From Mechanical Brains to Philosophical Zombies

Nathan Ensmenger
School of Informatics & Computing
Indiana University

Mind-mills, intuition pumps, and thinking machines

In 1714 the mathematician and philosopher Gottfried Leibniz published *The Monadology*, a short text outlining a new theory of matter that solved, or at least proposed to solve, the mind-body problem. The mind-body problem, most often associated with Rene Descartes, is one of the enduring conundrums in metaphysics: namely, how do purely material process produce spiritual (or at least intellectual) phenomenon? When (and how) does the physical body acquire its immortal soul (or, for modern philosophers, self-awareness)? Or, to put it in more earthy language, how to we more from the bloody mass of cells in a brain to the epiphenomenon of the conscious mind?

Descartes, of course, solved the mind-body problem by proposing a dualist ontology in which mind and matter were fundamentally distinct substances. Leibniz, on the other hand, was a monist: he believed that there was but one type of matter but that individual elements of this universal matter (which he called monads) could be infused with different qualities. In this sense, individual monads would contain within them the "programming" that would allow them to serve specific purposes: some monads would be essentially physical, and others essentially mental.[1]

In developing his argument for a monist metaphysics, Leibniz engaged in an

1. It is dangerous, of course, to impose on the past the categories and terminology (like "programming") borrowed from the present. This is a particular problem for historians of computing, given the ways in which computationalist discourse has become so hegemonic in so many disciplines, from biology to psychology to meteorology to ecology to economics. Nevertheless, in the case of Leibniz it does not seem inappropriate. In addition to his imagined thinking machine, Leibniz also believed firmly that binary arithmetic was the key to understanding the mind of God. See , for example, his *Mira numerorum omnium expressio* (1696).

extended thought experiment in which he imagined the human body as a kind of machine:

> "And supposing there were a machine so constructed as to think, feel, and have perception, it might be conceived as increasing in size, while keeping the same proportions, so that one might go into as into a mill. That being so, we should, on examining its interior, find only parts which work one upon another, and never anything by which to explain a perception..."[2]

The metaphysical conclusion that Leibniz drew from this thought experiment (which was that since perception could not be found in the whole, it must be contained in each and every of its individual parts) is not particularly important, at least for my purposes here. Speculation about mechanical minds and/or bodies was not unique to Leibniz, and has a long history within the Western intellectual tradition.[3] In some cases, the idea of artificial life served primarily as a vehicle for metaphysical speculation; in others, actual attempts were made to realize their existence. In this sense, Leibniz serves only as a useful case study, an indicator of a larger trend. But the mind-mill that Leibniz mentally constructed is remarkably similar to two of the paradigmatic thought experiments of artificial intelligence and cognitive science, namely the Turing Test and the Chinese Room Argument. This historical continuity is significant. In addition, the Leibniz version of the experiment is notable in that it served as the particular target of an attack on the epistemological value of thought experiments in general led by the philosopher and cognitive scientist Daniel Dennett, who (drawing on another curiously mechanical model of cognition) dismissed them as mere "intuition pumps," cognitive sleights of hand that distracted the rational mind with intriguing but impossible fantasies.[4] This tension between what *is* and what *could be* is a distinctive feature of the debate about the role of experiments of all kinds in cognitive science and artificial intelligence.

---

2. Gottfried Wilhelm Leibniz Freiherr von and Nicholas Rescher, *G.W. Leibniz's Monadology* (University of Pittsburgh Pre, 1991).

3. Phil Husbands, Owen Holland, and Michael Wheeler, *The mechanical mind in history* (The MIT Press, 2008); J Riskin, ed., *Genesis Redux: Essays in the History and Philosophy of Artificial Life* (University Of Chicago Press, 2007); Minsoo Kang, *Sublime Dreams of Living Machines: The Automaton in the European Imagination* (Harvard University Press, 2011).

4. Daniel Dennett, *Consciousness Explained* (Litle, Brown / Company,1991, 1991).

My interest in thought experiments in the history of artificial intelligence stems from a very practical set of problems I have been working on in the history of computer software. To a certain degree, software occupies the same relationship to the electronic digital computer as the mind does to the body. Like a thought in the mind, software in a computer is invisible, intangible, and ephemeral. The complexities of software are often emergent, meaning that they cannot easily be explained in terms of constitutive elements. Whereas computer hardware (and physical bodies) can be dissected, measured, and compared, software is difficult to isolate from its operational context. Software exists simultaneously as concept, text, technology, and practice. One of the reasons that it is so difficult to open up the "black box" of software to historical analysis is that software is at once a created object and an idealized abstraction. Software is clearly a technology, but it is a technology without an artifact. Software is hard to get your mind around, in part because it is so difficult to actually take in hand. And yet as with any technology, how software systems get built, and by whom, turns out to be enormously important from an historical perspective.

I have been thinking a great deal recently about the tangibility — or lack thereof — of software. In a recent paper published in *Social Studies of Science* I explored the role of computer chess as the so-called "drosophila" of artificial intelligence.[5] Like the mind-mill and the Turing test, computer chess began as a thought experiment. Believing that the ability to play good chess was a sign of general intelligence and strategic thinking, early AI researchers posited that a computer that could play chess would therefore be de facto intelligent. Turing himself believed this to be true; so did Herbert Simon, Allan Newell, John McCarthy, and most of the rest of the early AI community. But rather than leave computer chess as a purely metaphysical exercise, they actually set about building chess computers. My SSS paper addressed the way in which the specific way in which the computer chess thought experiment was implemented (using an algorithmic approach called minimax) came to structure the research agenda in AI for decades. As it turned out, the fact that a computer could play strong chess was not so significant; *how* that computer played chess was — enormously so. If the social turn in the history of science has told us nothing else, it is that the choice of experiments matters.[6] This is equally true of thought experiments.

5. Nathan Ensmenger, "Is Chess the Drosophila of AI? A Social History of an Algorithm," *Social Studies of Science* (2012).

6. Thomas S Kuhn, *The structure of scientific revolutions* (Chicago University of Chicago Press, 1962).

In this paper, I will provide a brief history of the role of thought experiments in the history artificial intelligence, will explore the recent turn towards biological metaphors in these experiments, and will suggest ways in which the study of such experiments illuminates key issues in the larger history of computing. I will argue that thought experiments play an especially important role in the history of these two disciplines, in part because of the unique ability of researchers in these disciplines to actually construct the machines that were once only dreamt of by philosophers. As the ability of artificial intelligence researchers to actually build machines that exhibit intelligent behavior has improved dramatically, cognitive scientists and philosophers of the mind have increasingly adopted biological, rather than a mechanical, metaphors through which to explore the mind-body duality. In fact, in recent years the dominant thought experiments have become downright gothic in their sensibilities: from the Swampman to the Blockhead to Twin Earthlings to the Philosophical Zombie, these experiments *appear* to have abandoned their traditional focus on paradoxes of the "thinking machine."

Turing's Test

While the Leibniz mind-mill might arguably be seen as the first thought experiment in artificial intelligence, the most famous by far is the Turing test. First introduced by Alan Turing in his 1950 paper "Computing Machinery and Intelligence," the Turing test is an adaption of a common parlor entertainment known as the "imitation game." In the imitation game, the designated judge attempts to determine which of two concealed participants is a man, and which a women, solely on the basis of their written responses to the questions he poses. In Turing's adaptation, the judge is charged with determining not the sex but the rather humanity of his unseen interlocutors, one of which is a machine. If the judge cannot tell the difference between man and machine, then the latter is assumed to be exhibiting intelligent behavior. (Whether or not this also means that the machine is also therefore self-conscious is left ambiguous, although the popular interpretation of the Turing test is that it tests for the existence of a true artificial intelligence.)[7]

The Turing test has been widely criticized for its deficiencies, both practical and metaphysical.[8] For example, although Turing specified that the test would

7. Alan M. Turing, "Computing Machinery and Intelligence," *Mind* (1950).

8. Mark I. Halpern, "The Trouble with the Turing Test," *The New Atlantis: A Journal of Tech-*

be successful if the machine could fool "the average interrogator," he provided no details on what he meant by "average." Some people are more easily deceived that others, or at least in certain contexts. For example, it is estimated that more than 70% of the sex chat on the Internet has at least one chat-bot as a participant. Within the narrow genre conventions of online sex talk, at the very least, the standards for acting "plausibly human" are apparently not particularly rigorous... In any case, human interrogators are also notoriously susceptible to falling for irrelevant "tricks," such as typing errors or juvenile insults, that mimic human, but not intelligent, behavior. And the Turing test does not even attempt to address the deeper philosophical question of whether "simulated intelligence" is the same thing as "real intelligence."

Within the professional AI community, the Turing test is generally dismissed as being irrelevant, a "publicity stunt" with no real value to serious academic research.[9] This casual dismissal of one of the field's most durable and memorable thought experiments, however, ignores is significant and lasting impact on the public perception of the fields of artificial intelligence and cognitive science. It may not be a useful as a tool for measurement, but as a rhetorical device is has proven extraordinarily effective.

What interests me most about the Turing test and other thought experiments in artificial intelligence and cognitive science is their ambiguous relationship to actual, real-world machines. Whereas classical gedanken experiments are by definition never intended to be attempted — Maxwell's Demon is clearly a creation of a lively imagination, only slightly less fantastic than the idea that Schroedinger's cat (or any cat, for that matter) would ever deign to be locked in a box by a mere physicist — thought experiments in artificial intelligence often blur the boundary between idealized logical demonstration and actual real-world technological development. On the one hand, the Turing test can be treated as one in a long line of thought experiments aimed at exploring the mind-body problem; on the other hand, Turing developed his test not as an exercise in abstract metaphysics, but because he truly believed that intelligent machines could and would soon be constructed. When Leibniz speculated that the organic body was, in fact, "a kind of divine machine, or natural automaton," he was thinking metaphysically; when a cognitive scientist describes the brain as a computer, he or she is thinking

---

nology and Society (2006).

9. It should be noted, however, the annual Loebner competition, which offers a $100,000 prize to the winner of a modified Turing test, is still hotly contested, and is widely covered in the popular media

technologically.

The blurriness between abstraction and reality that characterizes artificial intelligence is exemplified, and perhaps even enabled, by Turing's most influential thought experiment. In his 1937 paper "On computable numbers, with an application to the Entscheidungsproblem" Turing, in order to solve a problem in theoretical mathematics, invented an imaginary machine consisting of a paper tape and a read/write device that, provided with an appropriate set of symbolic instructions, could perform mechanical computations.[10] This "Turing Machine," as it is called, could itself be represented as a series of symbols. These symbols could in turn become the instruction set for yet another Turing Machine, allowing for the creation of a universal Turing Machine, which could compute .... well, anything that was computable. Although the universal Turing Machine has never been constructed in the form that Turing imagined it (which is infeasible from a practical perspective), it has become the theoretical basis for all of modern computer science. A computer is a machine that is logically equivalent to a universal Turing Machine — regardless of how (or even if) it is materially constructed. Such machines are often made out tubes, transistors, or metal-oxide semiconductors grown on a silicon wafer, but they could easily also be biological in nature.[11] The human brain is but one incarnation of a universal Turing Machine, according to computationally minded cognitive scientists, a computer *in vivo* rather than *in silico*.

By providing a logical abstraction of the computer divorced from any particular implementation, the concept of the universal Turing Machine encourages a dualistic distinction between hardware and software. Since all computers are, by definition, universal Turing Machines (and therefore logically equivalent), in theory software written for one computer could run on any (and every) computer. This has obvious implications for the mind/body problem. Indeed, for an extreme computationalist, the mind is simply the software of the brain, meaning that it could (again, in theory) be downloadable to a hard drive. Think about that the next time your computer crashes...

10. A M Turing, "On computable numbers, with an application to the Entscheidungsproblem. A correction," in *Proceedings of the London Mathematical ...* (1938).

11. Turing machines have also been constructed out of Legos, Tinker Toys, and a deck of Magic the Gathering cards, among other things.

Does the Chinese Room really speak Chinese?

Despite its many deficiencies, the Turing test has survived as the most recognizable symbol of what is usually referred to as the the Strong AI program. The central assumption of Strong AI is that machines can be intelligent: when and how such machines will be built, or how we will recognize them for what they are, are open questions for Strong AI proponents, but the assumption is that they will eventually, perhaps inevitably, be answered. The Strong AI program is not universally accepted, even within the AI community. In his 1980 paper "Minds, Brains, and Programs," the philosopher John Searle constructed an influential and compelling critique of the strong AI program built around a central conceit strikingly similar to that outlined several hundred years earlier by Gottfried Leibniz.[12] Searle's version was called the Chinese Room argument. In the Chinese Room experiment, an individual (or group) is locked into a room and given a set of instructions (an algorithm) to follow. Slips of paper are slid under the door, the individuals in the room apply the appropriate algorithm, and return under the door some transformed version of the original source data. Imagine that in applying this algorithm, asked Searle, that the person (or people) in the room were actually translating from Chinese into English. Would their ability to provide identical results to what is generally considered a cognitive function (language translation) mean that actual cognition had occurred? Or, in other words, did the ability of the room (and its residents) to successfully translate Chinese mean that it (or they) actually understand Chinese? Of course not, concluded Searle: purely functionalist accounts of human intelligence most therefore be insufficient.

There have been a number of compelling responses made to Searle. The three most significant are the Systems Reply, the Virtual Mind Reply, and the Robot Reply. The essential argument of the Systems Reply is that, while the man in the room might not understand Chinese, the overall system (which includes the man, the room, the instructions, and any intermediate conversions created during the implementation of the algorithm) does. The Virtual Mind Reply makes a similar argument, although rather than suggesting that the system understands Chinese, it situates this knowledge in some emergent virtual entity. The Robot Reply is an adaptation of both earlier replies to take into account the recognition that to truly understand Chinese (or any other language), the system (or virtual mind) must have some conception of the semantic relationship between objects

12. John R Searle, "Minds, brains, and programs," *Behavioral and brain sciences* (1980).

and their referent: that is to say, it must understanding between the relationship between the word "pig" and an actual pig. The best way to accomplish this, the Robot Reply argues, is to build a robot that can perceive and interact with the world. Such a robot, or at least the computer that controlled such a robot, would truly be intelligent, according to the proponents of the Robot Reply.

All three of these replies are articulations of what is called the functionalist approach to artificial intelligence. Whereas the Turing test reflects a purely behavioralist metaphysic (if it acts intelligently, it is therefore intelligent), the functionalist approach suggests a relationship between state of a system and what it does (intelligence is what intelligence does). The functionalist model of pain, for example, is that it is caused by damage to the body, is located in a body-image, and is aversive.[13] Where a functionalist would differ most from a biological naturalist (such as Searle) is in their willingness to allow for a multitude of different mechanisms for accomplishing cognitive functions. That is to say, a mechanical brain with the appropriate structure (such as the machine imagined in the Robot Reply) would be capable of intelligence. The functionalist is also a dualist. Mind and brain are distinct and separable phenomenon. A behavior naturalist, on the other hand, is a monist: all higher-level mental processes are caused by (and are inseparable from) lower-level neurobiological processes in the brain. Mind and brain are indistinguishable.

Zombie Attack!

For several decades Searle's Chinese Room Argument has served as the paradigmatic thought experiment in cognitive science and the philosophy of the mind, capturing neatly the central metaphysical conundrum that has defined the mind-body problem since the days of Leibniz and Descartes. More recently, however, a remarkable shift has occurred within the discourse these two disciplines. In the past decade, the literature in these fields have become infested with zombies.[14]

13. An excellent discussion of this distinction can be found in the Stanford Encyclopedia of Philosophy, available online at \url{http://plato.stanford.edu/entries/chinese-room/}.

14. P Skokowski, "I, zombie," *Consciousness and Cognition* (2002); Robert Kirk, *Zombies And Consciousness* (Oxford University Press, USA, 2005); J Connelly, "On Siamese Twins and Philosophical Zombies: A New Reading of Wittgenstein's 'Private Language Argument'," *Kulturen: Konflikt-Analyse-Dialog Cultures: Conflict- ...*; Andrew Bailey, "Zombies, Epiphenomenalism, and Physicalist Theories of Consciousness," *Canadian Journal of Philosophy* (2006); S Bringsjord, "The zombie attack on the computational conception of mind," *Philosophical and Phenomenological Re-*

The so-called "Zombie Attack" on the computational conception of mind began, surprisingly, with another seminal thought experiment proposed by John Searle. In Searle's original formulation, the zombie in question is the philosopher himself who, faced with the physical deterioration of his brain, has parts of it replaced with silicon-based work-alikes. Eventually there is nothing left in his cranium but a computer. With the brain gone, asks Searle, what is left of the mind? One possibility is that the newly reconstructed philosopher is exactly identical to the original, a perfect replica, proof of the functionalist conception of the mind as software and the brain merely a meat-machine implementation of an idealized computer. Another alternative is that the procedure would leave the philosopher as a non-cognitive vegetable. The third, and most interesting, possibility is that the philosopher would retain his ability to function in the world — that is to say, his external behavior and capabilities would remain the same — but that he would lose his inner cognitive self-awareness, his consciousness, would disappear. The philosopher would have become a zombie, and zombies, according to Searle, do not think, and do not have minds. The might be functionally equivalent to a human, but are metaphysically entirely distinct.[15]

Although Searle first launched his zombie attack in the early 1990s, it take another decade for zombie fever to assume epidemic proportions. In the intervening period, the defining characteristics of the philosophical zombie, or p-zombie, would be subtly but significantly reconceptualized. Searle's philosopher started out as human, and it was only when crucial physical components of his brain/mind were eliminated that he became a zombie. This in fact was the point of the thought experiment, to establish the irreducible relationship between the biological brain and the metaphysical mind.

The philosophical zombie, as it was reimagined by subsequent philosophers, was not a human without a brain, but rather a human without a mind. This is not a trivial distinction. Searle's zombie was a human body with a computerized brain. The p-zombie as it came to be understood was simply a human body/brain that lacked qualia. Qualia is the internal and subjective component of perception — a fancy philosopher's term for consciousness. The p-zombie was not so much a damaged or degraded philosopher as the philosopher's perfectly identical twin (or perhaps clone), with the sole difference between the two being that one experiences qualia and the other does not. The p-zombie is a perfect micro-

search (1999).

15. John R Searle, *The Rediscovery of the Mind* (MIT Press, 1992).

physical duplicate of the original philosopher, molecule for molecule, that is also functionally equivalent. It is not the replacement of the biological brain with the mechanical computer that is the key difference between the philosopher and his analogous zombie, at least in this version of the thought experiment, but rather the presence (or absence) of qualia.

A blueprint of a car is not a car

The purpose of this paper is not to explore in any detail the vast and growing literature on philosophical zombies for practitioners in the fields of cognitive science and the philosophy of mind. It is clear that the p-zombie has proven an extraordinarily productive experimental technology within these disciplines, at least judging from the number of books and papers published on the topic. In terms of settling the mind-body problem once and for all, the p-zombie has turned out to be any more conclusive an experiment that Searle's Chinese Room or Leibniz's mind-mill. Searle mobilized his zombie in support of monism; David Chalmers equally convincingly argued that p-zombie demonstrates the dualistic nature of mind-body relationship.[16]

Instead, let me make three important points about thought experiments in general, and the philosophical zombie in particular, that I think are extremely relevant to the history of technology.

The first is that, despite the increasingly biological focus of these thought experiments (the p-zombie is striking but not unique in this respect; the Swampman and Twin Earthling experiments were also constructed around biological organisms), they are, at their heart, essentially machinic. The origins of the mind-body problem itself lies in the attempt to establish (or refute) the possibility of a purely material and mechanistic universe. For Leibnitz and others, the use of mechanical and industrial metaphors was meant to highlight the difference between the technological and the organic. For more recent metaphysicians, it is used to blur this imagined boundary. The philosophical zombie is not a biological alternative to a thinking machine. The p-zombie is simply the *perfect* machine. For Leibniz, the idea of mind-mill was manifestly absurd; its obvious impossibility was the essence of his argument. In the age of the smart machine, however, when computers at least appear to demonstrate intelligent (or at

16. David J Chalmers, *The Conscious Mind* (Oxford University Press, USA, 1997); A Cottrell, "On the conceivability of zombies: Chalmers v. Dennett" (1996).

least intelligence-seeming) behaviors, the metaphysical significance and rhetorical power of the mind/machine distinction is no longer so self-evident. The metaphysical questions of traditional philosophy threatened to be overwhelmed by the practical achievements of artificial intelligence researchers. Any objection to the functionalist and computational conception of the mind could be countered with the argument that, while perhaps the current generation of mechanical minds might not be truly intelligent, it was only a matter of time before they were demonstrably so. Technology trumped philosophy. **By inventing in the philosophical zombie the ideal thinking machine, the perfect functional equivalent to a human being, philosophers of the mind and philosophically-minded cognitive scientists were able to reclaim the initiative from the technologists.** No matter the improvements made to computer technology, the philosopher's machine would always take precedent. The debate about the mind-body problem could therefore return to its focus on more traditional metaphysical questions.

That being said, my second point is that, as was mentioned earlier, thought experiments in cognitive science and AI are different from those in other disciplines in that they can, in theory at least, actually be performed. There is always the risk, when speculating about some new thought experiment, that some developer somewhere will actually build it. This is indeed what happened in the case of computer chess: for several decades, the grand rhetorical claims of the artificial intelligence community that, "If one could devise a successful chess machine, one would seem to have penetrated to the core of human intellectual endeavor," were shown to be deeply and fundamentally flawed, at least in terms of the metaphysics of cognition, not because the AI researchers failed to accomplish their goals, but because they eventually did.[17] The ability of computers to play good chess turned out to have less to say about the computers than it did about chess. What was originally imagined as the ultimate test of general intelligence (human or machine) was gradually redefined as a much narrower exercise in pattern recognition (on the part of humans) and deep searching through a decision tree (in the case of computers). Despite their name, thought experiments are not actually meant to be tested. They might be revealed to have logical or conceptual flaws, but their outcome is, by design, not supposed to be surprising.

It is this ambiguous relationship between thought experiment and real experiment that brings me to my third and final point. Implicit in the computational

17. Allen Newell, J C Shaw, and H A Simon, "Chess-Playing Programs and the Problem of Complexity," *IBM Journal of Research and Development* (1958).

conception of the mind (as in the computationalist perspective more generally) is the assumption that (with a few important exceptions) any complex system can be modeled (and indeed implemented) as a Turing machine. This means that all of the thought experiments of the philosophers can, in principle, be transformed in actual experiments. This would seem to argue that long-standing metaphysical debates, such as the mind-body problem, might eventually be decisively solved. In fact, the computationalists might argue that, in theory, they have been already. But the growing literature on the history of software reveals the vast gulf between theory and practice in the computer sciences. In theory, the mind is the software of the brain, and thus the functionalist perspective is demonstrated to be correct. In practice, how the logical abstraction of a Turing machine becomes embodied and enacted in the physical world matters significantly. This is one of the great lessons that the history of technology has to offer the larger set of disciplines subsumed until the umbrella of science and technology studies. The materiality of things matters, even when we are talking about intangible artifacts like software systems. A blueprint of a car is not the same thing as a car. The translation of ideas into technologies, functions (like the mind) into embodiments like the brain (in this case, quite literally) is not a frictionless process. Thought experiments — or thought technologies, like the philosophical zombie — are constructed objects, despite their apparent lack of substance. And like all technologies, they have histories that are significant.