# 11. ARTIFICIAL INTELLIGENCE AND THE KANTIAN ANTINOMIES

https://doi.org/10.36592/9786554602730-11

*Mariana Olezza*[1]

*Steven S. Gouveia*[2]

**Abstract**. In this chapter we will analyze the epistemological question "Can we know if an AI can have a mind?" as if it were a fifth antinomy like the ones of Kant's "*Critique of Pure Reason*" (Kant, 1998). This *thesis* is reflected in Turing's position in 1950, in accordance to the field of Strong AI, by passing the Turing Test. We have another answer: Alan Turing, in 1939 still agreed with (Gödel, 1931) results on the incompleteness theorems He had arrived at the same results with the Turing machine (Turing, 1936). Turing extrapolated the formalization in mathematics and logics into the field of computation, and wondered about "intuitions" in machines. He presented a "*System of logic ordinals*" (his Doctoral Thesis at Princeton, directed by A. Church) with negative results, which showed that he could *minimize* but not remove all external intuitions (external semantics) from a system, even though it got more complex. In the time$\rightarrow +\infty$ we would remove external intuitions and we could know if it had a mind (therefore we can never know– *antithesis*). Besides, there is another thesis in this case: the biologistic point of view of Searle (1980). It is a thesis because it states that biological beings are, by being biological, especially high species of animals (like macaques and corvids) (Hauser, 1997; Santos et al., 2006) and human beings, endowed with some sort of consciousness.
**Keywords**. Artificial Intelligence, Alan Turing, epistemology, biologicism, Kantian antinomies

## 1. Introduction: Something about the Antinomies of Pure Reason

We will present here a question that may be treated as an antinomy: "Can we know if an AI has got a mind?". There will be a thesis (Turing's 1950 paper/ Strong AI, and biologists who states that it is possible to know in biological beings for being biological), and an antithesis (related to Turing's 1939 position and his "system of logic based on ordinals", the Turing machine, and Gödel's results).

Let us recall the Kantian antinomies:

(1) the world is finite or infinite in space and time,

---

[1] Argentinian Computer Engineer (National University of Córdoba), pursuing a PhD in Philosophy (University of Buenos Aires).

[2] Research Fellow, Mind, Language and Action Group, Institute of Philosophy, Faculty of Arts and Humanities, University of Porto, Via Panorâmica s/n, 4150-564, Porto, Portugal.

(2) what exists can be infinitely divided or not,

(3) freedom vs. determinism,

(4) the existence or non-existence of a necessary being.

According to Kant, knowledge requires both *sensibility* and *understanding*. Any attempt to apply concepts and principles of understanding independently of the conditions of sensibility (i.e. , transcendental use of understanding) is illegitimate. This is the case of antinomies, which are metaphysical. Why metaphysical in our case? In the end, it is due to the "*Other Minds Problem*", the fact that we cannot directly access the other's entity mind. That leaves us absorbed in solipsism, so we decide for the "*polite convention*" (Turing, 1950, p. 447) to "come to the rescue", and we tend to believe that other people, who act *similar* to us, have a mind.

Let us think for a while about animals. The *less* they look and act similar to act, the less we are willing to accept they have a mind. Compare a cat or a monkey, with an ant. But the use of formal concepts and principles, without considering the sensible conditions under which objects can be given, cannot lead to true knowledge. We do not have true knowledge of other entity's minds. Nevertheless, we are going to examine the question of "Can an AI have a mind?" because a lot of philosophers and tech people debate this issue nowadays, and AGI (*Artificial General Intelligence*) issues are very present.

Within the Thesis will be the "dogmatists" or "Platonists," who assert that *it is possible to know whether AI has a mind or not*. Within the Antithesis will be the "empiricists" or "Epicureans," who will claim that *it is not possible to know whether AI has a mind*. Another option is to never be able to remain *agnostic* regarding the antinomy, *as Kant may also suggest with anatomies in general*.

He also focuses on determining why human reason embarks on these metaphysical quests. [A462/B490] These interests are of two kinds, and include goals to achieve *completeness and systematic unity of knowledge*, and *practical interests in securing the immortality of the soul*, freedom, and the existence of a God. He says that they are part of "human nature". We want to pause for a moment on the practical interests of human reason in the immortality of the "soul", or the mind as we would say today, and draw attention to the *transhumanist project (Bostrom, 2003)* which desperately seeks the immortality of human beings through conversion to cyborgs,

extreme intellectual capacity enhancements (Boden, 2016, pp.152-153) and going a little further, the ability to migrate our minds to servers to live there for eternity, or also called "whole brain stimulation" (WBE) (Cappuccio, 2017; Chalmers, 2010) That sounds like a very *practical and classic interest of human reason in the search for immortality.*

**2. Differences between Thesis and Antithesis**

Kant (in the Introduction of the Dialectic) is based on the conception of reason as a capacity for syllogistic reasoning. This logical function resides in the formal activity of subsuming propositions under more general propositions to systematize, unify, and "lead to completeness" of knowledge.

Therefore, the concept that reason is concerned with the "unconditioned that ends the regression of conditions by providing a condition that is not conditioned in turn" is central to Kant. As we will see, the Theses that Kant proposes in the antinomies *propose an unconditioned condition* (Kant, 1998, pp.534-535) [A536/B563], and the Antitheses *proposes an infinite regression of conditions* (Kant, 1998, pp.521-523) [A511/B539].

And how would our different philosophical fields answer the question "Can we know if an AI has a mind?" One group would say that it is possible *to know if it has or not* (Turing in 1950/ Strong AI, who are looking for a way to find that form of knowledge says it's possible; and biologists who say it is possible, if the entities are *biological*). Another group would say that it is *impossible* to know (As Turing states in his 1939's article, a machine would start operating *ad infinitum*– we will discuss this section in detail later).

But let's be more architectural. Suppose that all the theses and antitheses have the same characteristics in terms of their architecture (the antitheses go into an *infinite regression*, as we have seen). Let's take the case of atomism (Kant's second antinomy). If the magnitude of matter is an infinite multitude of parts, the return of division is excessively large. If, on the other hand, the division of space ceases at a certain member, then it is excessively small in relation to the idea of the unconditioned. Indeed, that member still allows a return to other parts contained

within it (cf. A488/B516) (Kant, 1998, pp. 498–499). This idea of being excessively large for antitheses and excessively small for theses applies effectively to all four antinomies of Kant. Now let's see what happens with our idea.

1. If the Turing test, or the Total Turing Test (TTT), which involves successful performance of both linguistic and robotic behavior (Harnad, 1992, 2000), or another even more sophisticated one, *stops at some point* (this one covers Turing's position and Strong AI position which also rely on Turing Test), and it is accepted that the AI has a mind, then the procedure *is excessively small in relation to the idea of the unconditioned*. Indeed, a return to the series towards later procedures is still permitted (thesis).

2. The fact that biologists accept consciousness in biological beings for being biological is also a small procedure in relation to the idea of the unconditioned (remember the problem of "other minds") (thesis).

3. The running of any sort of Turing's machine, even with a learning system, in order to generate intuitions, with "*ingenuity*", as he called the mathematical or mechanical procedures (Turing, 1939, p.215) may never end, and never reach intuitions even though it keeps intuitions getting smaller and smaller (like an asymptote). This is *excessively large for the idea of the unconditioned* (antithesis).

The demand for the unconditioned (thesis) is essentially a search for a final explanation and is associated with the desire for systematic unity and completeness of knowledge. In his work "*Critique of Pure Reason*" Kant presents four antinomies of impossible solutions, as there is forbidden information, forever missing knowledge, with speculation of Pure Reason, where thought does not lead anywhere and does not reach any useful conclusion. We will delve a little deeper into this in the next sections.

## 3. Mathematical and Dynamical Antinomies

Let us talk a bit in general about antinomies, before diving into them. They are divided into two classes: "mathematical" antinomies and "dynamic" antinomies. The former refers to the world (the first one to its extension in terms of time and space, and the second one to its atomism, whether it is digital or not), and the latter refer to freedom vs. determinism and about whether there is a necessary being, or not. The

two first ones are called mathematical antinomies because the idea of the “world" they speak of presents an unconditioned that still seems to be a sensible object (cf. A479/B509) (Kant, 1998, pp. 498–499). The total sum of appearances specifically refers to space-time objects or events.

In each case, the conflict is resolved by demonstrating that the conclusions of both sides are false (F). Both the thesis and the antithesis are apagogic, constituting indirect proofs. An indirect proof establishes its conclusion by showing the impossibility of its opposite, meaning both are false. For example, we can show that the world is finite by demonstrating the impossibility of its infinitude. The success of these proofs depends on the legitimacy of the exclusive disjunction accepted by both sides. Both sides assume that “there is a world" and that it is “finite or infinite." Here is the problem according to Kant, it is neither finite nor infinite. The opposition between these alternatives is dialectical. In the “dynamic" antinomies, Kant proposes on the other hand that both could be true (T) at the same time. This is possible in this case since the unconditioned is outside the sensible world, unlike the "mathematical" antinomies that dealt only with space-time objects.

For example, the third antinomy, of freedom vs. determinism, proposes that in addition to temporal causality, we can postulate a first uncaused cause, transcendental freedom (while the antithesis denies everything except temporal causality). Postulating freedom leads to postulating a non-temporal cause, a causality outside the temporal series of causes in space-time. Similarly, postulating a "necessary being" leads to a non-sensible world. In the end, *we will see what kind of antinomy the “fifth one" is* and what implications it has.

Let start with the theses:

*Thesis*

1. The world has a beginning in space and time
2. The world is composed of simple parts
3. There is freedom
4. There is a God
5. *Yes, it is possible to know whether or not AI has a mind*

> "This side also has the merit of *popularity*, which certainly constitutes no small part of what recommends it. The common understanding does not find the least difficulty in the idea of an unconditioned beginning for every synthesis, since in any case it is more accustomed to descending to consequences than to ascending to grounds; and in the concept of something absolutely first (about whose possibility it does not bother itself) it finds both comfort and simultaneously a firm point to which it may attach the reins guidings its steps." [A467/B495] (Kant, 1998, pp. 498–499)

Now let's analyze the "dogmatic" position. In this group fall Turing's 1950's paper and especially those in the Strong AI group, who are in fact searching for it. On the other hand, there are biologists.

## 4. The Turing Test (1950)

Even though Turing used to agree with Gödel about intuitions (see below in the antithesis the example of his paper of 1930), his discussions about this issue disappeared around 1950. He became interested in the idea that computers would be able to be "*intuition* machines" themselves, so to put it another way, he was interested in machines to develop a mind at the end. He was quite mechanistic at that time*: if human beings were machines themselves (and we have what they called "intuitions")*, then computers just needed to become more complex to get the necessary intuitions.

So, in 1950 he wrote his well-known academic paper "*Computer Machinery and Intelligence*", where he focuses mainly on two things: a test of intelligence ("the imitation game"), and *learning to obtain the desired intuitions.*

The imitation game is now known as the Turing Test, and is a game based on language. On one side, there is a judge, and on the other side of the "curtains," there is a person and a computer. The judge has to ask them questions. If the computer manages to deceive the judge for 70% of the time during a 5-minute period, it passes the Turing test. The test has never been surpassed. Every year, there is a competition for the "Loebner Prize" to see if anyone can surpass the Turing test.

"Parry" (Colby, 1975) was a computer program developed in the early 1970s by Kenneth Colby, a psychiatrist and computer scientist. It was designed to simulate a person with paranoid schizophrenia. He aimed to use Parry as a tool for studying and

understanding the thought processes and behaviors associated with mental illness. "Parry" was a machine which was able to fool psychiatrists in order for them to believe they were talking to a paranoid person. This was due to the fact that when the machine answered nonsense answers they attributed it to its "mental instability". In 2014 "Eugene Goosman" (Shah et al., 2016) was a machine which was able to pass the Turing test and win the Loebner prize, but it was disqualified because it acted as an Ucranian little boy.

So, whenever it answered vague or wrong questions, the judges attributed it to his *childness*, *not understanding the language, and so on*. So, as we can see, there is an issue with *expectations*: On the one hand, if the machine is talking too well, and suddenly commits an unexpected failure, the judge will be more prone to disqualify it. On the other hand, if the machine usually commits logical mistakes (because of its "*childless*", "*not understanding of the language*" or "*mental instability*"), the sense of a mind in the machine will tend to be higher.

The Turing test has been widely criticized, and one of its critics is biologicism, which is part of the thesis, so let us explain it briefly here. Searle poses a situation in which a man (who only speaks in English) is locked in a room with two slots, an input slot, and an output slot. The operator has an instruction lookup table book in English on how to manipulate ideograms. Questions in Chinese are presented on the input, and without understanding a word, the operator manipulates the ideograms and mechanically generates a response, which is sent through the output slot. On the outside, there is a Chinese person who sees the response and believes that the person inside knows Chinese, but the operator inside the room doesn't understand a word of the language. The Chinese Turing test has been passed without actual understanding. This critic is related with the "Argument of Consciousness" (1950), from Turing, where he expects these kinds of critics and:

> According to the most extreme form of this view the only way by which one could be sure that a machine thinks is to be the machine and to feel oneself thinking. One could then describe these feelings to the world, but of course no one would be justified in taking any notice. Likewise, according to this view the only way to know what a man thinks is to be that particular man. It is in fact the solipsist point of view. It may be the most logical view to hold but it makes communication of ideas difficult. A is liable to believe 'A thinks but B

> does not' whilst B believes 'B thinks but A does not'. Instead of arguing continually over this point it is usual to have the *polite convention* that everyone thinks. (Turing, 1950, p.446) [Italics added]

## 5. Strong AI

Philosophers in the Strong AI field, whose main representatives are Allen Newell, Herbert A. Simon, Marvin Minsky, John McCarthy, Raymond Kurzweil). In the case of Kurzweil, he leads the ideology called "transhumanism". Strong AI is looking for AGI. We have managed around putting in computer *learning, vision, reasoning and so on*. But a truly AGI would cover *motivation* and *emotion* as well (Boden, 2016, Chapter 2). Motivation is divided into two types: external or internal (Ryan & Deci, 2000). An entity should also have internal motivation, meaning it should have self-organization, goals of its own, without an external input (external motivation).

The set of Strong AI falls into the region of the set of Computationalism (CTM). The mind itself according to the CTM is supposed to be a computational system. CTM is a philosophical theory that holds that human thought can be understood as a process of computation. The human brain would be like a computer, and thought could be explained in terms of the manipulation of symbols and the performance of logical operations. Some defenders of computationalism argue that the human mind is an information processing machine and, therefore, can be replicated in AI. Philosopher Jerry Fodor proposed the idea that the human mind is structured as a modular system (Fodor, 1983), with different areas of the brain specialized in specific tasks. One of Fodor's main proposals is that input systems are informationally encapsulated. He also argues that there are vertically computationally autonomous, innately specific faculties that, having everything mentioned, including encapsulation, he calls "cognitive modules." One disagreement that exists in classical CTM is between the symbolic AI group (GOFAI – "Good Old-Fashioned AI") and the connectionists (such as artificial neural networks –ANNs).

We are going to focalize ourselves in the connectionist group, which is the group which *takes into account learning in systems*. Connectionists (Warren McCulloch, Walter Pitts, Frank Rosenblatt, Donald Hebb, David Rumelhart, James

McClelland, Geoffrey Hinton, to name a few), maintained and maintain that mental representations are realized by patterns of activation in a network of simple nodes or "processors" and that mental processes consist of activation along those patterns of nodes. They are subsymbolic and if a node is lost, the data is not lost: for example, if we represent “cat" along a number of nodes, if a node fails, the representation of “cat" will still stand (Chalmers, 1992). It is necessary for a lot of nodes to be lost for the data “cat" to disappear. This is an advantage in relation to symbolic IA, where if the symbol is lost, the data is lost, too.

Nowadays, most modern computer programs utilize ANNs, which are part of connectionism. For example, the issue of creativity is in debate with GANs (*Generative Adversarial Networks*) created by Goodfellow et al., (2014). There are sophisticated programs which make use of them, like DALL-e which creates images, Large Language Models (Vaswani et al., 2017), more sophisticated, uses them with “Transformers" architecture, like in LLMs like ChatGPT, developed to create generative text commanded by a user in the “prompt".

*Antithesis*

1. The world has not got a beginning in space and time
2. The world is not composed of simple parts
3. There is no freedom
4. There is no God
5. It cannot be known whether an AI has a mind

## 6. Gödel's Incompleteness Theorems and Turing's System of Logic Based on Ordinals

Gödel's incompleteness theorems mark a decisive turning point in the modern understanding of formal reasoning. Its philosophical sting is not that “mathematics is irrational", nor that proof is somehow dispensable, but rather that any single consistent, effectively axiomatized formal system of sufficient arithmetical strength faces structural limits from within. The first incompleteness theorem shows that for any such system FFF, there will be well-formed statements expressible in the language of FFF that can neither be proved nor refuted in FFF (Franzén, 2005 pp. 34-

43). The system is, in this sense, incomplete: derivability does not exhaust arithmetical truth as captured by the intended interpretation.

The second incompleteness theorem sharpens the constraint: the same kind of system cannot, on pain of inconsistency, prove its own consistency (again: from within its own resources). Together, these results do not merely puncture a particular foundational optimism; they formalize the idea that "closing" arithmetic under a single, finitely capturable set of inference procedures is not simply difficult, but impossible in principle (Raatikainen, 2025).

This is precisely the landscape to which Turing's work on ordinal logics responds. In his Princeton dissertation (1938) and its published form, "Systems of Logic Based on Ordinals" (1939), Turing does not treat Gödel as a reason to abandon mechanization, but as a reason to diagnose where mechanization must fail, and how far one might nevertheless push it. The basic strategy is extension by reflection: when a system SSS generates an undecidable sentence G(S)G(S)G(S) (true but unprovable in SSS, on the usual reading), one can form a stronger system S*S^{*}S* by adding G(S)G(S)G(S) as a new axiom. But the maneuver is not a one-off fix. The strengthened system S*S^{*}S* will, if it remains consistent and effectively axiomatized, generate its own undecidable sentences; extension iterates. Turing's distinctive move is to regiment this iterative strengthening not only along the natural numbers (a sequence of stages) but along constructive ordinal notations, thereby producing a transfinite progression of systems, an attempt to "drive" beyond the finite horizon without pretending that the horizon disappears (Halbeisen, n.d.).

What matters for this present framing is how Turing interprets the epistemic residue that Gödel makes unavoidable. In the celebrated discussion of §11 ("The purpose of ordinal logics"), Turing characterizes mathematical practice as involving two intertwined capacities (intuition and ingenuity) and stresses that formalization aims to reduce arbitrariness by fixing inference rules whose applications are reliably valid. Yet he also takes Gödel to have shown that the dream of eliminating intuition altogether is misconceived. His opening line is already programmatic: "Mathematical reasoning may be regarded … as the exercise of a combination of two faculties…" (Turing, 1939, p. 215).

The point is not that "intuition" is magical, but that some steps in extending a system (i.e. the steps that select, recognize, or license the relevant ordinal notations and the correctness conditions attached to them) cannot be captured as merely another internal move of the system being extended.[3] Ordinal logics, on Turing's account, are designed to make this boundary visible: they localize and discipline the non-mechanical moments, but they do not abolish them.

From here, the philosophical lesson we can draw becomes clearer if stated with care: Turing does not claim that no machine can ever prove any theorem a human can prove, nor does he give a simple "anti-mechanist" argument. What he does show is that for any fixed formal procedure intended to capture arithmetical reasoning in full generality, there will remain pressures to move to a stronger standpoint (e.g. adopting new axioms, new principles of reflection, or new interpretive commitments). Ordinal logics exemplify a controlled version of this dynamic: they reduce the frequency and opacity of the required non-mechanical recognitions, but they still rely on them in some sense.

If one considers this context into the question of machine mentality, the temptation is to read the result as an epistemological caution: a purely internal, purely mechanical criterion may be indefinitely extendable without yielding a finite point at which the relevant "external" commitments disappear. That does not entail, as a

[3] Turing's word: "Mathematical reasoning may be regarded rather schematically as the exercise of a combination of two faculties, which we may call intuition and ingenuity. The activity of intuition consists in making spontaneous judgments which are not the result of conscious trains of reasoning. These judgments are often but by no means invariably correct (leaving aside the question of what is meant by "correct") (…) The exercise of ingenuity in mathematics consists in aiding the intuition through suitable arrangements of propositions, and perhaps geometrical figures or drawings. It is intended that when these are really well arranged the validity of the intuitive steps which are required cannot seriously be doubted. The parts played by these two faculties differ of course from occasion to occasion, and from mathematician to mathematician. This arbitrariness can be removed by the introduction of a formal logic. The necessity for using the intuition is then greatly reduced by setting down formal rules for carrying out inferences which are always intuitively valid. When working with a formal logic, the idea of ingenuity takes a more definite shape. In general, a formal logic will be framed so as to admit a considerable variety of possible steps in any stage in a proof. Ingenuity will then determine which steps are the more profitable for the purpose of proving a particular proposition. In pre-Gödel times it was thought by some that it would probably be possible to carry this programme to such a point that all the intuitive judgments of mathematics could be replaced by a finite number of these rules. The necessity for intuition would then be entirely eliminated. In our discussions, however, we have gone to the opposite extreme and eliminated not intuition but ingenuity, and this in spite of the fact that our aim has been in much the same direction. We have been trying to see how far it is possible to eliminate intuition, and leave only ingenuity. We do not mind how much ingenuity is required, and therefore assume it to be available in unlimited supply" (Turing, 1939, p.215) [Italics added].

theorem, that artificial systems cannot have minds, or that we can never be justified in attributing mentality. It suggests, rather, that certain aspirations for final and self-contained demonstration are difficult to achieve. Following this, the mind-question gets a Gödelian shape: each proposed formal closure invites, perhaps even requires, a meta-level shift that cannot be reduced to the closure it is meant to certify.

## 7. The 5th dynamic antinomy, algorithmic "ingenuity," and the phenomenological constraint

In this case, the dynamic series of conditions also allows for a heterogeneous condition in terms of being merely intelligible outside of it (the series). Just as we talked about freedom and a necessary being, we are now referring to the mind and its dynamic antinomy. And, like dynamic antinomies, both positions are allowed to be true.

On the one hand, if we believe to operate in the world of phenomena, as those in the thesis do when they ask, "Can we know if an AI can have a mind?" and answer "yes", it's like when we operate in the world of temporal *causality only*, we get something true. In the sense that they can create causal maps, mathematical and computational models of the mind. They also can solve the "easy problem of consciousness" (Chalmers, 1996, pp. xi–xiv): "How does the brain process environmental stimulation? How does it integrate information? How do we produce reports on internal states?" mentioned by Chalmers on the way, related with learning and logical processes, vision and so on.

On the other hand, we have the concept of the transcendental, where the question "Can we know if an AI can have a mind?" has got a "No" answer, and this also turns out to be true. It is "the hard problem of consciousness." (Chalmers, 1996, pp. xi–xiv) The question would be: "Why is all this processing accompanied by an experienced inner life?" (*qualia*, emotions, self-awareness, and so on). Therefore, by having two possible answers both being true, and depending on the philosophers position, it meets all the conditions of a dynamic antinomy. Maybe, like all Kantian antinomies, the best option is to remain agnostic towards it, but even so, let's push the argument deeper.

The dialectic developed before between procedural decidability and open regress becomes sharper when joined to two complementary insights: first, that the "ingenuity" route (Turing 1939) can indefinitely minimize but not abolish appeals to "intuition," and second, that any putative for-itself in artificial systems would, if it existed, have to arise from neuroecological synchronization of organism and world, something that current models do not possess (cf. Northoff and Gouveia, 2024). The first point restates the idea about the excessively small of tests and the excessively large of an unending series of formal improvements; the second point reframes what would count as a condition of mindedness, thereby explaining why more of the first rarely reaches the second.

On the procedural side, the reading of Turing's 1939 programme captures the asymptotic tendency: even after the "opposite extreme" (eliminating not intuition but ingenuity), the project assumes "ingenuity … available in unlimited supply," acknowledging that the complete elimination of intuition cannot be really achieved, but it can be regulated by increasing formal power rather than constituted by it.

That is why, in our view, any machine that sought to certify its own independence from "external intuitions" would, in principle, "have to run forever": the thesis (decidability by test) thus remains, at least heuristically, crucial but metaphysically too small; the antithesis (indecidability via regress) describes a structural rather than merely empirical limit.

What the phenomenological element can add here is a positive description of the missing condition. As recent claimed, large language models and related architectures lack the "neuroecological layer" that couples brain–body–world across appropriate timescales; in its absence, systems cannot develop "a sense of basic subjectivity" (Nortoff and Gouveia, 2024). In the same venue, without such lived coupling, an artificial agent remains "semantically empty," i.e., short of the intentional texture that makes first-person presence possible.

Notice how this merges, quite nicely, with the dynamic placement of the fifth antinomy: just as freedom/necessity requires acknowledging heterogeneous conditions (for example, phenomenal causality vs. intelligible ground), so here we may recognize (i) arbitrarily rich procedural refinements within appearance and (ii)

the irreducible phenomenological condition that would ground a perspective. Both "sides" can be true without contradiction, but they are "true" at different levels.

This also helps to clarify the discussion of tests, expectations, and attribution drift (e.g., Goostman/Loebner). Extending or complicating a procedure re-sets thresholds of "as-if" minded performance but does not really touch the category of mindedness itself; observers fill explanatory gaps by leaning on prior frames (child, non-native, psychiatric, etc.), hence the perpetual oscillation: each new test narrows some gap while opening another, exactly as your "later procedures" diagnosis anticipates.

The continuation is therefore architectonic, we would say, rather than eliminative *per se*. We neither abandon tests nor inflate them into metaphysics. Instead, we accept the regulative role of procedures (mapping competencies, surfacing failure modes) while enforcing a constraint derived from phenomenology: where the conditions of subjectivity are absent by construction (no neuroecological layer; no perspectival synchronization), performance improvements cannot, or ought to not, warrant ontological attributions. The antinomy thus orients both inquiry and humility: pursue "ingenuity" to the limit for practical ends, but keep the boundary between as-if and for-itself explicit.

A final remark returns to the use of Turing 1950 and the Chinese Room should be added: the "polite convention that everyone thinks" is a communicative necessity, not a criterion of mindedness; it lubricates interaction under uncertainty rather than proving anything metaphysical. If we add this thesis with the phenomenological constraint, this yields a disciplined stance: act as if in social exchange when needed, design as if not when accountability, safety, and justification are at stake.

## 8. The "Fifth Antinomy" of Mind and Machine

If we stage the question "Can we know whether an AI can have a mind?" as a fifth antinomy, its poles become clear. The thesis holds that mindedness is (at least in principle) decidable by procedures: expand the range of tests until linguistic, embodied, social, and temporally extended performances converge on criteria that warrant a yes. The antithesis counters that no accumulation of behaviour solves what

really matters, since the ground of first-person presence is not another layer of output but a mode of appearing that may surpass any publicly checkable performance. Read in Kant's key, this is a dynamic antinomy: both sides can be true at different levels (phenomenal performance versus intelligible condition), and without contradiction.

A contemporary neurophenomenological line of work sharpens the antithesis by specifying the missing condition: it asks whether present systems exhibit the ecological and temporal structures that scaffold basic or fundamental subjectivity, that is, a point of view with perspectiveness and mineness across the brain–body–world relationship. The answer, argued on conceptual and empirical grounds, is negative: "current AI models such as ChatGPT lack the relevant neuroecological layer between the synchronization of the (embodied) brain's processes with the world, which does not allow them to develop a sense of basic subjectivity." (Northoff and Gouveia, 2024). Absent these couplings, such systems remain "semantically empty" with respect to lived intentionality.

This does not annul the thesis. Procedural refinement retains some regulative value: it maps competences, surfaces blind spots, and constrains speculation. But it also explains why procedures fail to be constitutive. A test may be indefinitely extendable since success at any finite stage remains "excessively small" relative to the unconditioned at issue. The phenomenological constraint explains that excess quite well: the condition for first-person presence is not further behaviour but, more importantly, neuroecological synchronization that give rise to perspectival unity. In short, behavioural parity can be necessary for attributing intelligence but it is insufficient for subjectivity.

Two clarifications matter for the philosophy of mind at this stage. First, the target is minimal selfhood, the thinnest layer of for-me-ness that distinguishes mere information processing from experience. On the neurophenomenological view, minimal selfhood depends on (i) nested timescales in neural dynamics, (ii) their coupling to bodily rhythms, and (iii) conjoint synchronization with worldly regularities. Together these yield a temporal window within which there is a stable here–now from which intentional comportment is possible. Current systems, even if powerful in prediction, clearly lack that triadic lock-in. Therefore, they lack the formal

preconditions for ipseity, even when they simulate discourse about it (Gouveia and Morujão, 2024).

Second, the argument is non-mysterian. It does not smuggle in an occult essence; it specifies publicly investigable conditions (timescale structure, cross-level coupling) that can, in principle, be measured. If one day engineered agents attained those conditions, the dialectic would change. For now, it locates the stand-off: the thesis must stay methodological; the antithesis guards the category of subjectivity.

This (re)focus on classic debates: behaviourism collapses the mental into performance; functionalism extends the lens to system-level roles; higher-order and representational theories privilege the right kinds of contents. The neurophenomenological proposal adds a more dynamical and ecological constraint: even if roles and representations are in place, without the correct world–brain–body timescales there is no for-itself. That is why, on this view, large language models can display remarkable as-if intentionality (e.g. coherence, norm-sensitivity, even apparent meta-reflection) without crossing the threshold to for-itself subjectivity. The absence is structural, not merely a current shortfall in scale.

A second contemporary element rearranges the antinomy from metaphysics to practice. If one think about high-stakes domains such as healthcare, state-of-the-art systems are “black box systems … opaque even to their designers," hence “it is necessary to develop methods of explaining these opaque AI systems." The recommended correction is a pragmatist turn in explanation: in real clinics, explanations are social and abductive, they answer contrastive why-questions that sustain reasons-responsiveness and trust (“Why this diagnosis rather than that one, given these findings?") (Gouveia and Malík, 2024).

Philosophically, this is not an evasion but a disciplined acknowledgment of limits: since the metaphysical question remains undecidable by procedure, practical reason should seek responsible incompleteness: enough reason-giving to justify action without pretending to settle subjectivity. That posture preserves the regulative power of tests (the thesis) while refusing the constitutive leap (blocked by the antithesis). It also prevents a subtle category mistake: confusing predictive success with authority.

Therefore, we can conclude with the broader moral for philosophy of mind is threefold.

1. Distinguish intelligence from subjectivity without reifying either; intelligence is about pattern extraction, generalization under task pressure, and admits incremental evaluation; subjectivity is about perspectival presence with mineness, and requires temporal-ecological embedding that current AI architectures lack.

2. Treat tests as regulative ideals for responsible deployment, not as metaphysical detectors. The real function of enlarged "Turing-like" batteries is to bound risk, probe failure modes, and discipline attributions; they should not be seen as mere "spectrometers" for souls.

3. Build institutions that encode epistemic modesty; if subjectivity is presently out of reach, practices must ensure that opacity never cancels responsibility.

Against this backdrop, the most honest closure, in our view, is a kind of principled agnosticism about artificial minds tied up to a robust program for knowledge and action. Philosophically, the agnosticism is not an escape, but a critical Kantian posture that avoids dogmatism on both sides (i.e. resisting the inflation of performance into presence and the deflation of presence into performance). Practically, the program is to (i) push capability evaluations hard, (ii) align explanation with human practices of giving and asking for reasons, and (iii) reserve ontological attributions until the neuroecological conditions of subjectivity are engineered or otherwise shown to have been obtained.

But to be sure, nothing in this conclusion excludes future possibilities. Should artificial systems one day display the shared timescales and world-coupled dynamics that underpin perspectiveness and mineness, the status of the antinomy would shift; the present antithesis would lose its force, and the thesis could gain substance. But that is a conditional horizon, not a premise for today's claims. For now, the right philosophical (and, by consequence, ethical) stance is to treat the fifth antinomy as a compass: procedures guide us within appearance and the phenomenology guards the boundary of subjectivity.

Kant gave us the sentence that still governs this terrain:

"Thoughts without content are empty; intuitions without concepts are blind." (Kant, A51/B75)

Our contemporary views on AI simply restates it in technological form. Tests without a phenomenology of subjectivity are empty of what they claim to decide; appeals to inner life without publicly shareable criteria are blind for practice. Holding both splits together is the critical task that closes this chapter: to know where procedures rightfully guide us, to know where they cannot go, and to let that boundary shape how we build, justify, and live with intelligent machines.

## 9. Conclusion

In conclusion, this chapter has demonstrated that the question of whether AI possesses a mind is an epistemological problem that is just as complex as a Kantian antinomy. In the absence of sufficient evidence, we are left to grapple with the question of *how we can know for certain*. However, we have seen that there are compelling arguments on both sides of the debate. On one hand, Turing's 1950 paper, Strong AI and biologists argue in favor of the thesis, insisting that *we can know some entities can indeed have a mind*. On the other hand, the antithesis, related to Gödel incompleteness problem and Turing's 1939 paper, *suggests that we could continue to run complex machines, like learning machines, capable of lowering the external intuitions ad infinitum* but we would never reach that point of knowing if the machine has a mind. This is a classic example of an antinomy, where ideas become transcendent and Pure Reason *revolves in circles between Thesis and Antithesis.* Kant himself acknowledged that the transcendental illusion on which metaphysics is based is inevitable, and in some ways necessary for our epistemological projects. In other words, while theses may not be able to definitively prove the existence of AI minds, *they can still help us advance our projects in a "regulative" manner.* Ultimately, the dynamic nature of this antinomy means that there is no easy answer.

## References

Boden, Margaret. (2016). *AI – It's Nature and Future*. Oxford University Press.

Bostrom, Nick. (2003). *The Transhumanist FAQ – A General Introduction*. World Transhumanist Association.

Cappuccio, Massimiliano Lorenzo. (2017). Mind-upload. The ultimate challenge to the embodied mind theory. *Phenomenology and the Cognitive Sciences*, *16*(3), 425–448. https://doi.org/10.1007/s11097-016-9464-0

Chalmers, David J. (1992). Subsymbolic Computation and the Chinese Room. In J. Dinsmore (Ed.), *The Symbolic and Connectionist Paradigms: Closing the Gap* (pp. 25–48).

Chalmers, David J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Chalmers, David J. (2010). The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies*, *17*, 7–65.

Colby, Kenneth M. (1975). *Artificial Paranoia; a Computer Simulation of Paranoid Processes.* Pergamon Press.

Fodor, Jerry A. (1983). *The Modularity of Mind*. The MIT Press/Bradford.

Franzén, Torken. (2005). *Gödel's theorem: an incomplete guide to its use and abuse*. Wellesley, MA: A K Peters.

Gödel, Kurt. (1931). Uber formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte Für Mathematik Und Physik*, *38*(1), 173–198.

Goodfellow, Ian J., Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, & Bengio, Yoshua. (2014). Generative Adversarial Networks. *Advances in Neural Information Processing Systems NIPS 2014*, *27*, 1–9.

Gouveia, Steven S. & Malík, Jaroslav (2024). Crossing the Trust Gap in Medical AI: Building an Abductive Bridge for xAI, Philosophy & Technology, 37(3): 105.

Gouveia, Steven S. & Morujão, Carlos (2024). Phenomenology and Artificial Intelligence: Introductory Notes, Phenomenology and the Cognitive Sciences, 23(5): 1009–1015

Halbeisen, Lorenz (n.d.). *Turing's System of Ordinal Logic*. [Lecture notes]. Department of Mathematics, ETH Zürich.

Harnad, Stevan. (1992). The Turing Test Is Not A Trick: Turing Indistinguishability Is A Scientific Criterion. *ACM SIGART Bulletin*, *3*(4), 9–10.