Aïda Raoult

Of Mills and Machines

Computing Thought Experiments on Consciousness

Abstract: *While there is no simple and consensual definition of consciousness, the topic has often been approached by "thought experiments", a peculiar process that serves to problematize a difficult question. As consciousness and machines have gathered renewed academic interest with the progress of computing, it is somewhat fascinating that one of the most historically known thought experiments on consciousness, "Leibniz's Mill", concerns something akin to a computer that would "think, sense and have perceptions". In recent years, as reality got closer to thinking machines, thought experiments have left computers and have been interested in zombies or bats. In this paper, we try to "translate" these more recent, refined and precise thought experiments in the same terms as Leibniz's so as to see what they add to the consideration of a conscious machine. We show that these experiments can be reduced to two "questions", one more complex that the other, and that they highlight two elements in the mind-body discussion, qualia and intentionality. These dichotomies nourishes which research avenues in machine consciousness can go beyond the current debates on self-awareness.*

Correspondence:
Aïda Raoult, Aix-Marseille Université, CETFI, Europôle de l'Arbois, Avenue Louis Philibert, Hall D Laennec, 13857 Aix en Provence Cedex 3 - France
Email: araoult@uchicago.edu

1. Introduction

With the recent progress of robotics and computing capabilities, computer scientists have had a renewed interest in machine consciousness, or strong AI as opposed to weak AI (the contemporary meaning of “artificial intelligence”). Since there is no short, functional and consensual definition of consciousness, in applied robotics, scientists generally refer to a particular aspect of what is discussed in literature as “consciousness”, most notably self-awareness[1]. When we look at how consciousness has been described in the western tradition, instead of a “definition”, we find an extensive usage of thought experiments[2]. In fact, their late recrudescence in the philosophy of mind has participated in initiating the on-going debate on their validity[3], some going as far as calling thought experiments “intuition pumps”[4]. From a more agnostic point of view, one can at least recognize thought experiments such as Leibniz’s Mill, Philosophical Zombies, Being a Bat, Mary’s Room, the Chinese Nation or the Chinese Room as interesting components of consciousness research, *problematizations* of the mind-body relationship.

What is fascinating is that one of the most historically known thought experiments on consciousness, “Leibniz’s Mill”, a thought experiment that may, as we will show, serve as a base to describe all others, is describing a conscious machine, something akin to a computer that would “think, sense and have perceptions”. This major thought experiment of the 18th century was already addressing the question of conscious machines, attesting that machines are one of the favored paths through which consciousness is initially

[1] Chella, Antonio, Marcello Frixione, and Salvatore Gaglio. "A cognitive architecture for robot self-consciousness." Artificial Intelligence in Medicine 44.2 (2008): 147-154. Takiguchi, Toshiyuki, Atsushi Mizunaga, and Junichi Takeno. "A study of self-awareness in robots." International Journal of Machine Consciousness 5.02 (2013): 145-164. Gorbenko, Anna, Vladimir Popov, and Andrey Sheka. "Robot self-awareness: Exploration of internal states." Applied Mathematical Sciences 6.14 (2012): 675-688. Kawamura, Kazuhiko, et al. "Development of a robot with a sense of self." Computational Intelligence in Robotics and Automation, 2005. CIRA 2005. Proceedings. 2005 IEEE International Symposium on. IEEE, 2005.

[2] Cole, David. "Thought and thought experiments." Philosophical Studies 45.3 (1984): 431-444.

[3] Ierodiakonou, Katerina, and Sophie Roux, eds. Thought Experiments in Methodological and Historical Contexts. Vol. 19. Brill, 2011. Gooding, David C. "What is experimental about thought experiments?." (1992). Machery, Edouard. Gendler, Tamar Szabó. Thought experiment: On the powers and limits of imaginary cases. Garland Pub., 2000. "Thought experiments and philosophical knowledge." Metaphilosophy 42.3 (2011): 191-214.

[4] Dennett, Daniel C. Intuition pumps and other tools for thinking. WW Norton & Company, 2013. Brendel, Elke. "Intuition pumps and the proper use of thought experiments." Dialectica 58.1 (2004): 89-108.

approached in modern western thought. Paradoxically, as reality got closer to thinking machines, thought experiments have left machines and have been interested in zombies, bats, sequestrated ladies and China. But there may be a way to "translate" these more recent, refined and precise thought experiments in the same terms as Leibniz's so as to see what they add to the consideration of a conscious machine. This is the endeavor this paper pursues.

After a review of the six aforementioned major thought experiments on consciousness (section 2), we will attempt to translate the most recent ones into thought experiments on machine consciousness (section 3). In the process we will identify two key elements in the mind-body problem discussion, namely *qualia* and *intentionality*, which point out how machine consciousness can be thought of beyond machine self-awareness (section 4).

2. A review of six thought experiments on consciousness

We will review six thought experiments on consciousness, which were all, aside from Leibniz's Mill, published between 1974 and 1996. Almost any scholarly book approaching the philosophy of consciousness refers to at least one of them[5], if not all of them[6], which we may see as amongst the most important one. Other related thought experiments, such as the "brain in a vat"[7] or the "inverted spectrum"[8], were not included in this work either because they focused on specific issues that we will not be addressing, or because they were too redundant in virtue of those we selected.

Leibniz's mill (LM)

Leibniz's philosophy takes place in the 18th century, during the Enlightenment, at a time when Descartes' ideas have given rise to rationalism[9]. Directed against Cartesian dualism[10], The

[5] Edelman, Gerald M., and Giulio Tononi. A universe of consciousness: How matter becomes imagination. Basic books, 2000. Penrose, Roger. The emperor's new mind: concerning computers, minds, and the laws of physics. Oxford University Press, 1999. Blackmore, Susan. Consciousness: an introduction. Routledge, 2013.

[6] Dennett, Daniel C. Consciousness explained. Penguin UK, 1993.

[7] Putnam, Hilary. Reason, truth and history. Vol. 3. Cambridge University Press, 1981. Dennett, Daniel. "Where Am I?" as reprinted in Brainstorms: Philosophical essays on mind and psychology. No. 8. Mit press, 1981: 310-23.

[8] Locke, J., Essay Concerning Human Understanding, Oxford: Oxford University Press, 1689/1975: II, xxxii, 15.

[9] Cassirer, Ernst, 1932. The Philosophy of the Enlightenment, tr. Fritz C.A. Koelln and James P. Pettegrove, Boston: Beacon, 1955.

[10] GP VI, 608/AG 214, §14. Sources of Leibniz's work normally refer to Die Philosophischen Schriften von Gottfried Wilhelm Leibniz. 7 vols. Edited by C.I. Gerhardt. Berlin: Weidman, 1875-1890. Cited by volume and page. (GP) and

Monadology contains Leibniz's unachieved work on his monist theory based on the concept of *unextended*[11] atomistic particles which he refers to as *monads*. Scholars have not reached on an agreement over the various interpretations of this theory[12]. The Monadology contains Leibniz's famous thought experiment, commonly known as "Leibniz's Mill". This is the translation given by Roger Ariew and Daniel Garber of the original quote: "[P]erception, and what depends upon it, is inexplicable in terms of mechanical reasons, that is through shapes and motions. If we imagine that there is a machine whose structure makes it think, sense, and have perceptions, we could conceive it enlarged, keeping the same proportions, so that we could enter into it, as one enters a mill. Assuming that, when inspecting its interior, we will find only parts that push one another, and we will never find anything to explain a perception. And so, one should seek perception in the simple substance and not in the composite or in the machine."[13]

Like the Monadology in general, several meanings are attributed to this paragraph. A thorough review of the discussion has been published less than two decades ago[14]. Amidst this debate, what does not seem to be the subject of a dissension is how crucial the notion of *unity* is to Leibniz's argument. Indeed, in a preceding paragraph, he defines a perception as a state which "involves and represents a multitude in the unity or in the simple"[15], and that is precisely the difference with the mill, which, as any artefact, would be made of parts or pieces put together. Whether this experiment is intended as a proof or a consequence[16] of this unity, the fact still remains that no matter the scale, the mill will never allow us to physically observe any perception, because these are not located in the material realm.

Philosophical zombies (PZ)

This thought experiment can be traced back to an article Robert Kirk published in 1974[17], but it has been taken over by several authors

---

Philosophical Essays. Edited and translated by Roger Ariew and Daniel Garber. Indianapolis: Hackett, 1989. (AG) for the translation.

[11] GP VI, 607/AG 213, §3.

[12] Lodge, P. and Bobro, M., 1998. "Stepping back inside Leibniz's Mill," Monist, 81: 554–73.

[13] GP VI, 609/AG 215, §17.

[14] Lodge, P. and Bobro, M., 1998. "Stepping back inside Leibniz's Mill," Monist, 81: 554–73.

[15] GP VI, 608/AG 214, §14.

[16] In Lodge Bobro 1998, the authors propose their own interpretation of the experiment claiming it should be taken as a corollary to a previous assertion Leibniz makes in The Monadology.

[17] Kirk, R. (1974). Sentience and behaviour. Mind, 83(329), 43-60.

since then, most notably by David Chalmers since the 90s[18]. We can find different kind of zombies in the literature, at different stages of "zombification", but they are all aimed at challenging materialism.

The experiment as posed by Kirk is "the story of Dan, who, it seems, turned into a Zombie by stages. One day Dan […] accidentally cut his hand and started to behave very strangely. True, his behaviour included features indistinguishable from the typical behaviour of a normal person suffering pain of the kind and intensity to be expected in the circumstances. He winced, exclaimed 'Ouch!', […] in addition […] [h]e expressed astonishment […]. He appeared to be astonished by two things: first, by the fact […] that he really felt no pain at all in spite of a fairly serious injury; and second, by the fact that he was nevertheless wincing, groaning, […] he said that so far as his […] 'pain-behaviour[s]' were concerned, they had been a series of happenings over which he had had no control: he had felt 'like a puppet'[…]. [I]f we were faced by Dan or someone behaving like Dan, it would be not merely intelligible, but very reasonable, indeed justifiable, to say that he had ceased to feel pain, despite his apparent pain-behaviour". After this incident, Dan never recovers his ability to feel pain, instead he goes through three more episodes of this kind, each time losing another of his senses: first his smell, then his sight and finally his hearing; and he keeps involuntarily behaving as if nothing had changed. In fact, when he reaches the final stage of his transformation, nobody can make a difference between the original Dan and his new self, D., the zombie. Aside from the irony that has been made regarding Dan's mental health[19], the point of the author is that the mere logical possibility of the existence of such zombies conflicts materialist theories, notably behaviorism and functionalism. Indeed, since such theories define sensations as pain by characteristics of their physical expression, how could they account for zombies which/whom display every physical aspects of the sensation but without the usual accompanying phenomenology? However, the sensation in itself (the *quale*), is not the core of Kirk's argument; he is more preoccupied by the logical possibilities that are allowed by theories based on physical descriptions of consciousness, which, from his

[18] Chalmers, D. J. (1993). Self-Ascription without qualia: A case study. Behavioral and Brain Sciences, 16(1), 35-36; Chalmers, D. J. (1996). The conscious mind: In search of a fundamental theory. Oxford University Press; Chalmers, D. J. (1997). Moving forward on the problem of consciousness. Journal of Consciousness studies, 4(1), 3-46; Chalmers, D. J. (1999). Materialism and the Metaphysics of Modality.

[19] Locke, D. (1976). Zombies, schizophrenics, and purely physical objects. Mind, 85(337), 97-99.

point of view, seems to indicate these are not satisfying models of materialism.

The work of David Chalmers deepened and popularized the issue of philosophical zombies. Sharing the same concerns as Kirk, Chalmers addresses them in terms of supervenience: “B-properties supervene on A-properties if no possible two situations are identical with respect to their A-properties while differing in their B-properties.”[20] His version of the experiment is simpler, and focuses on the absence of qualia from the zombie’s perspective: “Let us consider my zombie twin. This creature is molecule for molecule identical to me, and identical in all the low-level properties postulated by a completed physics, but he lacks conscious experience entirely.” [21]

This amounts to say that phenomenological properties do no logically supervene on the physical ones. The most obvious interpretation of this experiment is epiphenomenalism, the view that conscious experiences are an incidental byproduct of the physical world because they bear no effect on it. This doctrine has been criticized in length[22], and it is not what David Chalmers sought to demonstrate[23]. Apart from this, criticisms of the zombies experiment revolves around their conceivability. It has been claimed that the conceivability of zombies does not entail their possibility[24], or that it is not possible to conceive zombies because their definition hides a fundamental incoherence[25]; some even went as far as to argue that the same argument can be used in favor of materialism[26]. While the literature is divided, philosophical zombies have survived this criticism and are a significant part of the contemporary landscape of consciousness studies.

---

[20] Chalmers 1996, p.31

[21] Chalmers 1996, p.92

[22] Locke, Edwin A. "The contradiction of epiphenomenalism." British Journal of Psychology 57.1-2 (1966): 203-204. Block, Ned. "Evidence against epiphenomenalism." Behavioral and Brain Sciences 14.04 (1991): 670-672.

[23] The Conscious Mind, pp. 150-154

[24] Balog, Katalin. "Conceivability, possibility, and the mind-body problem." Philosophical Review (1999): 497-528. Hill, Christopher S. "Imaginability, conceivability, possibility and the mind-body problem." Philosophical Studies 87.1 (1997): 61-85.

[25] Marcus, Eric. "Why zombies are inconceivable." Australasian Journal of Philosophy 82.3 (2004): 477-490. Thomas, Nigel J. T. (1998). Zombie killer. In Stuart R. Hameroff, Alfred W. Kaszniak & A. C. Scott (eds.), Toward a Science of Consciousness II. MIT Press. Dennett, Daniel C. (1995). The unimagined preposterousness of zombies. Journal of Consciousness Studies 2 (4):322-26. Kirk, Robert. "The inconceivability of zombies." Philosophical Studies 139.1 (2008): 73-89.

[26] Frankish, Keith. "The Anti-Zombie Argument." The Philosophical Quarterly 57.229 (2007): 650-666.

Being a bat (BAB)
In 1974, Thomas Nagel confronts the reductionist trend that had spread through philosophical theories of the mind with the following experiment: to imagine oneself being a bat. The choice of this animal is not random: while we may relate better to mammals, the bat's sensory system is so radically different from ours that we cannot pretend to know how it feels to experience its perceptions. That is the point of the experiment: in his words: "[…]the essence of the belief that bats have experience is that there is something that it is like to be a bat. Now we know that most bats […] perceive the external world primarily by sonar, or echolocation, detecting the reflections, from objects within range, of their own rapid, subtly modulated, high-frequency shrieks. Their brains are designed to correlate the outgoing impulses with the subsequent echoes, and the information thus acquired enables bats to make precise discriminations of distance, size, shape, motion, and texture comparable to those we make by vision. But bat sonar, though clearly a form of perception, is not similar in its operation to any sense that we possess, and there is no reason to suppose that it is subjectively like anything we can experience or imagine. This appears to create difficulties for the notion of what it is like to be a bat."[27]

Following Nagel's idea, these difficulties stem from the *subjective character of experience*: what it feels to *undergo* the experience, regardless of any of its other manifestations; to answer that, the question of whether mental phenomenon could be fully explained by physical theories is not relevant. However, it has been objected[28] that this apparent obstacle is simply a consequence of the limitations of our current knowledge and that further scientific advancement may change this situation by providing us an adequate representational machinery for a foreign set of qualia[29]. Further critics merely reject the idea of an enigmatic feature of experience claiming such a misconstruction is the result of "conceptual confusions"[30], though this point had already been addressed has a "phenomenological fallacy" before Nagel's experiment[31], or that an

---

[27] Nagel, Thomas. "What Is It Like to Be a Bat?", The Philosophical Review, Vol. 83, No. 4 (Oct., 1974), pp. 435-450.
[28] Akins, Kathleen. "Science and our inner lives: Birds of prey, bats, and the common (featherless) bi-ped." Interpretation and explanation in the study of animal behavior 1 (1990): 414-427.
[29] Dennett, Daniel C. Consciousness explained. Penguin UK, 1993.
[30] Hacker, P. M. S. "Is there anything it is like to be a bat?." Philosophy 77.2 (2002): 157-174.
[31] Place, Ullin Thomas. "Is consciousness a brain process?." British Journal of Psychology 47.1 (1956): 44-50.

"objective account of consciousness is possible"[32]. In any case, in the end, we cannot share our *raw* experiences, we need something, maybe a revolutionary tool[33], but obviously more than words in order to apprehend subjective experiences.

Chinese nation (CN)

This experiment was designed by Ned Block primarily as an argument against functionalist theories of the mind[34] based on the intuition of the existence of qualia. He purposefully takes the functionalist reasoning to the extreme, proposing to generate the brain functions by equipping Chinese citizens with radios: "Suppose we convert the government of China to functionalism, and we convince its officials to realize a human mind for an hour. We provide each of the billion people in China (I chose China because it has a billion inhabitants) with a specially designed two-way radio that connects them in the appropriate way to other persons and to the artificial body mentioned in the previous example. We replace each of the little men with a citizen of China plus his radio. Instead of a bulletin board we arrange to have letters displayed on a series of satellites placed so that they can be seen from anywhere in China. The system of a billion people communicating with one another plus satellites plays the role of an external "brain" connected to the artificial body by radio."[35]

Ned Block then rejects functionalism on the basis that it leads to the "absurd conclusion" that such a system would have qualia. It is worth noting that, in the same article, Block started his argument with another experiment he referred to as "the homunculi-headed robot", supposed equivalent to that of the Chinese nation, where an apparently human body would have been animated on the inside by an army of little men, operating with cards and buttons to generate our internal functions. And this experiment strikingly resembles an alternate experiment of Robert Kirk, where a human puppet is controlled by what he calls "the brain team"[36], which is none other than a team of microscopically small scientists. Kirk achieves the same conclusion, but using another vocabulary: for him, the *causal roles* are performed by the work of the brain team, but the puppet is not *sentient*.

---

[32] Kekes, John. "Physicalism and subjectivity." Philosophy and Phenomenological Research (1977): 533-536.

[33] Among the recent advances in VR (Virtual Reality), the headset Oculus Rift has been used for swap gender experiments by the art collective BeAnotherLab.

[34] Theories reducing mental properties to their functional role as introduced by Putnam, Hilary. "Psychological predicates." Art, mind, and religion (1967): 37-48.

[35] Block, Ned. "Troubles with functionalism." Readings in the Philosophy of Psychology 1 (1980): 268-305.

[36] Kirk, R. (1974). Sentience and behaviour. Mind, 83(329), 43-60.

In his article, Ned Block considers several objections about the validity of the experiment, particularly concerning the influence of the space and time scales which are not taken into account even though they may be decisive features of human consciousness; but he dismisses them by pointing out that there might be non-human forms of consciousness acting on other scales, and any kind of consciousness is enough to serve his purpose of attacking functionalism, the "China brain" doesn't need to specifically mimic human consciousness. It has since been argued that the intuitive premises are wrong and that the China brain would indeed be conscious either by relying on the absence of any empirical evidence[37] or by calling to the concept of collective consciousness, in the tradition of East Asian cultures[38].

Chinese room (CR)

A.I. research aroused keen interest in questions concerning the scope of machine thinking[39]. In 1980, John Searle identified two opposite positions on the subject: *weak A.I.* is the view of computers as very advanced calculators whereas according to *strong A.I.*, with the right program or architecture, computers can eventually have minds comparable to ours. As a supporter of weak A.I., Searle proposed this thought experiment to demonstrate that computer programs will never be capable of human understanding: "Suppose that I'm locked in a room and given a large batch of Chinese writing. Suppose furthermore (as is indeed the case) that I know no Chinese, either written or spoken[…].Now suppose further that after this first batch of Chinese writing I am given a second batch of Chinese script together with a set of rules for correlating the second batch with the first batch. The rules are in English, and I understand these rules as well as any other native speaker of English. They enable me to correlate one set of formal symbols with another set of formal symbols, and all that 'formal' means here is that I can identify the symbols entirely by their shapes. Now suppose also that I am given a third batch of Chinese symbols together with some instructions, again in English, that enable me to correlate elements of this third batch with the first two batches, and these rules instruct me how to give back certain Chinese symbols with certain sorts of shapes in response to certain sorts of shapes given me in the third batch. Unknown to me, the people who are giving me all of these symbols

[37] Dennett 1993, p 435

[38] Huebner, Bryce, Michael Bruno, and Hagop Sarkissian. "What does the nation of China think about phenomenal states?." Review of Philosophy and Psychology 1.2 (2010): 225-243.

[39] Weizenbaum, Joseph. "Computer power and human reason: From judgment to calculation." (1976). Dreyfus, Hubert L. "What computers can't do." (1972). Moor, James H. "Three myths of computer science." British Journal for the Philosophy of Science (1978): 213-222.

call the first batch "a script," they call the second batch a "story. ' and they call the third batch "questions." Furthermore, they call the symbols I give them back in response to the third batch "answers to the questions." and the set of rules in English that they gave me, they call "the program."[…] Nobody just looking at my answers can tell that I don't speak a word of Chinese. […] I simply behave like a computer; I perform computational operations on formally specified elements. […] I am simply an instantiation of the computer program."[40]

This experiment encompasses many issues relating minds at the same time: consciousness, understanding, syntax vs semantics, and intentionality. Consequentially, there are several levels of interpretation for this experiment, each about a different aspect of the same problem - the possibility/impossibility of artificial minds - through the question of human understanding: is it possible for a computer to understand the meaning of the words it is manipulating? This experiment has raised a number of objections[41], possibly stronger than the others, the most vocal being Douglas R. Hofstadter, who called it a “religious diatribe against AI, masquerading as a serious scientific argument [and] one of the wrongest, most infuriating articles [he had] ever read in [his] life”[42]. Part of these critics rejects the interpretation of this experiment, while the other focuses on the implementation of the experiment by explaining why the room cannot understand Chinese or how it could. Among the first group, the most common argument, known as “the system reply”, claims that the man does not understand Chinese only because he is part of a bigger structure, like the CPU[43] in the computer[44], in addition they hold that the system as a whole understands[45], or that it has produced a virtual mind that understands[46]. Critics from the second group attack the outcomes of the experiment by blaming the program for being too rudimentary,

---

[40] Searle, John R. "Minds, brains, and programs." Behavioral and brain sciences 3.03 (1980): 417-424.
[41] In fact, BBS (behavioral and brain sciences) published the original article along with 27 commentaries.
[42] Hofstadter, Douglas R. "Reductionism and religion." Behavioral and Brain Sciences 3.03 (1980): 433-434.
[43] The central processing unit (CPU) is a hardware device that is responsible for the calculations in the computer.
[44] Rey, Georges. "What's really going on in Searle's “Chinese room”." Philosophical Studies 50.2 (1986): 169-185.
[45] Block, Ned. " The Mind as the Software of the Brain " Science Fiction and Philosophy: From Time Travel to Superintelligence (2009): 126.
[46] Hayes, Patrick, et al. "Virtual symposium on virtual mind." Minds and Machines 2.3 (1992): 217-238.

inducing a lack of intentionality because it is deprived of initiative[47], not being able to form its own representations of the world -from which meaning would emerge[48]- nor to recall them later, and in this sense CR's program seems to address the behavior of an unsophisticated chatbot[49,50]. These arguments are only a fraction of the diversity of replies that have been made for and against CR and that contributed to John Searle's paper being "BBS's most influential target article"[51]. In the end, even though the experiment has not closed the subject of "strong A.I." as it intended to, it has raised many questions and led to further discussions in in the scientific community.

Mary's room (MR)

As a strong advocate of the epiphenomenal character of qualia, and self-proclaimed "qualia freak"[52], Frank Jackson exposes his views in a paper unequivocally entitled "Epiphenomenal Qualia" where he formulates the *knowledge argument*, illustrated by the following thought experiment: "Mary is a brilliant scientist who is, for whatever reason, forced to investigate the world from a black and white room via a black and white television monitor. She specializes in the neurophysiology of vision and acquires, let us suppose, all the physical information there is to obtain about what goes on when we see ripe tomatoes, or the sky, and use terms like 'red', 'blue', and so on. She discovers, for example, just which wave-length combinations from the sky stimulate the retina, and exactly how this produces via the central nervous system the contraction of the vocal chords and expulsion of air from the lungs that results in the uttering of the sentence 'The sky is blue'. (It can hardly be denied that it is in principle possible to obtain all this physical information from black and white television, otherwise the Open University would of necessity need to use colour television.) What will happen when

---

[47] Kugel, Peter. "The Chinese room is a trick." Behavioral and Brain Sciences 27.01 (2004): 153-154. Kugel, Peter. "Computing Machines Can't Be Intelligent (... and Turing Said So)." Minds and Machines 12.4 (2002): 563-579.

[48] Thagard, Paul. "The emergence of meaning: how to escape Searle's Chinese room." Behaviorism 14.2 (1986): 139-146.

[49] Chatbots, or chat robots, are computer programs designed to have a conversation with human beings; in its most basic form, a chatbot scans the user input for keywords and statistically determines the corresponding answer from a pre-constructed database of replies.

[50] Proudfoot, Diane. "Wittgenstein's anticipation of the Chinese Room." (2003).

[51] Harnad, Stevan. "Minds, machines and Searle 2: What's right and wrong about the Chinese Room Argument." Preston and Bishop (eds.) (2002): 294-307.

[52] Jackson, F. (1982). Epiphenomenal qualia. *The Philosophical Quarterly*, 32(127), 127-136.

Mary is released from her black and white room or is given a colour television monitor? Will she learn anything or not?"[53]

With no surprise, Jackson's answer is positive; yes, by actually seeing the color itself, Mary learns something that no colorless book, no black and white video course or any other material available in her achromatic prison could have ever taught her. Therefore, *knowledge* fails to encompass the whole of consciousness, which indicates that qualia must lie beyond the scope of physicalism.

Dennett answers that the premises abuse our intuition, and without the entire physical knowledge about color we fail to imagine how it could allow Mary to discriminate between blue and red even though she never saw anything but black and white before[54]; there is no reason *a priori* to discard the scenario where Mary walks out of the room and learns nothing. In the same line of thought, a Russellian monist[55] could also claim that the premises are inconsistent because the sensory input corresponding to the wavelength of red hitting our retina *is* an actual physical information that is relayed in our brain through a series of purely physical processes, even though we have not identified all the details of the chain of events yet. Hence, for Mary to acquire all the physical information about color she would unavoidably require to see every color. Nevertheless, even then MR still presents the interest of highlighting how information can be conveyed through a number of channels and take various forms. Jackson himself later revised his opinion, explaining Mary's new knowledge in terms of representational states[56]. At any rate, this experiment makes for a striking illustration of what qualia are and how unseizable they seem.

3. Thought experiments recast for machine consciousness

The thought experiments that are studied in this paper can be understood as diverse problematizations of the mind-body relationship. And while there is no consensual formulation of *a* mind-body problem[57] – sometimes it is not even agreed on why

[53] Jackson, F. (1982). Epiphenomenal qualia. *The Philosophical Quarterly*, 32(127), 127-136.
[54] Consciousness Explained.
[55] Alter, Torin. "What is Russellian monism?." Journal of Consciousness Studies 19.9-10 (2012): 9-10.
[56] Jackson, Frank. "Mind and illusion." Royal Institute of Philosophy Supplement 53 (2003): 251-271.
[57] McGinn, Colin. "Can We Solve the Mind--Body Problem?." Mind (1989): 349-366.

there should be a problem at all[58] – we can examine each thought experiment and see what type of problem they can be reduced to, what relationships they describe.

Rephrasing the experiments

As we noted earlier, LM's conception makes it especially relevant to analyze the problem of consciousness in the terms of machines. Taking it as a reference, we will rephrase the other experiments in LM's structure: if an observer walks into a mill-sized machine, what will he find (see table 1)? Next, visualization charts of each situation, based on the model of figure 1-0, will expose the mind-body dichotomy as implied by each of the experiments, and will also serve as a support to bring forward how the newly expressed experiments can be divided among two groups with respect to their conclusions (see figure 1). This will in turn help us identify the specific aspects of consciousness on which the experiments focus.

LM	The observer walks into a mill-sized machine that has thoughts, sensations, and perceptions. He can observe all the mechanisms but fails to observe the machine's perceptions.
BAB	The observer walks into a giant bat. The anatomy of the animal offers nothing equivalent to the subjective experience of flying or echolocating. The observer cannot know "what it feels like" to be the bat.
MR	The observer is imprisoned in a black and white machine that recreates every step of human vision, so he can directly observe the mechanisms that are involved in details, and from there he is taught every physical fact about vision and colors. He knows what "red" is, but he will not be able to experience "redness" until he leaves the mill.
TZ	The observer walks into Leibniz's mill, but this time, the mill has no qualia. The observer will not be able to notice the difference.
CN	The observer walks into a giant machine where all the brain functions are recreated through Chinese citizen sending radio signals. The machine is not conscious because it lacks qualia.
CR	The observer, who does not speak Chinese, maneuvers the mechanisms of a giant machine to make it have a conversation in Chinese. He will be able to observe the symbols, but he will not understand the conversation.

Table 1. Reformulations of the experiments based on the model of Leibniz's Mill.

Leibniz's Mill. In LM, when the observer (figure 1, the eye), enters the mill, he can look at the thinking, feeling and perceiving machine's mechanisms (figure 1-1, right arrow), but he fails to witness its thoughts, sensations, or perceptions (figure 1-1, left arrow).

[58] Searle, John R. Intentionality: An essay in the philosophy of mind. Cambridge University Press, 1983: p.15.

Being a Bat and Mary's Room. Two thought experiments share the structure of figure 1-1, BAB and MR. A way of rephrasing BAB in LM's terms is to state that by visiting the cerebral system of a giant bat, while the observer will have access to each corner of the bat's brain (figure 1-2, right arrow), he will not be able to experience what it feels like to fly or to echolocate (figure 1-2, left arrow). In the case of MR, Mary is impersonated by the observer, and her room, by a black and white machine that duplicates every step of human vision. The observer can look at the mechanisms that are involved in great details (figure 1-3, right arrow), and from there he is taught every physical fact about vision and colors, but he will not be able to experience "redness" until he leaves the machine (figure 1-3, left arrow).

Philosophical Zombies. PZ, as well as the two remaining experiments, follows a different scheme. In fact, PZ can be translated in an inversed LM. While Leibniz's experiment depicts the construction, *ex nihilo*, of a conscious machine where one will fail to observe the physical mechanisms of consciousness, the zombie is a human whose consciousness was taken away, and an observer will fail to see anything different about that person. Leibniz's starting point is the opposite of the PZ experiment. To rephrase PZ in LM's terms: the observer enters a zombified machine, he is able to observe all the mechanisms associated with consciousness (figure 1-4, right arrow), and he is doomed to fail in witnessing the machine's qualia because it was never there (figure 1-4, left rectangle).

Chinese Nation and Chinese Room. Two thought experiments share the structure of figure 1-4, CN and CR. A simple approach to build CN on LM's structure is to replace the physical components of the mill by radio signals exchanged by Chinese citizens. The result in this case is that by building a machine where all the brain functions are performed by Chinese citizen exchanging radio signals, the machine is not conscious because it lacks qualia. Therefore, while walking inside this gigantic factory, the observer might manage to intercept radio signals or to deduce their functions (figure 1-5, right arrow), but he will never observe the machine's qualia which is missing (figure 1-5, left rectangle). One way to recast CR after LM is by asking the observer, who does not speak Chinese, to maneuver the mechanisms of a machine to make it have a conversation in Chinese. In this scenario, the machine does not understand because semantics is not inherent to the mechanisms. Consequently, the observer will be able to see the symbols (figure 1-5, right arrow), but the process he will follow is not constitutive of neither their meaning nor their intentionality, which will therefore never appear (figure 1-5, left rectangle).

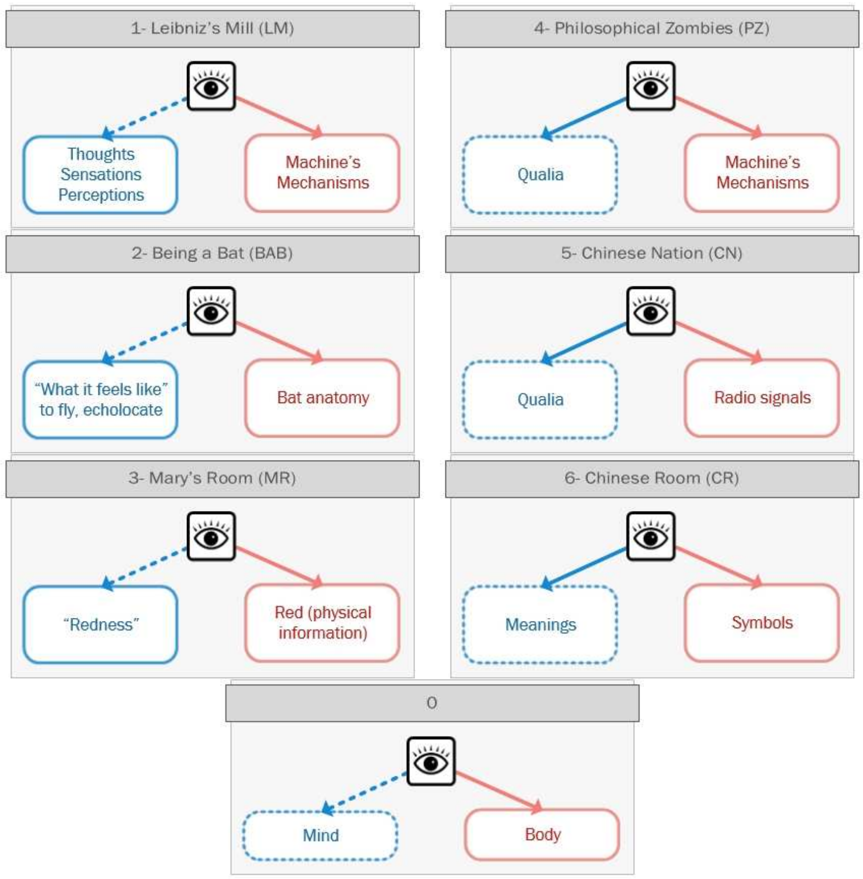

Figure 1. In 1-3, the external observer (the eye) cannot observe the "mind" level because of a missing link; in 4-6, the "mind" level is missing. In either case (0), the observer sees only the "body" level.

All experiments rest on the fact that it is impossible to observe the conscious stream from an external point of view, regardless of whether the machine effectively has one or not (see figure 1-0), thereby leading to the *other minds problem*: how to know if anything has a conscious mind outside of ourselves?

Besides, the mind-body division initially proposed by Leibniz, with an observable physical level on one side and an imperceptible mental level the other side, is still of relevance today; and we see that it can be applied to recent conceptions of the problem. Under this unified perspective, the contemporary experiments appear as more specific versions of LM, centering the mind-body problem on qualia (subjective experience) or semantics (meanings and intentionality), each one stressing out a different aspect of consciousness (see figure 2).

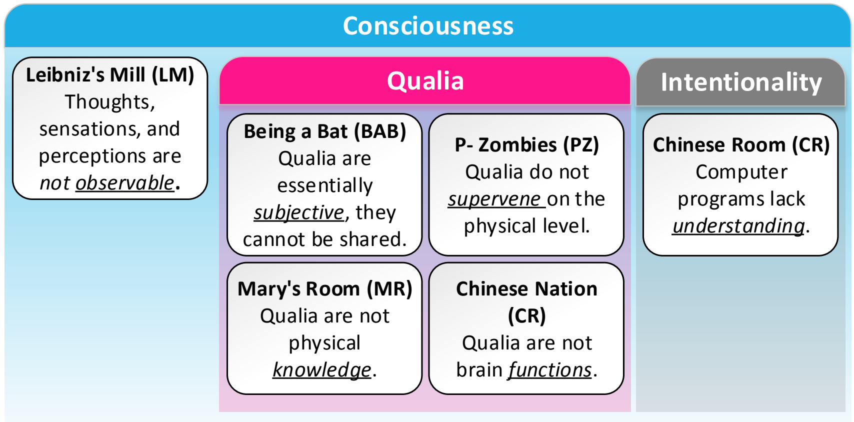

Figure 2. While LM addresses consciousness globally, the other experiments refine the notion of consciousness, focusing either on qualia or intentionality.

Identification of two problems and their complexity

Computational complexity is the evaluation of problems in terms of difficulty. The complexity of a given problem corresponds to the amount of resources (generally, time) needed by the most efficient algorithm that solves the problem. The theoretical process of complexity research is to first find the most efficient algorithm to solve a given problem and then to calculate the complexity of this algorithm. In practice, computational complexity usually is assessed by reducing a given problem A to another problem B for which the most efficient algorithm is already known, which is noted $A <_m B$. To operate such a reduction one must find a mapping of all the instances of A, i.e. all the possible inputs of A, to instances of B. This mapping is a computable function that converts A-instances into B-instances, making it possible to solve any A-instance by first converting it into a B-instance, and then running the known algorithm for solving B. From this process we learn at least two things: first, A is solvable, second, A is not harder to solve than B.

Group A - Figure 1 What is the nature of qualia?	**Group B – Figure 2** How do the mental and physical world interact?
LM	TZ
BAB	CN
MR	CR

Table 2. The experiment diverge on their approach of the mind-body problem.

The translation and representation of thought experiments on consciousness resulted in in two groups of problems (table 2). Group A involves experiments that are designed to bring out the existence of qualia, or subjective feeling of experience. Group B, on the other hand, comprises experiments that inquire non-physical properties as well as their relations with the material world by inferring from their circumstantial absence. The problem posed by group A is “What is the nature of qualia?” (ontology), while the one raise by group B is

"How do the mental and physical world interact?" (causality). Thinking in terms of set theory, problem A involves defining whether there are two distinct sets, the physical set P, and the mental set M, and if they somehow overlap. On the other hand problem B consists in defining F, a set of functions f associating elements from P with their image in M, and G, a set of functions g associating elements from M with their image in P (see equation 1 for a formal definition of F and G).

$$F = \{f_{i \in \mathbb{N}}: P \to M, f_i(p) \to m, p \in P, m \in M\}$$
$$G = \{g_{i \in \mathbb{N}}: P \to M, f_i(p) \to m, p \in P, m \in M\}$$

Equation 1. Formal definition of F and G.

There are only four possible solutions to problem A, illustrated on

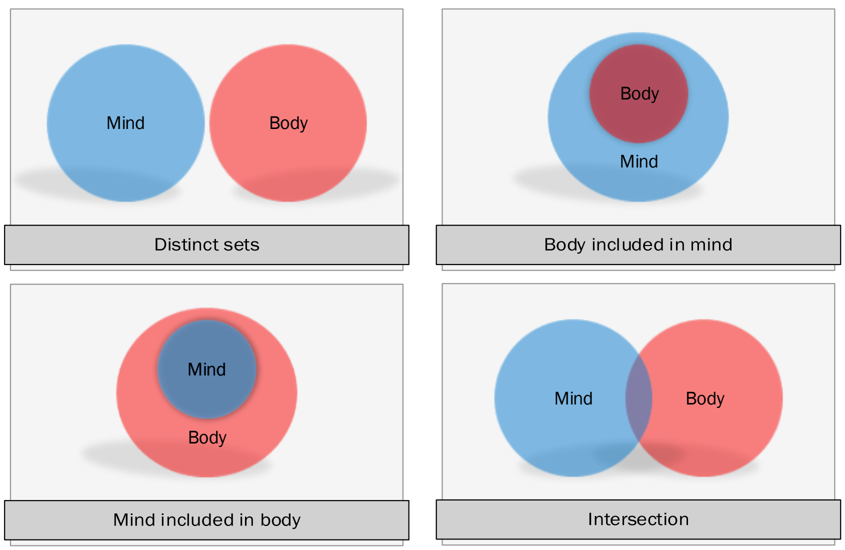

Figure 3 Four possible solutions to the ontological question.

figure 3, whereas for problem B, the number of solutions depends on the cardinality of each set, expressed on equation 2.

$$\text{Max-Card}(F) = \text{Max-Card}(G) = 2^{Card(P) \times Card(M)}$$

Equation 2. The number of possible solutions to problem B is the total number of possible mappings from one set to another. This number grows with the cardinality of each set.

If we knew every element of the world as we tried to solve both problems, they would become:

Problem A: for each element, does it belong to P? Does it belong to M?

Problem B: for each element from P, is there a relation that maps it to at least one element from M? For each element from M, is there a relation that maps it to at least one element from P?

To theoretically solve problem A, we need a test that will certify whether an element belongs to P and/or M. For problem B, we need to be able to determine whether a relation between P elements and M elements is relevant or not, but get to there, we first need to know which elements belong in P, and which ones belong in M: we need the same test that was used to solve problem A. Therefore if we can solve problem B, we can solve problem A, we have $A <_m B$. What that means is that in theory, if we had access to every element necessary, thought experiments of group B (PZ, CN and CR) would be harder to verify than those of group A.

In the end, all these experiments are compatible with a division between the mental and the physical, but they do not draw the line at the same level. Leibniz's definition of the mental realm is vague, while most contemporary views have identified qualia as the essential non-physical component of consciousness, with the exception of CR for which intentionality is the real problem.

4. Machine consciousness beyond self-awareness

These thought experiments emphasize perceptions, thoughts, feelings, qualia and intentionality as essential components of consciousness. Apart from these components, awareness[59] and the sense of self[60] are also found in consciousness literature. Those traits are not all equal when we talk about machine consciousness. *Perceptions*, when considered as plain sensory inputs, can be provided through the appropriate devices to a machine[61]. As for *thoughts*, *emotions*, *feelings*, we can easily imagine how they could be programmed, translated in algorithms[62]. *Awareness* and

---

[59] Rosenthal, David. "Higher-order awareness, misrepresentation and function." Philosophical Transactions of the Royal Society B: Biological Sciences 367.1594 (2012): 1424-1438. Schwenkler, John. Levine, Joseph. "Conscious awareness and (self-) representation." Self-representational approaches to consciousness (2006): 173-198. Anderson, Michael L., and Donald R. Perlis. "The roots of self-awareness." Phenomenology and the Cognitive Sciences 4.3 (2005): 297-333.

[60] Bermúdez, José Luis. The paradox of self-consciousness. MIT Press, 2000. De Vignemont, Frederique. "Habeas corpus: The sense of ownership of one's own body." Mind & Language 22.4 (2007): 427-449. de Haan, Sanneke, and Leon de Bruin. Kriegel, Uriah. "Consciousness and self-consciousness." The Monist (2004): 182-205.

[61] Robertsson, Linn, et al. "Perception modeling for human-like artificial sensor systems." International journal of human-computer studies 65.5 (2007): 446-459. Irie, Robert Eiichi. Robust sound localization: An application of an auditory perception system for a humanoid robot. Diss. Massachusetts Institute Of Technology, 1995. Russell, R. Andrew. "Survey of robotic applications for odor-sensing technology." The International Journal of Robotics Research 20.2 (2001): 144-162.

[62] Breazeal, Cynthia. "Emotion and sociable humanoid robots." International Journal of Human-Computer Studies 59.1 (2003): 119-155. Pierre-Yves, Oudeyer. "The production and recognition of emotions in speech: features and algorithms." International Journal of Human-Computer Studies 59.1 (2003): 157-

especially the *sense of self* are currently central topics of interest in robotics and many studies and experiments in the past decade have been dedicated to solve these particular issues, with encouraging results[63]. A machine built with all those components would pass the Turing test[64], however, what we can conclude from part 2 is that such a machine will not be called conscious as it lacks the two most problematic components of consciousness: *qualia* and *intentionality*[65].

Russellian grinders and zombie machines

The presence of qualia may be seen as an anthropocentric requisite to acknowledge consciousness in a machine. As a matter of fact, PZ emphasizes that there is no apparent need for any conscious system to have qualia, apart to mimic human consciousness. If qualia were the only missing part of consciousness in a machine, this machine would be as conscious as a zombie. Zombies are intimately convinced of their consciousness, indistinguishable from any human being, and animated with the same complex system of perceptions and feelings, and capable of the most sincere empathy and the most brilliant intelligence. Ultimately, this begs two questions: first, are qualia necessary for consciousness? Second, are they sufficient?

According to Russellian views of the world[66], phenomenal properties are fundamentally embedded in the physical level, just as mass or electrical charges. Under this assumption, qualia are associated to every kind of physical interaction, at every level. We will all agree that an electrical coffee grinder is not conscious, however isn't there something it is like for the blades to crush the coffee beans, for the circuitry to be plugged in and have the electricity flowing, for the buttons to be pushed? There will be no

---

183. Dias, João, and Ana Paiva. "Feeling and reasoning: A computational model for emotional characters." Progress in artificial intelligence. Springer Berlin Heidelberg, 2005. 127-140. Oatley, Keith, and Philip N. Johnson-Laird. "Towards a cognitive theory of emotions." Cognition and emotion 1.1 (1987): 29-50.

[63] Chella, Antonio, Marcello Frixione, and Salvatore Gaglio. "A cognitive architecture for robot self-consciousness." Artificial Intelligence in Medicine 44.2 (2008): 147-154. Takiguchi, Toshiyuki, Atsushi Mizunaga, and Junichi Takeno. "A study of self-awareness in robots." International Journal of Machine Consciousness 5.02 (2013): 145-164. Gorbenko, Anna, Vladimir Popov, and Andrey Sheka. "Robot self-awareness: Exploration of internal states." Applied Mathematical Sciences 6.14 (2012): 675-688. Kawamura, Kazuhiko, et al. "Development of a robot with a sense of self." Computational Intelligence in Robotics and Automation, 2005. CIRA 2005. Proceedings. 2005 IEEE International Symposium on. IEEE, 2005.

[64] Turing, Alan M. "Computing machinery and intelligence." Mind (1950): 433-460.

[65] For a discussion on the subject, see Chalmers 2004 "The representational character of experience"

[66] Chalmers, D. "Panpsychism and panprotopsychism." The Amherst Lecture in Philosophy.

witness, no recording, no possible reminiscence, but the event has occurred nonetheless, physical entities have undergone physical experiences; they just lack an ascendant structure to interpret anything of what happened. Therefore, unless we consider that under these conditions the coffee grinder is conscious, qualia alone is not sufficient to account for consciousness. Furthermore, under a Russellian perspective, if the coffee grinder was also programmed to display a conscious behavior, the way robots are conceived today, the program which grants the conscious states to the machine would not have access to any of these inherent phenomenal properties, all the while, paradoxically, each one of the machine's physical states would really be associated with a specific corresponding quale. Without a clear understanding of how these properties give rise to biological consciousness as we know it, it is hard to see how we could provide the machine with adequate tools to "read" these properties. Russellian solutions to the mind-body problem let little hope to achieve machine consciousness in the near future, unless we reconsider the necessity of qualia and consider zombie machines conscious.

Abstraction layers and intentionality

Intentionality can be understood as an abstract relationship between an object and its meaning which is entirely dependent on the observer and the situation. The only physical connection between the letters that the reader is looking at and the reader's brain is relayed by the photons hitting the reader's retina. The general pattern of a word and its meaning are associations that are made inside our brain, which physically correspond to specific neural activities[67]. Therefore, as suggested by neuroscientific theories of consciousness[68], intentionality can be physically described as a network of neural connections inside the brain, which are in theory reproducible through adequately programmed machines. In fact, it is somewhat fascinating that while some neuroscientists are puzzled by the fact that there is no fundamental difference between the neural activity responsible for distinct processes[69] (i.e., neurons firings) from a computer scientist's perspective, this similarity is not a surprise: at the programming level, all data structures are essentially the same - networks of nodes - they only vary on their connections and the information they contain.

---

[67] Bookheimer, Susan. "Functional MRI of language: new approaches to understanding the cortical organization of semantic processing." Annual review of neuroscience 25.1 (2002): 151-188.

[68] Edelman, Gerald M., and Giulio Tononi. A universe of consciousness: How matter becomes imagination. Basic books, 2000.

[69] Edelman, Gerald M., and Giulio Tononi. A universe of consciousness: How matter becomes imagination. Basic books, 2000: 17-18.

However, we may also choose to consider intentionality as an emergent property, and not as the causal result of neuronal activity. That is, instead of considering the effective connections that are established inside the brain, we decide to only take into account a higher-order level of comprehension. In this sense, intentionality is inextricably intertwined with semantics, being the link between any actual object of thought and its meaning. At first, with a computer program, meanings seem to be encoded with binary bits, but this is only looking at the machinery, this is like stopping the observation at the neuronal level. Reducing computers to electrical signals or binary bits or Turing machines is missing some undeniable consequences of their physical existence, for they also manifest higher-order properties. Distinctions of this kind are well known to computer scientists who resort to *abstraction layers* to separate the different levels that interact in a computer process. For instance, the computer's architecture can be decomposed roughly in three layers: the hardware layer, the software layer, and the graphical user interface (GUI) layer. Just as with LM, by observing the physical components that operate the computing, there is nothing like a "window". And a program never embeds a window itself, only its attributes. The window which appears on the screen is located on the upper layer, known as the GUI. If we think about the understanding process as an abstraction layer, then meanings and intentionality constitute the upper layer, underneath them are elementary processing like shape recognition, and after those come databases and algorithms. Understanding is a very complex process that we are not even sure to fully grasp for humans, therefore it developing it in computers is a premature endeavor. What is sure is that it will not consist in feeding the machine with a dictionary, but providing it with the right tools to come up with its own interpretation.

5. Conclusion

Six major thought experiments on consciousness, five of them published in the last four decades, can be translated into thought experiments on machine consciousness. We could expand this approach to other thought experiments, to include further related issues such as the sense of self. We classified these experiments according to the problem they pose (ontological or causal), and inferred that the causal problem is more complex than the ontological one. The complexity analysis could be completed by reducing the identified problems to computational problems the complexity of which we know. Ultimately, these experiments reveal the problematic nature of qualia and intentionality, two characteristics of consciousness that offer fascinating perspectives on machine consciousness, beyond current self-awareness research.

For intentionality, representationalism is a serious research avenue that additionally serves as a basis for synthetic phenomenology, an attempt at replacing qualia.