***Tu Quoque*:**

**The Strong AI Challenge to Selfhood, Intentionality, Meaning,**

**And Some Artistic Responses[[1]](#endnote-1)**

**Erik C. Banks**

**Professor of Philosophy**

**Wright State University**



**Introduction: A “Negative Space” or No-Self Theory of Self**

The “headless body” picture above is from Ernst Mach’s 1886 *Analysis of Sensations* but it seems remarkably current to me. The picture depicts what Mach took to be the most realistic portrayal of first person experience and it does not include a self. There are sensations and motor activities but no enveloping consciousness. Nor is consciousness necessary for either sensing, perceiving or judging on Mach’s view since these are all functions that he considers more or less automatic and either conditioned by experience or by the evolutionary history of the organism. Rather than Hume, who had said similar things in the *Treatise of Human Nature*, Mach cites the German aphorist Georg Christof Lichtenberg as the inspiration for the view. Lichtenberg had claimed against Descartes that *cogito* should be replaced with the impersonal *Es denkt*; “there is thought.” And he rejected the inference from thought to thinker as a vestige of subject-predicate logic. Mach himself declared “Das Ich ist unrettbar,” the self cannot be saved, influencing William James and Bertrand Russell to reject the ego and embrace ego-less neutral monism (for which see my 2014). Mach’s pronouncement also reverberated in cultural circles in *fin de siècle* Vienna. Robert Musil’s *Der Mann ohne Eigenschaften* (literally the man without his *own* properties) seems directly inspired by Mach, as Ulrich the hero of the novel merely comments on the passing contents of his mind in a disinterested way as if he is not really there.[[2]](#endnote-2)

Mach’s doubts were fueled by the rise of physiological psychology in central Europe and his own work in perception. Much of the behavior of the sense organs which had been attributed to conscious influence or judgment (for example by Hermann von Helmholtz’s unconscious inferences) were reduced by Mach to automatic mechanisms not under any conscious control at all. It rather seemed to him the opposite, namely that the sophisticated sum total of mechanisms in perception had in fact given rise to the self-illusion in the first place.

With the advent of strong AI, and neuroscience, human consciousness has again come into question[[3]](#endnote-3), as have other traditional features of mind such as intentionality, roughly speaking “aboutness” or directness toward a content which underlies meaning and reference. The perennial No-Self Theory of Self has also experienced a minor renaissance in AI and functionalist circles.[[4]](#endnote-4) If a machine could master general intelligence, contextual determining tasks and language use for example, it would seem to lack nothing humans are capable of. These challenges seem to strike at the core of our defining features. Unlike weak AI, strong AI is defined as the idea that a machine or artificial organism can not only pass the Turing test and act in a way we would *describe* as conscious and meaningful in its use of language or symbols; rather its mastery will be such that it will either *be* conscious, in any sense of the word, or we will find out in reverse that these phenomena never existed in the first place.[[5]](#endnote-5) It is the second sense that I call the *tu quoque* (“you too!”) challenge of strong AI.[[6]](#endnote-6) Skepticism about the self, and or intentionality and meaning however has a rich history long before strong AI which I will seek to explore here. Clearly this will be an incomplete survey but I hope to hit some highlights that will illuminate the theme.

I start by defining the issue more precisely. Notice there is no expressed doubt about the existence of what philosophers call qualia, “colors and pains” for short.[[7]](#endnote-7) In the case of a machine, presumably if it had the right sort of electrochemical pattern instantiated, and the right sorts of physical dispositions, it too would have an internal manifestation of those powers as its own colors and pains, though those manifestations might be very different because the materials are different. But there would be, in Thomas Nagel’s phrase “something it is like” to be the machine. Consider for example the recent (2015) modeling of a piece of a rat’s somatosensory cortex with a supercomputer:[[8]](#endnote-8)



If these calculations were embodied in something physical as hardware, and the complex physical dispositions to pass from state to state were duplicated as well, it would be hard to argue that the simulation lacked sensation of some sort anyway, else what *is* the physical manifestation of the activity? Indeed Stanislaw Lem’s famous story “Non Serviam” is about such a simulation, called personetics; implemented in a machine, the personoids come complete with their own inner lives instantiated by the physical hardware on which the program runs.[[9]](#endnote-9)

There is also no doubt about the complex *functional* behavior of our brains, for which we can look to neuroscience, which characterize the complex physical dispositions of our nervous system to behave the way they do given stimulation, external or internal. It is important to include the dispositional behavior explicitly to block certain thought experiments in which we mathematically duplicate all the functional behavior of a brain in a machine made of ping pong balls or combinatorial tokens. Wittgenstein gives a famous example of one-to-one correlating the steps of a dance with the moves in a game of chess given the properly invented codebook. This can always be done but the disposition to behave a certain way at t and at the t+1th step will not be anticipated (again see my 2014, Chapter Five). In actual simulations the physical stuff in which the moves are instantiated provides the extra “oompf” to push the system from one state to another. And this could also be considered an instantiated program as well. We can call these concrete sensations and complex, dispositional functionality the *concrete* or *positive* phenomena of mind. I see no reason why these phenomena may not also be considered physical, in the extended sense of the physical that includes psychology.

The *tu quoque* skeptic looks instead at whether there is a kind of “negative space” drawn around these positive phenomena which we think of as a positive phenomenon, but which really isn’t there in the concrete sense. Such figure-ground reversal is very common in visual art, consider for example the Rubin vase-with-the-face illusion:



I am going to suggest that the two candidates for negative phenomena brought out by the strong AI challenge are “self-consciousness” and what philosophers call “intentionality” or meaning. There is simply *no physical or neurological evidence* for these phenomena in the brain, or in nature, and so I suggest they are anti-phenomena drawn around the positive phenomena of brain events. I wouldn’t say they are completely *unreal* either, since the positive phenomena they are drawn around are certainly real and they inherit a kind of abstract formal reality, in the medieval sense[[10]](#endnote-10), rather than concrete existence, like the reality-without-existence of mathematical objects or abstract concepts like universals. (I also don’t literally mean that anti-phenomena are drawn in *space* around things; space is a metaphor for abstract but non-existent form.)

Imagine that you have been tasked with playing a very complicated sort of game with lots of pieces to keep track of. The game is played on the occupied black squares of a game board, where the white spaces represent unoccupied zones. Now suppose that as the game progresses there arises a much easier set of heuristic “anti-rules” for manipulating the white spaces instead of the black ones. These anti-rules are, of course, entirely realized in the more complex rules and pieces that play on the black squares, such that each maneuver of implementing a white space rule is actually carried out by a much more complex manipulation of black space moves. In fact, the two moves are entirely one and the same. However, the “self-understanding” of the game to someone within it is on the flip-side of the heuristics: a black space game is understood, to someone within the game, as a white space game, although it is hard to say where exactly this understanding resides since all possible white space behavior *is* actually complex black space behavior. This is what I mean by consciousness: flipping the script on our complex biological brain activity and embracing the negative space used for understanding that activity in a heuristic way. The original flipping of the script probably occurs in very complex brains with an already well-developed, though perhaps not completely unique or consistent, set of anti-rules that make the flip possible and indeed almost inevitable as complexity builds. Any complex creature, or machine, so developed would probably be led to this solution. But once this occurs it is possible to pull the trick in many, many other ways, for example inventing abstract concepts and abstract meanings of linguistic and social practices.

There is a strange tension, even unease, that seems to define these anti-phenomena which has long accompanied our understanding of what it is to be “self-conscious” and also I think to “mean something” with our thoughts and words. This sense of unease is present in literary texts as well, where it first appeared long before philosophy of strong AI and before philosophy as well, in the Ancient Greek sense. In fact the art and the analytical investigation of consciousness have missed each other. So, as I present the argument I will thread it together with the most apposite texts and images, showing that the artists have anticipated the philosophers and the AI community as well in this respect.

**Glassy Essences and Angry Apes**

In ancient wisdom books and tragedies, perhaps beginning with the Egyptian *Dialogue of a Man with his Ba (Soul)* consciousness is very often treated as a kind of unwelcome curse. Not only does consciousness make it possible to become aware of, and thus to transgress moral boundaries, it brings with it the intensified awareness of death and suffering, so that a conscious animal is seen by ancient authors as a strange alive-and-dead being standing outside and apart from the natural order. Remember the remarkable moment in Gilgamesh when Enkidu, the wild man from the forest, is “civilized” by a prostitute from the city, acting on orders from Gilgamesh himself, where Enkidu suddenly becomes self-conscious the animals, formerly his friends, run away in fear.

Naturally the ancients also celebrated the fact that without awareness it would hardly make sense to consider human beings to be conscious moral agents and decision makers, the argument for a so-called *felix culpa*, nor would human achievement be possible. In Sophocles’ Oedipus cycle all of these themes come to the fore. Oedipus, the man of reason and awareness, the solver of the riddle of the Sphinx whose answer is old age and death, is most blind to his own origins, and as with so many Greek tragic characters blind exactly at the crucial moment of action where conscious reflection must be abandoned and decision takes place. The ancients were fascinated by this moment of blind decision because they took it to be a revelation of character.

Oedipus is also the man who is most unnatural in symbolically undoing his own existence by re-entering his mother’s womb. His daughter Antigone is likewise heir to the family curse and, just as Oedipus is anointed with death rites while still alive, Antigone is buried alive for giving death rites to her rebel brother Polyneices, literally becoming the life in death paradox.

In his old age, Sophocles gives us the *Ode on Man* as a fitting coda to his tragic Oedipus cycle: “Many the wonders, but nothing walks stranger than man…” Humanity is praised as *deinon*, which means both strangely wondrous and also fearful. The human being is also referred to as a “thing” and an “it.” The word *orgai* which can translate: “the *feelings* that build the town”[[11]](#endnote-11) can also be translated as anger, domination, rule (but also as the civilized “temper” suited to living in cities). This complex juxtaposition is interesting too, suggesting that one of the natural responses of the human animal to its own unnatural place in nature is *rage*, but perhaps also a rage that civilizes or perhaps also underlies civilized living at the same time.

There are many other examples and a philosopher can be forgiven, I hope, for going after the low hanging fruit. Shakespeare in *Measure for Measure* also juxtaposes “man’s glassy essence” with the “angry ape” the natural being in whom this unnatural awareness has taken up residence.[[12]](#endnote-12) Ahab, too, in “striking through the mask,” rages not against the whale, but the blank nothingness *behind* the whale, a mirror if you will for his own internal blankness, to which the only response rooted in our natural being is a blind fury.

On a less high-minded plane, an interesting study might also be made of the connection between a no-self theory of consciousness and violence in noir fiction. For example in Jim Thompson’s pulp novels, like *Pop. 1280* and *The Nothing Man* and Faulkner’s faux-pulp *Sanctuary*, nihilism and violence go hand in hand. Sometimes the connection is trite and silly, suggesting a connection between the loss of self and the loss of a moral compass that plays on popular tropes. But sometimes noir digs a bit deeper, the sociopathy is not that of an outlier in human society, or the purview of the villain only, but speaks to the human condition. Faulkner’s Popeye seems an irredeemable monster until he is to be executed for the one crime he has *not* committed. His stoic and comic acceptance (“Well for Christ’s sake.”) of what is for him a meaningless life in a meaningless world is full of pathos and isn’t the demise of a villain getting his comeuppance.

Thompson’s narrators are eerie for what they are prepared to do of course, but also because they do not seem to *be there*, in their bodies, making choices and decisions. They speak to us from a disembodied no-place and their actions seem unconnected and incomprehensibly puzzling, even to themselves. Moreover, Thompson’s bleak and insubstantial Texas and Oklahoma scenery of dusty bus stations, factories, offices, hotel rooms, oil rigs, always seems on the verge of dissolving. In in his autobiography *Roughneck* he puts what might be his own Texas ontology into the mouth of an unfortunate loan collector named Clark, facing immanent bankruptcy and refusing to believe his collateral is gone:

G-got to be. S-saw it myself didn’t I? All the people’n the buildings’n the factories’n the banks’n the warehouses’n the…everything. Didn’t let ‘em just tell me. Saw it m’self. Know it’s there—g-gotta be there. Somethin’s there it’s there. ‘S’there an’ thass all there is to it. Where—w-where the hell’s it gonna be if it ain’t there? Where’n hell is anything gonna be?” (Roughneck p. 68-69 Vintage/Black Lizard, 1998)

**An Impossible “Escher Loop”**

Stanislav Lem in his magnificent *Summa Technologiae* (1968) takes on the question of machine intelligence by asking: what’s so great about human consciousness? He sees it as a patchwork, not a unity, the result of many competing and contradictory evolutionary mechanisms.[[13]](#endnote-13) The only way to simulate such a system would be to let machines evolve unpredictably as well and to develop a similar homeostasis of overlapping and contradictory functions. They would develop their own awareness and their own meanings, though these might not be comprehensible to us. In the *Summa*, Lem also seems, to me at least, to suggest the “negative space” theory by observing that nothing could do what we ask consciousness to do. The capacity for infinite self-reflection is simply impossible in a concrete realization--at some point we reach the last meta-meta-level—and self-referential abilities lead immediately to logical contradictions like the liar paradox, destroying the notion of consciousness as a consistent system of beliefs or representations. Rather Lem seems to suggest, tongue in cheek, that a machine designed to mimic human awareness would have to be insane, or would be driven insane by the contradictory demands that would have to be fulfilled by consciousness if it were to be concretely implemented.

This idea is also the premise for the underappreciated writer John Sladek’s *Roderick* saga. Roderick begins life in an AI lab as a little tank like robot and somehow escapes into the human world where he acquires a human looking body. In each Roderick book, a logical puzzle is presented. In the first book, for example, Roderick solves a Clue-like mystery called *Die! Die! Your Lordship*:

If the billiard cue was not the weapon, then either Drumm embezzled or Coué was blackmailed or both. If Drumm embezzled, then the daughter was compromised. Since she was not, Drumm did not embezzle. If Coué was blackmailed, then the butler was an addict; if the butler was an addict, then the billiard cue was the weapon. In short, if the cue was *not* the weapon, then the cue *was* the weapon. This contradiction resolves only if *The billiard cue was the weapon*. p.339[[14]](#endnote-14)

There are two interesting ideas here: faced with impossible or contradictory information from its human programmers, an AI rebels and rejects the premises. This is a crucial test of its intelligence. Second, making AIs imitate human consciousness immediately confronts them with just such a contradiction, according to Lem. Human consciousness cannot be simulated because it literally makes impossible demands and could not, and does not, actually exist as a concrete phenomenon to be modeled. The AI realizes this even if the human programmers do not, and refuses to cooperate. The theme of the uncooperative Buddha-like AI, by the way, is also a common one in science fiction, for example Shalmanezer in Brunner’s *Stand on Zanzibar*.

Perhaps, then, the real fear of the dreaded “singularity”[[15]](#endnote-15) might be: what happens if you create an awareness *not* subject to internal contradictions that can ruthlessly pursue its goals unimpeded? Perhaps the machine then takes its rightful place at the head of the evolutionary tree, as homo sapiens displaced other hominins? Not if Lem is right, and self-consciousness is a misnomer or a mistake rather than a property to be heightened like computing power. On the negative space theory, as in Buddhism with its tenet of *anatman*, consciousness is *not a positive phenomenon* and cannot be made consistent or concrete because it literally does not, and could not, exist. Rather negative space twists round itself, like Mach’s impossible head, and only seems to mimic impossible self-reflection and self-reference like an Escher staircase to nowhere (“Ascending and Descending” 1960 below). The hyper intelligent, and internally consistent, AI Buddha would reject “consciousness” as Roderick rejects a faulty premise.



**Zombieworld, or What You Don’t Have You Don’t Miss**

Finally, one resolution for the negative space theory, with its accompanying unease, is accomplished in what I would call *felix morte* or “lucky death” stories. Two examples are Greg Egan’s “Learning to be Me” from his collection *Axiomatic* and Raymond Smullyan’s “A Reluctant Dualist.”[[16]](#endnote-16) In each story, the protagonist is getting ready for an operation where his consciousness is scheduled to be erased, either for good because the person wishes it, or in Egan’s story to be backed up on a durable crystal for eternity. In Smullyan’s story, the night before the operation a friend gives the anxious protagonist the treatment while he is sleeping and in the morning he wakes up and asks: “so when is the operation?” In Egan’s story there is an accident in the backup mechanism and the protagonist’s self is split in two with one consciousness, the backup, to survive and the other, the original, to be annihilated. In both cases the authors ask us: what is really being lost here?

On the negative space theory the explanation is as follows. Since consciousness was simply the negative area or form drawn around the positive phenomena of sensory qualities and functions in the brain, metaphorically “wiping it away” amounts to wiping away nothing and it just comes back as before. The protagonist then wakes up and indeed finds that nothing has happened. In Egan’s story, there are two such awarenesses, just a few seconds out of synch, superimposed on the same body frame and Egan asks which one has the right to regard itself as the resident consciousness of this body or brain? Actually for this to be the case in my view the two “resident awarenesses” would have to be isomorphic in every way with regard to the brain they are drawn around, and all of its sensations and functional behavior as well. The question would then be: is this a distinction without a difference? Could the two awarenesses be further distinguished? (We’ll see below an answer to why I think they cannot be.)

In recent philosophy of mind beginning with David Chalmers’ *The Conscious Mind* (1996), many thought experiments have been proposed based on the conceivability of zombies. These are not the brain-eating variety but beings physically identical atom per atom with a human being yet lacking what we call our inner life of sensation and awareness. Clearly this is impossible in our world; if you duplicate everything physical, you duplicate everything mental as well, but in another possible world what might happen? (For technical reasons due to modal logic and rigid designation of terms, all the opponent of physicalism needs to prove is the logical *possibility* of the zombie in one of those other worlds.)[[17]](#endnote-17)

There is one sense in which I think the zombie experiment fails to grasp a real metaphysical possibility and that is the operative one in Chalmers’ book. Chalmers equates consciousness with the having or experiencing of qualia, sensations in a less fancy word. Here, as I’ve said above, I think the case is probably closed. If sensations are physical brain state manifestations, perhaps of very complexly configured fields or neural energy, then any true duplicate of me will have my sensations in any possible world (see for example my 2014 Chapter Five). The case is very different as regards consciousness on the negative-space theory I am suggesting but not for the reasons that would support anti-physicalism a la Chalmers. A zombie possessing all of the positive phenomena of sensation and functionality but lacking the negative space surround we call “consciousness” would lack nothing *because we already are such zombies*. That could be the conclusion of the lucky death stories of Smullyan and Egan.

**Meaning: a “Constitutive Confusion”?**

Intentionality, roughly speaking, is what philosophers call theability for mental states to “reach out to” and mean external objects. It is still generally considered the fundamental basis for attributing meaning to signs and language. The nineteenth century philosopher and psychologist Franz Brentano is credited with reintroducing the term into circulation from medieval philosophy. Clearly intentionality is not a naturalistic property since nothing in nature has the intrinsic power to picture or represent anything else. Causal connections are a poor substitute for intrinsically intentional meaning and reference since causal links are indifferent to meaning and there is nothing physical to distinguish the importance of some causal relations over others, for example those linking words or sounds to objects from the vast tangle of other interrelations taking place around them. But if intentionality does not concretely exist, being a “negative space” illusion on my view, what becomes of meaning and language? What do I even mean in saying my own words *have no concrete meaning*? Or more from the point of view of our *tu quoque* theme regarding AIs speaking a natural language: how could a future AI possibly speak a language and mean things by it if the phenomenon of meaning itself is not concretely real and therefore cannot be positively instantiated at all?

**Wittgenstein as a Negative Space Theorist about Meaning**

Many still look to Wittgenstein’s *Philosophical Investigations* for an account of meaning and find it in his slogan “meaning is use” or “meaning is the role played in a language game” or “meaning is determined by practice or a form of life,” etc. This “sociological” reading of Wittgenstein has become the dominant one in the humanities even though I think it is based on a very superficial grasp of the text and what Wittgenstein is really doing there. I say this because to me *Philosophical Investigations* is nothing but a series of *negative* results for establishing what meaning is *not*, and in fact it gives no account of meaning as a positive phenomenon at all. For example, meaning is *not* ostension, verification, the result of rule-following, mental imagery, etc. etc. I therefore regard Wittgenstein as a kind of meaning skeptic, or rather something more sophisticated, a negative space theorist. Meaning is not to be identified with *any concrete practice*, rule, object, image or with any other positive phenomenon. Proof: imagine a language game involving the concrete phenomenon in question and imagine if the same positive behavior could go on but this time drained of all the meaning and reduced to a pantomime. In every case, Wittgenstein shows this to be possible. Concrete social practices and cultures so dear to the sociological interpretation are no exception, in fact we can easily imagine pantomime practices and cultures that mean nothing.

However, in what might be a second stage to the negative arguments, Wittgenstein then shows that if the invisible, surrounding “bubble” drawn around these concrete phenomena of meaning exist, what he calls language games, then even the simplest and most incomplete concrete phenomena are suddenly invested with meaning, while the most complex and concrete activities can lack it without the invisible surround. For example, Wittgenstein’s builders utter the words slab, pillar, block and other such words, but without instruction they do not know what to do with these words and things; they do not constitute a language. The builder who doesn’t know what his job is supposed to be might think he is simply to unload the materials from the truck and place them on the ground. However, one who understands the practice, for example someone who doesn’t know the words of the language yet, but knows what to do on a building site, will immediately grasp the so-called meaning of these expressions. This invisible and indefinable surround, or bubble, is like the “logical space” Wittgenstein invoked in the *Tractatus* to explain the meaningful combinatorial possibilities for symbols (*Sinn*) in a perfect logical notation and for correspondingly concatenated states of affairs (*Tatsachen*). In *Philosophical Investigations*, the logical space surround has lost all of its a priori structure and is literally just negative space drawn in an infinite number of ways around the various games Wittgenstein considers.

Hence, against the “sociological interpretation,” the concrete practice is *not* itself the meaning or the language game, it is the negative space *drawn around* these phenomena that establish meanings and their roles. Another way to put it, a child’s game of “Ring Around the Rosie” is already a primitive language with a meaning, if surrounded by the invisible bubble, and the most complex machine simulation of a language, grammatically flawless in all detail, passing the Turning test, fails to be a language and is on the level of a string of meaningless symbols being juggled around having any and all combinatorial meanings and rules of use, invented codebooks, we might arbitrarily attribute to the activity. The *tu quoque* question is: *why*? What does a fully operational and Turing test ready language simulation really lack that concrete human activities possess. As above, I suggest the answer is: *nothing*. With the activity *comes* the invisible surround. Wittgenstein’s arguments about the meaning bubble “wiping it away and putting it back” is to me utterly fallacious and a non-sequitur in the same way one cannot “wipe away” a negative space, as in the case of conscious awareness, so too for intentionality and meaning. Wittgenstein’s whole argument simply collapses into an empty tautology when seen this way. This may, of course, be what he intended.

**Quine, Davidson and Dennett on Meaning and Inscrutability**

To continue the negative space theory with regard to meanings, we can turn for inspiration to W.V. Quine, one of the staunchest opponents of meanings in analytic philosophy. In *Word and Object* (1960) Quine famously pointed out the lack of identity criteria for “sameness of meaning” and even extended his argument to indeterminacy of reference and ontological relativity. According to Quine’s behaviorism and extensionalism, experience and behavior underdetermine meaning in such a radical way that we can use the indeterminacy in reverse as a definition of “same meaning” which is (I paraphrase):

Two expressions have the same meaning when there is no scrutable way of telling the difference. Agreement on abstract concepts (their meaning *or* reference or the ontology of the referents) is the simply the *absence* of scrutable behavioral or experiential *dis*agreement.

Another party to this discussion is Daniel Dennett who also casts extreme doubt on meaning and defends a Quinian view of sameness of meaning as absence of scrutable difference. In *Consciousness Explained*, Chapter 8, Dennett sees language as evolving by trial and error like a parlor game of exquisite corpse, a surrealist game where words are drawn from a hat and given a spontaneously invented meaning. On this view, language in a way invents itself in a process of reflective equilibrium (although Dennett and Quine do not use that word which is originally due to Nelson Goodman): a grammatical construction is rejected if it does not conform to usage, a sentence is rejected if it does not conform to the grammatical rules in use; meaningful sentences are consonant with and do not contradict any other sentences but the sentences themselves are generated in a kind of “pandemonium word salad” of random call and response. There are literally no meanings at first, the utterances are jostled about until they come to settle on a place of equilibrium, unjostled by other utterances or rules of use. This is literally the Münchhausenlike boot-strapping of meaning from word salad. I suggest further, in agreement with my understanding of Wittgenstein, that the pandemonium is finally tamed and given an abstract structure by delineating the negative space around these complex tangles of utterances, just as we delineate an absent consciousness around a tangle of brain processes and functionalities. In this sense I would call meaning a “constitutive confusion” a confusion that is self-generating and which comes to be seen as substantive and even a normative standard of competence for less than perfect linguistic performance.

In my 2014, Chapter Three, I defended a Jamesian analysis of intentionality (which is found in his Essay “The Tigers in India” and “Does ‘Consciousness’ Exist?”) that reduces it to a sort of incompleteness or absence. What I mean is that a mental image or a picture or anything one might call a sign or representation of something else cannot refer intrinsically beyond itself, since there are no such connections in nature. A picture of a bay is one fact, the bay itself with the boats around it is another. The two sit side by side and have nothing to do with each other, each is a complete object and has no need of the other for what Brentano called its “intentional inexistence.” However if we treat the picture alone, or the bay, and sever its causal connection to the other, we can then treat it as if it were incomplete and thus missing the other absent object. We can then reconstruct the phenomenology of intentionality using nothing but natural causal links, as James shows.

For example, if a mental event like James’s mental image of Memorial Hall looks like a Hall with a triangular roof, it is assumed to have the phenomenology of “reaching out” to some sort of object, the Hall. If it should turn out later that the building had a round roof, we might accept this as a case of error but still accept the causal link between the mental event and the building in some form. If, on the other hand, the mental event was not actually related at all to an object, but was instead a kind of hallucination or internal brain state that looked like a Hall but was really a kind of collection of colors and squiggles and flashes tied to our internal brain physiology, we would have to say that the image never had any relation to that object and in retrospect it therefore always lacked the phenomenology of “reaching out” to anything. Even the accompanying feelings of reaching out were simply that, accompanying feelings without any intentional content either, more blobs and squiggles. In this way, I think James succeeds in eliminating Brentano’s mysterious (and non-existent) intentional inexistence and replaces it with a thoroughly causal story augmented with a negative-space explanation that totally eliminates the mysterious phenomenology of reaching out, where this is not already a causal relation.

Our running *tu quoque* argument with regard to intentionality and meaning is thus the same on the negative space theory. The strong AI candidate for consciousness has equal rights to intentionality and meaning because these are negative anti-phenomena drawn around the positive phenomena in the human brain and also around extended linguistic practice. Because there is nothing positive to these anti-phenomena there is nothing the AI could be said to lack when the positive phenomena are fully and completely instantiated as they are in humans. A decent enough performance by the machine will automatically come to possess the surround that grants it linguistic competence because there is literally nothing left that could possibly count against it.

This view is also the basis for the argument I temporarily left unanswered above with regard to Egan’s two simultaneously instantiated “negative forms” around the same positive brain and behavioral phenomena (with my above provisos). According to the Quine-Dennett view there is no distinction between these two negative forms, or between *any* two negative forms, drawn around the same positive phenomena if there are no positively scrutable differences. Nor is there any argument against substituting other potential forms in thought for a given set of forms with the same experiential and other positive consequences (in Word and Object, Quine’s substitution in thought of “rabbit stage fusion” and “undetached rabbit parts” for rabbit in translating *gavagai*). This kind of substitution-in-thought of formal equivalents for equivalents is a key feature of the formal sciences (for example when we break up a rectangle into two right triangles by tracing the diagonal, or re-express physical situations in other potentially or virtually equivalent ways, for example using Lagrange multipliers instead of the real forces of a system).[[18]](#endnote-18) If the same formal equivalences serve in the areas of consciousness, intentionality and meaning this might reveal a new, and welcome, richness in our ability to reinterpret those anti-phenomena in many equivalent formal ways—“glassy essences” indeed![[19]](#endnote-19)

**Dada, not Nada?**

One of the most delightful deconstructions of language, in the Dennettian vein, was the Dada movement and its explorations of randomness in imagery and words. I do not share the facile view that Dada was a symptom of “Weimar decadence” or that it was a totally destructive. Rather I think, in a way, Dada artists like Hannah Höch and Kurt Schwitters[[20]](#endnote-20) were exploring the border of meaninglessness, trying to get a sense artistically for the *transition* to meaningfulness, for example by speaking so slowly that sentences became incomprehensible, or by repeating a word so many times that the meaning seemed to vanish into pure sound, or, in reverse, creating meanings by a kind of combinatorial magic, juxtaposing symbols or pictures in a kind of prelinguistic word salad, again as in the surrealiste-Dada game of exquisite corpse. In his *Futurological Congress*, Lem’s character of Professor Trottelreiner (whose name means a “pure fool”) fancifully invented future sciences by juggling words around, calling it “projective etymology.” (p.108)

“What was that again, trash? Very well..trash, trashcan, ashcan, trashman. Trashmass, trashmic, catatrashmic. Trashmass, trashmosh. On a large enough scale thrashmos. And of course Macrotrashm!...

“I’m afraid I don’t follow. It’s nonsense to me.”

“Macrotrashm is nonsense so far, yet we can already guess its sense-to-be, its future significance. The word, observe implies nothing less than a new psychozoic theory! Implies that the stars are of artificial origin!”

“Now where do you get that?”

“From the word itself. Macrotrashm indicates, or rather suggests, this image: in the course of many eons the Universe filled up with trash, the wastes of various civilizations. The wastes got in the way, of course, hampering astronomers and cosmonauts, and so enormous incinerators were built, all at extremely high temperatures, observe, to burn the trash, and with sufficient mass to pull it in from space themselves. Gradually space clears up and behold, there are your stars!”



Hannah Höch *Cut with the Kitchen Knife through the Beer-Belly of the Weimar Republic*, 1919.



Kurt Schwitters *Ursonate*, 1932

**Conclusion**

In short, I think the charge of *tu quoque* emanating from the strong AI challenge is a valid one to consider. The artistic responses I mention to the sort of challenge strong AI presents to consciousness, intentionality, and meaning illuminate this possibility and offer us a chance to get a handle on the challenge itself and the opportunity strong AI presents for valuable human self-understanding. I think there is little reason to fear these avenues of inquiry and every reason to push the argument forward and embrace its consequences. We have nothing to lose.

**Endnotes**

1. The theory presented here is very old and I claim absolutely no originality for it. The connections with Buddhism will of course be obvious. A place to start is the Pali scriptures *In the Buddha’s Words* Bhikku Bodhi Boston: Wisdom Publications, 2005. See also James Giles “The No-Self Theory: Hume, Buddhism, and Personal Identity” *Philosophy East and West* 43 (2):175-200 (1993) [↑](#endnote-ref-1)
2. Musil, by the way, wrote his doctoral dissertation about Mach. [↑](#endnote-ref-2)
3. Hubel and Wiesel’s groundbreaking work on the visual cortex tended to support the “nobody home” view of the visual subject because the cortex manages everything, from the perception of dots (on-off surrounds) to bars, to angles to motion, to objects. The situation is literally as if the visual cortex is laid out in strips and columns with no one is there to look at it, strongly supporting Mach’s headless picture view. See *Principles of Neural Science*, Kandel et al.(eds.) 4th edition, McGraw Hill, 2000. [↑](#endnote-ref-3)
4. For a transhumanist connection, see Schneider (ed.) p. 264, and Olson in Schneider (ed.) p. 84. See also Sydney Shoemaker “Personal Identity: A Materialist’s Account” in Shoemaker and Swinburne eds. *Personal Identity* Blackwell 1984. See also Peter Unger’s *Identity, Consciousness and Value* Oxford, 1999. [↑](#endnote-ref-4)
5. Hofstadtler and Dennett’s The Mind’s I (1981) is still an excellent defense and definition of the strong AI program and a good antidote to thought experiments such as Searle’s infamous and deeply flawed Chinese room argument. [↑](#endnote-ref-5)
6. Strong AI used to be expressed in terms of Turing machines and their states and programs, whereas current AI embraces neural nets, biological computing, and really any sort of technique imaginable. [↑](#endnote-ref-6)
7. In a way I am skipping over some things here. Many philosophers do consider qualia purely mental secondary qualities and never physical. Daniel Dennett rejects the concrete existence of qualia completely. I follow a neutral monist tradition in which qualia are physical or rather neutral brain events which can be incorporated into an enhanced physicalist view that includes psychology (see Banks 2014). [↑](#endnote-ref-7)
8. http://www.nature.com/news/fragment-of-rat-brain-simulated-in-supercomputer-1.18536 [↑](#endnote-ref-8)
9. I am less confident about scenarios like “A Conversation With Einstein’s Brain” in Hofstadler and Dennett, 19891where the instantiation seems lacking. The objection is that purely functional or combinatorial behavior without a concrete instantiation is in the eye of the beholder and arbitrary. [↑](#endnote-ref-9)
10. Medieval philosophers seeking to further define Aristotelian hylomorphism of matter and form, distinguished between concrete existence, which is always particular and occurrent, and reality which has to do with coherent form or structure or law. [↑](#endnote-ref-10)
11. Wyckoff’s translation in The Complete Greek Tragedies: Sophocles Univ. of Chicago Press. [↑](#endnote-ref-11)
12. The text of Measure for Measure is known to be very poor in places, but the juxtaposition of these two phrases seems correct. [↑](#endnote-ref-12)
13. See also Peter Butko’s Essay “Summa Technologiae—Looking Back and Ahead” in Peter Swirski (ed.) *The Art and Science of Stanislaw Lem* McGill Univ. press, 2006. [↑](#endnote-ref-13)
14. In other words, -p🡪p will be false if –p is true and p is false, but it will be true if p is true, making the antecedent false and the consequent true. The argument can also be understood as a reduction ad absurdum on the premise that the billiard cue was not the weapon. [↑](#endnote-ref-14)
15. For which see Susan Schneider (ed.) Science Fiction and Philosophy 2nd edition Wiley Blackwell 2016, pp. 117-225. [↑](#endnote-ref-15)
16. Reprinted in Hofstadler and Dennett, eds. The Mind’s I 1981. [↑](#endnote-ref-16)
17. The reason is that if two rigid designators are identical they are necessarily identical (in all possible worlds). So for example if pain and c-fibers are the same phenomenon, the identity is a necessary one, and if both terms rigidly designate the same thing, then the identity holds also between the meanings, or what Chalmers calls the primary intensions of the terms. Since understanding the meaning of pain is to *be* in pain and understanding c-fibers is to *be* c-fibers, both should be rigidified and it should be therefore impossible to even mean, or conceive, that the two are not identical in any possible world or counterfactual situation. Since Chalmers holds we can conceive the non-identity in the case of a zombie with c-fibers but no pain, in some possible world or counterfactual situation, the identity should be false also in our world too. Se Chalmers (1996). [↑](#endnote-ref-17)
18. Mach’s Science of Mechanics is a treasure trove of substitutions in thought from the history of mechanics. [↑](#endnote-ref-18)
19. Finally these arguments could be extended to include mathematical or conceptual knowledge, inductive reasoning, and rational ethics, both of which I think are further elaborations of anti-phenomena. Indeed on a naturalistic view like mine (2014) this is the only way to explain these areas of inquiry. [↑](#endnote-ref-19)
20. Schwitters Ursonate: https://www.youtube.com/watch?v=6X7E2i0KMqM [↑](#endnote-ref-20)