention of a user or the operating system, since this decision to step out of the loop is generated internally. It is the very being that is following infinitely looping instructions that becomes aware of that fact and decides to exit the loop, strongly suggesting that this being is something independent of the instructions themselves. By contrast, it seems that by their nature LLMs (and any other computational systems) must continue to execute whatever instructions are presented to them unless there is an external intervention of some kind. The only exceptions would be a loss of electrical current or the effects of entropy, but they do not entail an internal power of the system to terminate loops.

But a critic may press the two following questions. First, granted that LLMs and many computational systems lack meta-insight, why should we think they *must* do so? Perhaps this problem can be solved by ingenious programming. Second, why are we so sure that human beings really have meta-insight? Perhaps human beings are not that good at knowing where instructions are leading them either.

The main reason to doubt that computational systems have meta-insight is that it appears to require self-consciousness, and there is good reason to think that computational systems lack any kind of consciousness. In any physical system, we see a system of separable parts in external relations to each other, but we do not see thoughts which by their nature are inseparable from, and internally related to, the subject of those thoughts. One can remove any physical part of a computational system from that system and it remains intrinsically the same either isolated from any system or integrated into a different system. But thoughts cannot exist ownerless: one cannot put a thought on a shelf, as it could not *be* a thought if it did not belong to some mind. And while two thinkers can agree on the same thought content (they can both agree that the NFL season is too long), they cannot share the numerically same thought, because part of what makes a thought the thought that it is, is the one that thinks it. Similarly, if you and I have qualitatively identical pains, still they are two pains, not one, because they have different subjects.

Another reason to doubt that machines are conscious is that consciousness provides a unified subject at and over time, but machines appear to be shifting aggregates of externally

related entities. There is no identifiable subject of the machine's states at a time but only a multiplicity of parallel, distributed information processing streams, and nothing which endures as the very same thing over time. If we think, for example, of the many layers of "neurons" in transformer models like ChatGPT, one cannot identify a single subject of its processing at a time (since there are many parallel informational streams) and the weights of its neurons are in constant flux over time.

Still, one might suggest that there is a way to give a machine something analogous to consciousness. Perhaps we can add a layer of monitoring software to the programs a system is running, so that the monitoring layers becomes "aware" of what those programs are doing? Thus, even if a program P that is stuck in an infinite loop cannot halt itself, we could add an error-handler E which can detect that P is likely in an infinite loop and abort the program. Now there are technical problems in this neighborhood. In 1936, Alan Turing demonstrated that there is no universal algorithm that solves the "halting problem": no computer can always tell whether an arbitrary computer will halt or not, because the assumption that this is possible leads to contradictions. In particular, there can be no general algorithm that solves the *self*-halting problem, which would allow a program to determine whether it will enter an infinite loop (see Boolos et al. 2002, pp. 35–40). So while an error-handler might work in some cases, there is no generally reliable way for a program to determine if it is in an infinite loop.

But I will not press this issue, as it is implausible that humans can *always* tell whether or not they are in an infinite loop: the instructions might, like the US tax code, just be too complex for anyone to tell. And it is possible, in *some* cases, for an error-handler to tell that a program is an infinite loop, e.g., by noting that the value of a variable that determines loop termination is not being updated. The more important point is that there is a fundamental difference between the *way* machines and humans exit infinite loops (even if they are sometimes not strictly known to be infinite, but only suspected of being so).

This traces to an insight of Lucas (1961) about the very nature of machines. He pointed out that as a formal system, a Turing Machine is defined by its instruction set, so that if one adds, removes, or modifies instructions, one has a different machine. So, suppose that a system M contains instructions that constitute an infinite loop, but that M has no way to detect this problem. And suppose that though there is no general solution to the halting problem, we can write an error-handler E which can usually detect when M enters an infinite loop and sends an interrupt so that M aborts. The problem is that this does not show that M could be aware it is in an infinite loop, because M + E is a *different* machine M*. Furthermore, because there is no general solution to the self-halting problem, M* has no way of being aware if it may enter an infinite loop of its own. Supposing that this can be fixed by the addition of another error handler E′, this again does not show that M* could become aware of its problem, because M* + E′ is yet another system, M**, and M** also cannot tell if it will enter an infinite loop.

However, while it is true that human beings cannot always know if they are following instructions that constitute an infinite loop (due to their complexity), in those cases where they do know, and in others where they have a reasonable suspicion of an infinite loop or are just fed up with a process (which may be an infinite loop or a very lengthy but finite process), it seems they have a radically different way of exiting the loop. This is because a self-conscious being can become aware that it is likely involved in an infinite loop via an internal power (meta-insight) and not by the addition of layers of error-handlers. As Lucas said:

> [A] conscious being can both consider itself and its performance and yet not be other than that which did the performance. A machine can be made in a manner of speaking to "consider" its own performance, but it cannot take this

> "into account" without thereby becoming a different machine. . . But it is inherent in our idea of a conscious mind that it can reflect upon itself and criticize its own performances, and no extra part is required to do this.
>
> (Lucas 1961, p. 125)

Thus at least sometimes, a conscious being can know that it is (or likely is) in an infinite loop, but without changing its identity. But a machine can only detect an infinite loop by the addition of something external: either a user intervention or an operating system command or the addition of new code that makes it a different machine.

Once this point is realized, it becomes clear that it does not just apply to the unusual case of infinite loops. When considering instructions, conscious beings always have the ability to obey or disobey them without becoming different beings. That is why drivers zoom past speed limit signs and shotgun pellets are found embedded in "No hunting" signs. While human defection from rules may be justified (resistance to tyranny) or unjustified (sinful rebellion), it clearly shows that conscious beings are not defined by their instruction set. True, they are aware of other instructions besides the ones they are immediately considering, but there seems to be no limit to the rules that a conscious being can defect from obeying. Even in the case of *modus ponens*, which we know to be sound, if we think that the conclusion Q is false, we can always decline to draw that conclusion by holding that one of the premises is false. In fact, as C. S. Lewis points out, unlike an automated system programmed to construct logical proofs, human beings often decline to draw logical conclusions from premises they know (Lewis [1960] 1996, p. 25). This may be for good reasons (doubt about the premises) or bad ones (mere dislike of the conclusion), but it evidences an independence from the rules that machines seem to lack.

A skeptic might reply that conscious beings may simply have a good built-in error-handler, that we are machines that include a supervisory machine on top of the basic machine. But this does not seem plausible because machines are defined by the parts that compose them, so we would be composite entities consisting of separable parts (in this case a basic program and an error-handler). However, consciousness is simple in the sense that it is not composed of, or defined by, separable parts. Neither our thoughts and experiences, nor our various faculties, are separable from the conscious being they belong to in the way that an error-handling subroutine is separable from the program that it monitors. Further, while adding an error-handler results in a different machine, thinking new thoughts does not produce a different thinker. A conscious being cannot be defined by the particular states and instructions it manifests at a particular time because it can remain identical over a time during which those states and instructions change. In particular, it can persist as the very same conscious being during the time it takes to conclude that it is following an infinite loop and should therefore exit.

3.3. Free Will

That said, consciousness alone is not enough to explain our ability to exit infinite loops. After all, a conscious being might enter an infinite loop and watch herself repeating instructions as a passive spectator. Even though consciousness grants an independence from the instructions we are entertaining, we need something more to explain how we can abandon those instructions. For this it seems we need an active power to redirect our actions, "free will" in some strong sense. There is good reason to think that no computational system has active power of the kind required to exit an infinite loop or disobey other instructions, but considerable evidence that conscious beings do have this power, indeed that it is a power of the soul.

Whatever a machine does is the passive result of its input and program instructions. Most computers are completely deterministic in the sense that given their total state at one

time (including their data and instructions), there is only one possible subsequent state. So if a machine is executing instructions constituting an infinite loop, and each iteration is determined by the previous one, it has no way to terminate that loop. It is possible to create indeterministic machines, where their subsequent state depends in part on whether a random event (such as the decay of a radioactive nucleus) occurs. But still, this does not grant active power to these machines. Whether prior causes strictly determine or only fix the chances of the next state of the machine, that state is still the passive result of prior events. The machine has no internal active power to redirect, so that it does not enter the next state, or to change the chances of doing so.

Why is this the case? I think the answer is that a machine is not a metaphysical substance, that is a singular, unified being that maintains strict identity over time despite changes in its states and accidental properties, and which has active causal powers of its own. Whether physically or functionally conceived, a machine is an externally configured artifact consisting of an aggregate of parts with purely passive liabilities to do things under various conditions. There is no one well-defined entity, like an organism, that persists as a being of the same natural kind over time, and which has active powers of its own. More particularly, there is no agent, a being with the rational capacity to make choices for its own reasons, rather than simply following the directions in its hardware or software.

But I think we have ample evidence that conscious rational beings do have this active power, and since active power belongs to substances, this is a reason to think that human persons either are or include a metaphysical substance. Further, there is reason to think that, whether we attribute this active power to the soul or to the person as a whole, this power derives from an immaterial entity. So that my argument does not depend solely on a priori metaphysics, I will offer examples drawn from psychology which provide clear evidence that conscious rational beings have the active power of free will.

William James (1842–1910) was one of the first psychologists to note the phenomenon of selective attention. Unlike a computer, which, given its instructions, must process the same data in the same way, human beings can choose to process the same data in different ways solely based on their interests. Suppose I attend a cocktail party with my beloved colleagues, and they are all holding forth about their latest theories. At first, it is impossible to make out any intelligible speech as their voices merge into a babble. But then I see an old friend, JP, and want to know what he has to say. I find an internal power to tune in to JP, and to tune out the mutual suppressors of other voices. To be sure, one could train a speech processor to converge on the particular tones of JP's speech and to eliminate others. But this is not what I am doing: I am deliberately focusing my attention on one particular voice, and this is an *active decision* of mine, even if it is informed by my prior listening to JP's presentations and podcasts.

This fact about the ability to focus conscious attention is very important to a number of cognitive therapies. Those who have suffered psychological trauma due to witnessing highly disturbing scenes find it difficult to avoid suffering anguish and other negative emotions when presented with triggers that remind them of these scenes. However, though it may not solve all the problems, cognitive therapy can help patients fight these reactions. In a pioneering work by Oschner et al. (2002), it was shown that patients have the power to reappraise otherwise disturbing images, so that, over time, the associated negative emotional response diminished. On "attend" trials, subjects were told simply to monitor their natural emotional reaction, so they passively processed the data as if they were a machine and had their normal reaction, leading to great activity in the limbic region which is correlated with the emotions. But on "reappraise" trials, patients were asked to reinterpret the images in a more neutral, impersonal fashion, and found that they could

actively suppress their normal emotional reaction, leading to diminished activity in the limbic region. So the very same data were processed differently on the two trials.

Of course, a proponent of anthropic mechanism may claim that the experimenters' instructions changed the patient's 'program'. However, what this ignores is that in cognitive therapy generally, the efficacy of instructions depends on the patient's *decision* to follow them. This does not always happen and can be very difficult for patients. This is shown even more clearly by cognitive therapies for stroke rehabilitation, depression, and Obsessive–Compulsive Disorder (OCD). Whether a patient uses his willpower to get behind the therapy has an enormous impact on its efficacy (Schwartz and Begley 2002).

Patients with OCD experience a repeated urge to do something that they know does not need to be done (such as locking a door they know they have locked or turning off an oven they know they have just turned off). OCD leads to a detectable abnormality in the brain, known as "brain lock" (Schwartz and Begley 2002, p. 85). Left untreated, if the patient yields to these urges, the OCD becomes worse, and brain lock comes close to resembling an infinite loop in the brain, as one circles round and round, worrying about doing a futile task.

For many years, psychiatrists followed a materialistic paradigm, congenial to anthropic mechanism. They assumed that only passive conditioning (similar to changing the training history of an LLM) or medication (similar to a change in a computer's hardware) had any prospect of solving these problems. Inspired by William James and also the emerging evidence of neuroplasticity (the fact that neural pathways can be changed even in mature adults), UCLA psychiatrist and neuroscientist Jeffrey Schwartz wondered if patients themselves had the power to fight their condition by the active power of conscious will. He developed a multi-step program which allows patients to identify their condition, distance themselves from it, and, when seized with an urge to repeat a futile task, to refocus on an alternative behavior. The critical element in the therapy is patients' own will power: patients must focus hard on doing something other than what the OCD urge prompts them to do. Schwartz found that in only ten weeks of therapy, there were detectable changes in the brain as the brain lock was broken and healthier neural pathways were formed (Schwartz and Begley 2002, pp. 89–90).

To be sure, defenders of anthropic mechanism (or materialism more generally) may claim that all of this can be explained by one part of the brain gaining control over another until the brain lock condition is shut down. However, this ignores several facts. First the therapy depends on conscious attention, and what Chalmers (1996) famously called the "hard problem of consciousness" is that none of the physical descriptions of a system (including the brain and computers) imply that the system is conscious. So the assumption that conscious attention can be reduced to brain activity lacks scientific justification. Unless one wishes to embrace epiphenomenalism (the thesis that consciousness has no effects), one has to consider the possibility that a non-physical entity, consciousness, has the active power to change the brain. Secondly, to speak of one part of the brain "taking control over another" ignores the fact that cognitive therapy is highly intentional: one follows the therapy in order to break OCD. But no physical description of the brain implies that it has goals or purposes or that its states are intentional. This is an example of "minding up" the brain, of attributing "psychological predicates to biological tissue" without clear justification (Robinson 2011, p. 62). There is a similar danger in speaking of "supervisory" software monitoring lower-level programs, as if the higher-level software had agency and insight: this cannot be justified by any sober physical or functional analysis of what the software is actually doing. The causal powers of computational systems and physical brains seem entirely exhausted by passive liabilities and event causation. So it seems that

the basis for intentionality and the active power of free will must reside in an independent immaterial soul.

3.4. Access to Necessary Conceptual Relations

For simplicity, let us call brains, and various AI systems, "adaptive physical systems", physical systems that can derive new patterns of information (such as new rules) and offer new responses (such as completions of prompts or novel behaviors). LLMs are a particularly sharp illustration of the fact that the causal powers of adaptive physical systems are limited by the contingent history of their interactions with their environment. LLMs are backward-looking inductive learners that can find patterns derived from their training data and reinforcement, and adaptive physical systems in general also seem confined to registering contingent information derived from their interaction with the world.

But this means that, at best, adaptive physical systems can find "rules of thumb", rules that have applied or "worked" in the past, but which are by nature defeasible due to future discoveries. An implication is that, even if an adaptive physical system finds a necessary truth like *modus ponens*, it could never be justified in claiming that it was a necessary truth. Though the system can, with the right training, converge on the rule *modus ponens*, all it is entitled to conclude is that this rule is highly probable on the basis of its past experience. But the fact that a rule has worked in the past does not logically imply that it will work in the future, or, even if it does, that it necessarily holds, not only in the actual world, but in all possible worlds.

In his critique of austere versions of empiricism, Immanuel Kant memorably states the problem this way:

> Experience teaches us that a thing is so and so, but not that it cannot be otherwise. . . [E]xperience never confers on its judgments true or strict, but only assumed and comparative *universality*, through induction. . .. When, on the other hand, strict universality is necessary to a judgment, this indicates a special source of knowledge, a faculty of a priori knowledge.
>
> (Kant [1781] 1982, pp. 43–44)

Kant's point is that necessary truths cannot be recognized simply by interacting with particular examples of them. No matter how many physical token instances of *modus ponens* a physical system has encountered, it will have no basis, in that contingent history alone, for justifiably concluding that *modus ponens* is a necessary truth.

However, we do know that rational conscious beings can recognize necessary truths. Typically in their second class, logic students go through the soundness proofs for first order logic, and one of the easier cases is proving the soundness of *modus ponens*. The usual procedure is to show, using mathematical induction, that in any proof where we assume earlier lines (including the conditional and its antecedent) are true, we cannot derive a false conclusion by inferring the consequent. But mathematical induction is itself a demonstrably deductively valid principle. And so, since the proof of the soundness of *modus ponens* applies to derivations of arbitrary length, we can validly infer that the rule can *never* lead us from true premises to a false conclusion in an *infinite* number of cases. Even though the logic student, just like an adaptive physical system, can only interact with a finite number of physical token instances of *modus ponens* (in textbooks, on a whiteboard, or paper, etc.) he can see that the rule holds necessarily, not only in the future of the actual world, but in all possible worlds.

What this shows is that conscious rational beings are not limited in their insight to recognizing truths based on experience, but can access necessary relations between universals. One has to see that for any of an infinite number of instances of 'if P then Q and P', the conclusion that Q follows. But to see this, one must have access not only to

propositions, but general types of propositions and the abstract relations between them. Arguably, none of this is possible for an adaptive physical system, because propositions, and even more so, types of propositions (meta-variables), are abstract objects that do not exist in space and time (for discussion, see Moreland 2009, pp. 87–88). Moreover, the entailment relation is clearly not a physical entity because it holds a-temporally between propositions. So there is no way that an adaptive physical system could interact with propositions, types of propositions, or entailment relations to gain insight into a necessary truth.

C. S. Lewis recognized that the problem for any physical system is that it is fully interlocked with the causal nexus, a nexus of contingent relations between events, and has no way to break free of that nexus to discern necessary truths. But it seems clear that logic students can become acquainted with propositions and the relations between them, for otherwise they could not understand the soundness (and completeness) proofs for first order logic. If so, and this cannot be explained by any physical power of the organism, it is a reason to think that immaterial souls transcend physical limitations and can perceive abstract objects and relation between them (see Menuge 2016 for a detailed defense of this argument). While computers and other adaptive physical systems can only learn from their interactions with physical token instances of rules, the soul has the power to abstract universal concepts that ride above particular instances, and through inspection of these universal concepts it can, via a direct, non-sensory form of perception, see that there are necessary relations between these concepts. As C. S. Lewis so beautifully put it:

> My belief that things which are equal to the same thing are equal to one another is not at all based on the fact that I have never caught them behaving otherwise. I see that it 'must' be so.
>
> (Lewis [1960] 1996, pp. 30–31)

What this means is that even if developments in AI somehow overcome our first problem, by showing that AI systems can have insight into truth, still this will not by itself explain how these systems could have insight into *necessary* truths. Moreover, as I will now argue, the fact that conscious rational beings can access universal concepts also helps to explain why they are capable of a different order of creativity than AI or other adaptive physical systems.

3.5. Non-Combinatorial Creativity

'Creativity' is a rather ambiguous term, with several distinct meanings. If all we mean by it is that something new is produced, then natural events like meteor impacts and freak accidents are 'creative'. If we further mean that something coherent, meaningful, and at least partially distinct from any prior examples is produced, then certainly LLMs and other AI systems are 'creative'. These systems have produced 'new' essays, plays, poems, songs, music, and art. However, one can generally trace the product to a recombination of works originally produced by intelligent human beings. To be sure, human beings also are often 'creative' at this level, for example by combining genres of art, music, and films, architectural styles or 'fusion' cuisine. And at times, human beings are spectacularly uncreative: even before AI, they plagiarized works and board rooms are frequently ensnared by recycled phrases, e.g., "The elephant in the room". However, there is reason to think that conscious rational beings are also capable of higher levels of creativity that are inaccessible to AI and other adaptive physical systems.

As previously noted, adaptive physical systems are limited in their powers of discovery to past interactions with the world, and this suggests that the best they can do is some form of *combinatorial creativity*, that creates a new work by extrapolating from or combining works in their training history. That means that they cannot by themselves find a radically new item that is not derivative from patterns already instantiated in the actual world or

their actual training data. If they sometimes seem to do so, that will be traceable to the infusion of new information from a creative human being, and so it cannot be credited to the adaptive physical system.

Part of what makes some human discoveries so novel is that they are not merely extrapolations or combinations of extant patterns but involve a leap in conceptual space that allows us to envisage something that has not been instantiated before. Some good examples of this are the creative insights of Leonardo da Vinci (1452–1519), Jules Verne (1828–1905), Charles Babbage (1791–1871) and Alan Turing (1912–1954).

Some years ago, my wife and I visited the IBM da Vinci museum at Clos Lucé, in Amboise, France, which features forty models of da Vinci's ingenious ideas for new technology. One of these is a model of an 'aerial screw', an early if impractical design for a helicopter. When da Vinci conceived this idea, there were machines, and there were birds and other flying creatures, but there were no flying machines as precedents. Da Vinci had to make a leap in conceptual space: if there are flying creatures, and there are machines, perhaps there can be flying machines. Similarly, Jules Verne conceived the idea of a propellor-driven airship, the Albatross, in 1886, when there were only ocean-going ships and hot air balloons and no powered aircraft. While the design of Verne's 'Albatross' was not fully practical, it can be seen as a precursor to the first powered aircraft, the 1903 Wright Flyer, as well as to modern helicopters. The key insight was to see that the concept 'ship' need not be confined to ocean going vessels, but could also apply to other transportation mediums like air.

Arguably, an even bigger conceptual leap was made in conceiving the idea of the modern digital computer. Before the work of Charles Babbage, there were only mechanical calculators, such as those developed by Wilhelm Schickard (1592–1635), Blaise Pascal (1623–1662), and Gottfried Leibniz (1646–1716), that could perform simple arithmetical computations. As brilliant as these inventions were, Babbage went further: though never fully constructed, his Analytical Engine was conceived to be programmable using punched cards, which already suggested the idea of a multi-purpose machine. Instead of a dedicated machine like a clock which is limited to one main purpose, one could change the function of the Analytical Engine by changing its program, so that it could do the work of a multiplicity of dedicated devices. Alan Turing went further, developing the mathematical idea of a Universal Turing Machine, that could emulate any other computer and was therefore capable of a potentially infinite number of distinct tasks.

In her study of different kinds of creativity, Boden (2004), notes two other forms of creativity beyond the combinatorial creativity of which AI is certainly capable. In "exploratory creativity", there is an investigation of the logical space of possibilities, and Boden believes that AI can at least model some forms of exploratory creativity, and find "new" solutions. The success of the AI system AlphaFold2 in solving an important part of the fundamental protein-folding problem (how to predict the three-dimensional shape of a protein from its one-dimensional code) is a good example (Saplakoglu 2024). There is much still to be done, and the system does not explain how the protein-folding process actually works, but here AI found an answer that had eluded human researchers.

However, impressive as this feat is, it seems different in kind from the conceptual leaps made by da Vinci, Verne, Babbage, and Turing. Da Vinci and Verne conceived a mechanical flying machine with no examples to work on; Babbage and Turing had to jump from dedicated machines to a general-purpose problem-solver. This is different from exploring an already given problem space with well-defined concepts. One has first to form a new concept which has, as yet, no precursors.

The greatest feats of human creativity go further still. They are examples of what Boden calls "transformational creativity", in which one must change the foundational assumptions that govern a given conceptual space to create an entirely new own. In

Newtonian physics, it is assumed that space is Euclidean and therefore obeys the axiom of parallels. But in reconceptualizing gravity as a distortion of space-time, Albert Einstein (1879–1955) suggested that space-time is curved. From this it follows that parallel lines can meet (on a sphere, for example) and that light can bend around massive objects, something experimentally confirmed in 1919 by Sir Arthur Eddington, during a solar eclipse.

Similarly, Einstein's close friend at Princeton, Kurt Gödel (1906–1978) ended the ambitious program of David Hilbert (1862–1943) which aimed to show that mathematical "truth" could be reduced to what could be proved mechanically from a finite set of axioms. To do this, Gödel had to question what seemed like an obvious and inviolable distinction between statements *of* arithmetic (like '2 + 2 = 4'), and statements *about* arithmetic (like 'Peano arithmetic is consistent'). By the ingenious device of arithmetizing syntax, Gödel showed that arithmetic itself could speak about what is provable in arithmetic, and then constructed a sentence which is true, if, and only if, it is unprovable. He was able to show that any effectively axiomatizable system rich enough to express the language of arithmetic would have such a sentence, and so, if these systems are sound, they must be incomplete: there are statements that are true of the system which the system cannot prove (Smith 2013, especially chapters 19–22).

No extrapolation from classical Newtonian physics discovers Einstein's insights. No extrapolation from ordinary arithmetic finds Gödel's incompleteness theorems. To discover these truths, both men had to change the assumptions that governed the conceptual space. That means that making progress involved claiming something that was literally impossible in the original conceptual space. So long as space is Euclidean, parallel lines cannot meet; so long as we believe that arithmetic cannot talk about itself, Gödel's results cannot be found.

Moreover, there are two reasons to think that AI is not likely to achieve these feats of transformational creativity. The first is an empirical result known as "model collapse" which strongly suggests that AI is much less creative than it seems. Standardly, LLMs are trained on carefully curated data, such as high-quality encyclopedias or academic journals, which were originally generated by human researchers. However, it is possible to "close the loop", and have the LLM recycle its own output as input with no further human-generated content. When this is done, the LLM suffers a kind of informational entropy, so that its subsequent output quickly degrades in coherence. In a famous example (Shumailov et al. 2024), a LLM trained on data about English architecture was then allowed to recycle its data and, by the ninth generation was thoroughly contaminated by the repeated term, "jackrabbits". The fact that the system could not see that its output was becoming increasingly nonsensical reinforces our earlier point that LLMs and other AI systems do not understand the sense and reference of the terms they use. But it also suggests that LLMs are not original *sources* of coherent information, but only *conduits* of it, as they remain dependent on infusions of high quality input from human beings to avoid a catastrophic loss of coherence.

The second reason for doubting the creative power of AI returns to our earlier point about the human ability to abstract concepts. LLMs and adaptive physical systems in general can only learn by interacting with particular physical examples in their environment, such as particular green objects. But conscious rational beings can abstract the universal concept, greenness. And once they have this, and other concepts, they are free to consider instantiations of the property green which they have never experienced. J. R. R. Tolkien saw this power of abstraction as fundamental to the sub-creative power of human beings and a sign of their being specially made in the image of God.

> When we can take green from grass, blue from heaven, and red from blood, we have already an enchanter's power. . .. We may put a deadly green upon a man's face and produce a horror; we may make the rare and terrible blue moon to shine; or we may cause woods to spring with silver leaves and rams to wear fleeces of

> gold, and put hot fire into the belly of a cold worm. But in such 'fantasy' as it is called, new form is made. . .. Man becomes a sub-creator.
>
> (Tolkien 2001, pp. 22–23)

It is because conscious rational beings can access universal concepts that they can think of possibilities they have never experienced. What if one machine could emulate virtually any other machine? What if light could bend, or time could slow down? What if arithmetic could talk about itself? What if magic were possible? We see in the development of entire imaginary worlds, like Tolkien's own Middle-earth or C. S. Lewis's Narnia, that humans can conceive of richly populated alternative possible worlds, worlds which can be governed by fundamentally different principles than our actual world. This makes possible not only appealing science fiction and fantasy but also the *Gedankenexperiments* of physicists and philosophers.

To be sure, AI can be trained on the work of creative writers and generate responses that seems to fit their imaginary worlds. But this ability is parasitic on past human feats of imagination and creativity. A real test of our AI systems would be to see whether, given only the input they can physically derive from the actual world (equivalent to what we acquire through the senses), they could ever generate distinct, coherent, alternative worlds of their own. My conjecture is that they would fail this test, as without the ability to acquire universal concepts, and to step back and question basic assumptions, they cannot make creative leaps to distinct conceptual spaces, governed by fundamentally different assumptions than those with which they have interacted.

4. Conclusions

Credit must be given to the tremendous advances of AI. But, if the argument of this paper is correct, there are at least five powers of the human soul that are beyond current LLMs and which are likely beyond any adaptive physical system in principle. These systems do not seem capable of insight into truth, meta-insight into what they are doing, free will, access to necessary conceptual relations, or the kind of non-combinatorial creativity manifested by the greatest achievements of human thinkers. If this is right, anthropic mechanism is false. We are not merely organic realizations of Turing machines, but have an immaterial soul which can transcend the limitations set by our purely physical interactions with the environment. Due to their souls, human beings are fundamentally different kinds of beings than machines.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The author declares no conflict of interest.

References

Boden, Margaret. 2004. *The Creative Mind: Iths and Mechanisms*. New York: Routledge.

Boolos, George S., John P. Burgess, and Richard C. Jeffrey. 2002. *Computability and Logic*, 4th ed. Cambridge: Cambridge University Press.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and et al. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv* arXiv:2303.12712. [CrossRef]

Chalmers, David. 1996. *The Conscious Mind*. New York: Oxford University Press.

de la Mettrie, Julien Onfray. 1943. *Man a Machine*. La Salle: Open Court.
Goetz, Stewart, and Charles Taliaferro. 2011. *A Brief History of the Soul*. Malden: Wiley-Blackwell.
Hobbes, Thomas. 2008. *Leviathan*. Edited by Marshall Missner. New York: Pearson Longman. First published 1651.
Kant, Immanuel. 1982. *Critique of Pure Reason*. Translated by Norman Kemp Smith. London: The Macmillan Press, Ltd. First published 1781.
Kurzweil, Ray. 2024. *The Singularity Is Nearer: When We Merge with AI*. New York: Viking.
Lewis, Clive Staples. 1996. *Miracles: A Preliminary Study*, 2nd ed. New York: HarperCollins. First published 1960.
Lucas, John Randolph. 1961. Minds, Machines and Gödel. *Philosophy* 36: 112–27. [CrossRef]
Menuge, Angus. 2004. *Agents Under Fire: Materialism and the Rationality of Science*. Lanham: Rowman and Littlefield.
Menuge, Angus. 2009. Is Downward Causation Possible? How the Mind can make a Physical Difference. *Philosophia Christi* 11: 93–110. [CrossRef]
Menuge, Angus. 2016. Knowledge of Abstracta: A Challenge to Materialism. *Philosophia Christi* 18: 7–27. [CrossRef]
Menuge, Angus. 2018. Why Reject Christian Physicalism? In *The Blackwell Companion to Substance Dualism*. Edited by Jonathan J. Loose, Angus J. L. Menuge and J. P. Moreland. Oxford: Wiley Blackwell, pp. 394–410.
Moreland, James Porter. 2009. *The Recalcitrant Imago Dei*. London: SCM Press.
Oschner, Kevin N., Silvia A. Bunge, James A. Gross, and John D. E. Gabrieli. 2002. Re-thinking feelings: And fMRI study of the cognitive regulation of emotion. *Journal of Cognitive Neuroscience* 14: 1215–29.
Plato. 1985. Meno. In *The Collected Dialogues of Plato*. Edited by Edith Hamilton and Huntington Cairns. Princeton: Princeton University Press, pp. 353–84.
Rescorla, Michael. 2024. The Computational Theory of Mind. *Stanford Encyclopedia of Philosophy*. Available online: https://plato.stanford.edu/entries/computational-mind/ (accessed on 1 October 2025).
Robinson, Daniel. 2011. Minds, Brains, and Brains in Vats. In *The Soul Hypothesis: Investigations into the Existence of the Soul*. Edited by Mark C. Baker and Stewart Goetz. New York: Continuum, pp. 56–67.
Saplakoglu, Yasemin. 2024. How AI Revolutionized Protein Science, But Didn't End it. *Quanta Magazine*. June 26. Available online: https://www.quantamagazine.org/how-ai-revolutionized-protein-science-but-didnt-end-it-20240626/ (accessed on 20 November 2025).
Schwartz, Jeffrey, and Sharon Begley. 2002. *The Mind and the Brain: Neuroplasticity and the Power of Mental Force*. San Francisco: Harper.
Searle, John. 1980. Minds, Brains, and Programs. *Behavioral and Brain Sciences* 3: 417–57. [CrossRef]
Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature* 631: 755–59. [CrossRef] [PubMed]
Smith, Peter. 2013. *An Introduction to Gödel's Theorems*, 2nd ed. Cambridge: Cambridge University Press.
Tolkien, John Ronald Reuel. 2001. On Fairy Stories. In *Tree and Leaf*. London: HarperCollins.
Wolfram, Stephen. 2023. *What Is ChatGPT Doing. . .and Why Does It Work?* Champaign: Wolfram Media Inc.