Expertise
and insight
for the future

Teemu Rytilä

Turing Machines and Intentionality

A Search for the Essence of AI

Metropolia University of Applied Sciences

Bachelor of Culture and Arts

Media

Thesis

20 May 2019

Author(s) Title Number of Pages Date	Teemu Rytilä Turing Machines and Intentionality: A Search for the Essence of AI 45 pages 20 May 2019
Degree	Bachelor of Culture and Arts
Degree Programme	Media
Specialisation option	Digital media
Instructor(s)	Juhana Kokkonen, Senior Lecturer

The history of artificial intelligence (AI) is characterized by the difficulty to provide a concise definition for what AI actually entails, and so at different times it has stood for different priorities and motivations. This work is predicated on the postulation that running through the history of AI is a common thread—an *essence*—which characterizes the different approaches taken to defining AI over the course of its history, and the main goal of this work is to elaborate on what that common thread is and why.

This work is primarily based on the ideas and works of Daniel Dennett, Douglas Hofstadter, Alan Turing and John Searle, and expands upon them in an attempt to provide a cohesive picture of AI as a *pursuit to create intentional systems,* where "intentional system" is defined as a system whose behavior is effectively predictable by attributing intentional states to it, such as beliefs, desires, and intentions. It is argued that intentional states are necessarily attributed rather than innate, and the utility of this attribution comes across through use of intentional vocabulary in descriptions of the systems that are treated as intentional.

In addition, various problems of philosophy of mind which relate to the question "can the human mind be simulated by a machine?" are discussed. These include, among other things, the implications of the Church–Turing Thesis on computability of the mind, Searle's Chinese room argument, and the philosophical zombie argument. Discussions on these topics act as a narrative thread to help establish the main thesis of this work in a logical and cohesive manner.

Keywords	AI, artificial intelligence, machine intelligence, Turing machine, Church–Turing Thesis, intentionality, simulation, philosophy of mind

Contents

1. Introduction

Artificial intelligence, or AI for short, is a pretty hot topic these days. You may have noticed. There is all this doom-and-gloom talk about AI taking over the world, fascinating philosophical discussions about whether or not the mind can be fully simulated by a machine, and then there is all the AI in your devices that you are told improves the experience in some way, even though you cannot tell what aspect of it is supposed to be intelligent—it is just something that invisibly helps with your daily life. Yet somehow the big picture stuff and the practical-but-invisible implementations in your devices are reconciled under the same umbrella term. But how?

Your phone's camera, for instance, may now come with an AI that, somehow, improves the quality of the photos you take. If you try to imagine what that actually means, you might envision some kind of rational actor that makes meaningful and intelligent decisions on your behalf—some discrete *being* that operates within the software on your phone. After all, one cannot just "improve" an image without first having a subjective idea of what an improvement *is*. So whatever is doing that must be showing some kind of intelligence… right? But on the other hand, you do not actually witness any of the reasoning that goes into those decisions. You just get the output.

AI has been subject to a lot of revisionism throughout its history. In its early days researchers wanted to use AI to figure out how the mind works. Many have argued that intelligence is not simply a matter of single inputs and outputs, which is what your phone's camera does. In the current landscape of software development, however, AI seems to be a thing unto itself, unshackled from its ties to our conception of the mind. It is a thing that does things for you and sounds good in marketing. Whether or not it actually is intelligent seems secondary, or even completely irrelevant.

What is the *essence* of this thing we call AI? What is the element that holds all the disparate conceptions of AI together? What can we call intelligent? These are some of the questions that will be tackled in this text.

This work deals primarily with philosophical theories, and as such is in itself a work of philosophy. Therefore, it does not follow a scientific method, nor a structure similar to most other works of this level. Keep that in mind as you read this text. My conclusions are not empirical observations, and as such you may disagree with a lot of what I say. That is an inevitability for any philosophical text, and that is what makes philosophy fascinating—it is never solved, and it is constantly moving and shifting and evolving,

with disagreements creating new philosophy. My hope is that this text will provide a new perspective to enrich the ideas you already have about AI. It is not meant to act as a declaration of truth about the matters discussed within.

In another sense, this text is a look at some philosophical ideas and theories which I find fascinating, strung together to paint a picture of a broader theme. It primarily discusses the ideas of Douglas Hofstadter, John Searle and Daniel Dennett, who, as contemporaries, had a lot to do with each other and discussed many of the same topics, although each from their own framework. This work could be construed as an observation of the dialectic that emerges from their collective works. Whichever way you prefer to think of this work, I hope that at least you find it interesting at some level and are entertained while reading it.

2. Turing, Hofstadter, Searle

2.1. Alan Turing

Figure 1. Alan Turing c. 1951 (Elliott & Fry 1951).

In any discussion about AI, one cannot ignore Alan Turing's accomplishments and what effect they had on the field of AI research. He was highly influential in the early days of computer science, and his involvement in World War II as a codebreaker is estimated to have saved millions of lives (Copeland 2012). Tragically, despite his accomplishments, he was ostracized by his home country of the UK due to his homosexuality, which at the time was a crime (Biography.com Editors 2019).

One of the most prominent of Turing's accomplishments is the *Turing machine,* which he invented in 1936. The Turing machine is, essentially, a theoretical model of an abstract machine which mechanically operates on a tape, which Turing used as a proof for what is and is not computable. A subcategory of Turing machine is the *universal Turing machine,* which is able to simulate any other Turing machine, and is considered analogous to the "best possible" computer. Thus, if a system is as capable as a universal Turing machine, it can effectively compute anything that *can* be computed. (De Mol 2018)

> A Turing machine can do everything that a real computer can do. Nonetheless, even a Turing machine cannot solve certain problems. In a very real sense, these problems are beyond the theoretical limits of computation. (Sipser 2006, p. 137)

Systems of data-manipulation rules that can simulate any Turing machine are said to be *Turing-complete,* which include virtually every programming language in use today —as well as some other systems that are Turing-complete by accident, like video games such as *Minecraft* and *Dwarf Fortress* (Gwern 2019). These accidents demonstrate a rather unintuitive truth about how simple Turing machines really can be in order to be able to compute anything that is computable:

> One might think that such universality as a system being smart enough to be able to run any program might be difficult or hard to achieve, but it turns out to be the opposite and it is difficult to write a useful system which does not immediately tip over into TC [Turing completeness]. (Ibid.)

The formulation of the Turing machine led, in part, to the formulation of the Church–Turing Thesis, which Turing and mathematician Alonzo Church postulated independently of each other. The thesis (very generally speaking) postulates that any "effectively calculable" function is computable by a Turing machine. As the notion of effective calculability is an informal one, the thesis cannot be definitively proven, but it regardless has near-universal acceptance, and all attempts call it into question so far have failed. (Copeland 2017)

Another invention of Turing's that is relevant in AI to this day is the *Turing test.* Turing formulated the test in his paper *Computing Machinery and Intelligence* (1950) to deal

with the question “can machines think?”, which he found to be “too meaningless”. So, he took a different, but related, approach by describing a variant of “the imitation game”, in which a person is tasked to figure out, by written communication, the sex of two players hidden from view, who are each trying to convince the interrogator that they are the other player. If one of the players is replaced by a machine, and is able to fool the interrogator as often as a human player would, then it is said to pass the test and demonstrate actual intelligence. (Oppy & Dowe 2016) The paper has since become one of the most frequently cited in modern philosophical literature, even though Turing himself never described himself as a philosopher (Hodges 2013).

2.2. The Church–Turing Thesis

When AI was formally conceived of in a Dartmouth workshop in 1956 (McCarthy, Minsky, Rochester, Shannon 1955), it was done so in a setting where Turing had already done much of the groundwork. Turing’s formulation of the Turing machine had paved the way for the invention of modern computers (De Mol 2018), and by 1945 Turing himself had become convinced that computable functions encompassed all mental functions of the brain (Hodges 2013). Hence, AI research started off in a place where researchers had good reason to be optimistic. In addition, a beneficial consequence of Turing’s formulation of computability was that as long as the researchers were working with Turing-complete languages, no advancements in computer technology could ever make their work obsolete, since general computability is encompassed in Turing-complete systems. There were limitations in the hardware at the time, of course—namely in terms of speed and memory—but no newer computer could calculate more things than whatever systems they already had access to. This meant that the pertinent questions that needed answering the most were not “*can* this be simulated?”, but rather, “*what* needs to be simulated?” As such, these questions pertained far more to philosophy of mind than they did to computer science.

To better illustrate why that was the state of affairs, we ought to discuss what implications the Church–Turing Thesis had for philosophy of mind, and consequently, AI. Douglas Hofstadter, author and (among other things) cognitive scientist, crafted a series of analogies in *Gödel, Escher, Bach: an Eternal Golden Braid* (first released in 1979) to illustrate what those implications were. In his analogies he makes reference to a programming language of his creation called FlooP, which for all intents and purposes is the same as any other Turing-complete language. Hofstadter provides different “versions” of Church–Turing Thesis to illustrate how differences in viewpoint give rise to differences in interpretation of the thesis. The “standard version” of the thesis, for instance, goes as follows:

> Suppose there is a method which a sentient being follows in order to sort numbers into two classes. Suppose further that this method always yields an answer within a finite amount of time, and that it always gives the same answer for a given number. *Then:* Some terminating FlooP program (i.e., some general recursive function) exists which gives exactly the same answers as the sentient being's method does. (Hofstadter 1999, p. 561)

This version implies that mental processes are representable in a Turing-complete language, but it rests on the "intuitive belief [...] that there are no other tools than those in FlooP" (Ibid.). But, as Hofstadter goes on to demonstrate, there are stronger versions which make the case more evidently. The standard version is weak and easily dismissible because it makes the mistake of supposing that FlooP functions are isomorphic to mental processes at the level of consciousness and thought (what Hofstadter called the "symbol" level). We intuitively make comparisons at that level, because it is the level which we can meaningfully interpret.

> In arithmetic, the top level can be "skimmed off" and implemented equally well in many different sorts of hardware: mechanical adding machines, pocket calculators, large computers, people's brains, and so forth. This is what the Church–Turing Thesis is all about. But when it comes to real-world understanding, it seems that there is no simple way to skim off the top level, and program it alone. (Hofstadter 1999, p. 569)

The level at which mental processes are isomorphic to FlooP functions is far below the level of consciousness and thought. The problem is that no meaningful interpretation of the top level can be derived from operations at low levels, nor can we tell how operations at low levels give rise to the operations at the top level. But, by applying some "reductionistic faith", we can assume that the low-level operations do regardless give rise to the top-level functions. A fundamental property of Turing's formulation of computability is that any Turing machine can be simulated by a (universal) Turing machine (De Mol 2018). If there is a level of mental processes that can be represented in FlooP, those processes are equivalent to Turing machines. Hence, if that level of mental processes gives rise to the highest level of mental processes (through however many layers), then each step of the way the processes at each given level are *also* equivalent to Turing machines, including the very highest level, a Turing machine composed of a vast network of Turing machines.

Consider, for instance, this alternative to the sorting problem given above: suppose that there is a method which a presynaptic neuron follows to determine whether or not a neurotransmitter is released, and suppose further that identical variables in that method always lead to the same answers. *Then:* Some terminating FlooP program exists which gives exactly the same answers as the presynaptic neuron's method does.

By reducing brain operations to the level where a deterministic view of the mechanism becomes apparent (or even inevitable), we can see clearly that there *is* a level for which an effective method exists, and is therefore computable by a Turing machine. Then we repeat the process the next step up—let's say, the level at which interactions between presynaptic and postsynaptic neurons happen—and ask, again, does an effective method exist for the process at this level? Well, the neurons are already known to be computable by Turing machines in that equation, and they in turn must be simulatable by a Turing machine, so we only need to concern ourselves with the interaction between those Turing machines.

The question remains, if we repeat this process all the way up to brain level, will we ever encounter any process for which an effective method does not exist? At each given step up, we are only looking at the interactions between Turing machines, so is there anything to those interactions that eludes computability? Turing came to believe that the answer was "no".

As Hofstadter puts it, you cannot "skim off" the top level and program it on its own, but if that top level is treated as part of a larger structure in which lower levels are expressible in FlooP, then ultimately the top level should be as well. This is a culmination which Hofstadter calls the "AI version" of the Church–Turing Thesis:

> Mental processes of any sort can be simulated by a computer program whose underlying language is of power equal to that of FlooP—that is, in which all partial recursive functions can be programmed (Hofstadter 1999, pp. 578–579).

We can see how evident this version of the thesis is if we break it down to its base propositions:

> Proposition 1. The mental substrate can be simulated by a Turing machine.
>
> Proposition 2. The mind is built up from the mental substrate.
>
> Proposition 3. Turing machines can simulate any other Turing machine.
>
> Conclusion 1. The mind can be simulated by a Turing machine.

Of course, Hofstadter called this a result of "reductionistic faith", as it does not take into account how mental processing is affected by the soul, Cartesian mind-substance, or even the possibility that some interactions do turn out to be uncomputable (see Hofstadter 1999, p. 574). That is certainly true enough, and it makes this version of the thesis inconclusive (in fact, no version of the thesis is fully conclusive), but *it is as conclusive as it can be.* AI researchers could not have made a more conclusive case for computability of the mind by simply tackling the Church–Turing Thesis directly; the

thesis had already existed and was as well-defined as it reasonably could have been. What they could have done is proceed with the faith that the thesis works and see where we can get from that. The alternative is supposing that the soul could account for any amount of mental activity, or that some uncomputable interaction proves to be a roadblock, and hence simulating the mind at any level is inevitably a fool's errand.

With that in mind, the *Dartmouth Summer Research Project on Artificial Intelligence* of 1956 set off on the conjecture that any feature of intelligence can be simulated by a machine, as is stated in their proposal:

> We propose that a 2-month, 10-man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. *The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.* An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer. (McCarthy et al. 1955; my emphasis)

Rather than asking "can the human mind be simulated", they simply proceeded with the assumption that the answer is yes, and they had good reason to do so. They could also be assured that any Turing-complete language would be sufficient in the task they set out to do, so the problems at the level of computer science were mainly down to implementation. Hence, what remained were ultimately *philosophical* questions, like "what is intelligence?" A conclusion that can be made from the "AI version" of Church–Turing Thesis is that any serious attempt at simulation must be built bottom-up, with the lowest level functions programmed first, and working upwards from there. The question of what intelligence is then becomes "what are the discrete constituents of intelligence that need to be simulated?"

Marvin Minsky, who was one of the participants of the Dartmouth workshop and became one of the most prominent figures in AI research, released *The Society of Mind* in 1986. It is a book that, as the name suggests, conceives of the mind as a "society" of discrete components, which he called "agents" (Minsky 1988). It makes apparent what AI was primarily concerned with during that early period of AI research: the book, while considered a seminal work in the field of AI, actually says very little about AI or computer science. It is almost entirely a philosophical pursuit into unraveling the mysteries of the mind, which actually seems very fitting. If a proper simulation of intelligence can only be programmed at the agent-level, then obviously one ought to find out what the agents actually are before doing anything else.

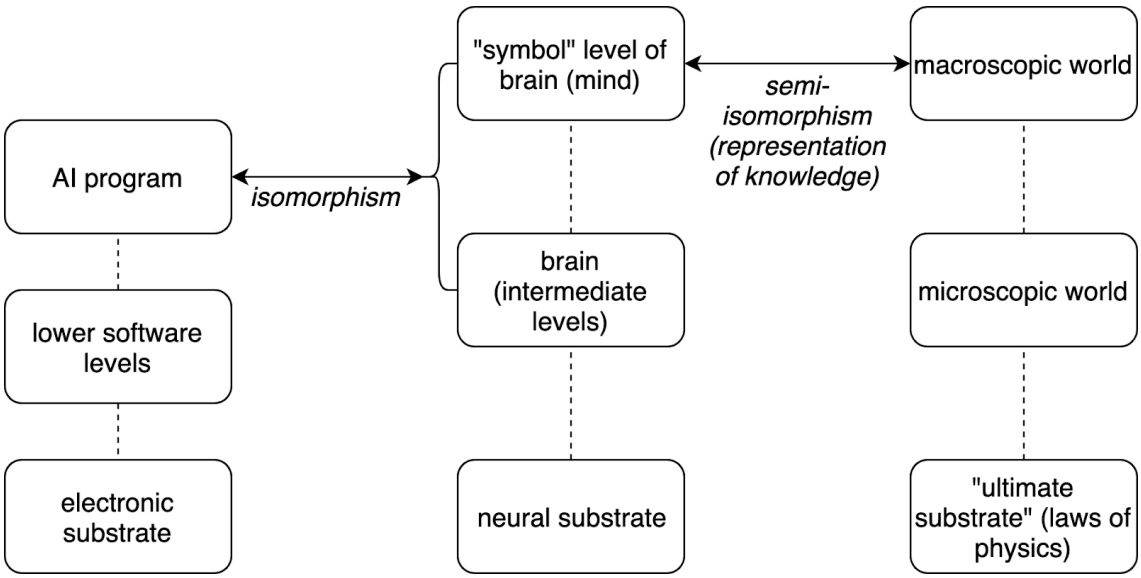

Figure 2. A recreation of a diagram found in *Gödel, Escher, Bach.* The original caption for the diagram reads: “Crucial to the endeavor of Artificial Intelligence research is the notion that the symbolic levels of the mind can be “skimmed off” of their neural substrate and implemented in other media, such as the electronic substrate of computers. To what depth the copying of brain must go is at present completely unclear.” (Hofstadter 1999, p. 573)

2.3. Tesler’s Theorem

As a philosophical pursuit, this kind of conception of AI has the unfortunate side effect of diminishing its own accomplishments. When breakthroughs in AI happen, they are often dismissed on philosophical grounds, even if in terms of computer science they are spectacular—this is because, paradoxically, in managing to “reveal” something about the nature of intelligence, the appearance of that revelation is that it must have not been a factor in intelligence in the first place. People expect intelligence to be something so mysterious and elusive that the models we have come up with could not possibly explain the “whole story”, so to speak. Hofstadter put it this way in his diagnosis of this paradoxical situation:

> There is a related “Theorem” about progress in AI: once some mental function is programmed, people soon cease to consider it as an essential ingredient of “real thinking”. The ineluctable core of intelligence is always in that next thing which hasn’t yet been programmed. This “Theorem” was first proposed to me by Larry Tesler, so I call it *Tesler’s Theorem:* “AI is whatever hasn’t been done yet.” (Hofstadter 1999, p. 601)

This phenomenon in general is called the “AI effect”. Rodney Brooks, former director of MIT’s Artificial Intelligence Laboratory, lamented on it, “every time we figure out a piece of it, it stops being magical; we say, ‘Oh, that’s just a computation’” (Kahn 2002). The AI effect meant that when AI researchers would come up with new technologies, those technologies would find a lot of uses in other fields and consumer products—but not in

constructing intelligence. Hence, achievements in AI were often seen as achievements in other fields. Patrick Winston, another former director of the MIT AI Laboratory, said on the matter:

> AI has become more important as it has become less conspicuous. […] These days, it is hard to find a big system that does not work, in part, because of ideas developed or matured in the AI world. (Swaine 2007)

AI did, for the longest time, have the problem that the hardware was simply not there to produce real-time intelligence, so much of the work had to be done on theoretical grounds. Researchers had cause to be optimistic thanks to Turing, but it would not help if there was no money to fund that optimism. We now live in a world where the hardware does exist, and the innovations in AI are near-ubiquitous. This has, seemingly, had the rather interesting effect that AI has found itself to be a very marketable term, and a lot of things are AI simply by the virtue of using certain technologies (namely, machine learning and artificial neural networks). AI has shifted from an ultimately philosophical pursuit to understand intelligence where the goal preceded the methods, to a consumer enterprise with real-world applications where the methods themselves define what can be called AI. Does it make sense to even keep these two things under the same umbrella term?

In their assessment, Andreas Kaplan and Michael Haenlein (2018) define AI as “a system’s ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation.” For a system to fall under this definition, it very likely has to incorporate technologies such as machine learning and artificial neural networks, which are both fairly recent innovations (at least to the extent that they have been used in practical applications). Kaplan and Haenlein even make a note that many systems once held to be major breakthroughs in AI fall outside this definition, naming Deep Blue, the chess computer that defeated Garry Kasparov in 1996, as simply an expert system and not AI (Ibid.). Google’s AlphaGo, which was the first Go program to defeat a professional player without handicaps (Silver & Hassabis 2016), does fall under the definition, despite the fact that both it and Deep Blue are very similar in their purpose: they were both meant to be better than humans at their respective games, and they both succeeded. A conclusion can be made that in the current climate of AI, implementation is given precedence over purpose and function, whereas the reverse used to be the case.

Deep Blue did offer an interesting instance of Tesler’s Theorem in that prior to Deep Blue’s victory, chess was commonly held to be a game requiring genuine creativity, and when that view was shattered by Deep Blue’s brute forcing methods, many were left

disillusioned. One of those people was Hofstadter himself: he predicted in *Gödel, Escher, Bach* that defeating the best human players would require general intelligence, and as such a chess machine which does nothing but play chess would not be able to do it (Hofstadter 1999, p. 678). After Deep Blue's victory over Kasparov, Hofstadter commented, "It was a watershed event, but it doesn't have to do with computers becoming intelligent." (Weber 1996). It goes to show that even with the optimists in AI, Tesler's Theorem holds true.

2.4. The Chinese Room

Back in 1980, philosopher John Searle stirred the pot with a thought experiment that would come to influence discussions about AI for the decades to come. His "Chinese room" thought experiment was published in an article titled *Minds, Brains, and Programs,* and it involved a distinction between what he called *weak AI* and *strong AI,* which he distinguishes as follows:

> According to weak AI, the principal value of the computer in the study of the mind is that it gives us a very powerful tool. For example, it enables us to formulate and test hypotheses in a more rigorous and precise fashion. But according to strong AI, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really *is* a mind, in the sense that computers given the right programs can be literally said to *understand* and have other cognitive states. (Searle 1981, p. 353; Searle's emphases)

Searle was strongly opposed to strong AI, and crafted a thought experiment to support his stance. Searle asks the reader the suppose that he is locked in a room and given a large batch of Chinese writing, and furthermore, that he does not understand any of it. Then he is given a second batch of Chinese writing, alongside rules in English for correlating the second batch with the first. Then he is given a third batch, again with rules in English for correlating the third batch with the first two batches, and these rules also instruct him how to create certain Chinese symbols in response to the symbols within the third batch.

> Unknown to me, the people who are giving me all of these symbols 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." (Searle 1981, p. 355)

The idea behind the thought experiment is that for the people outside the room, the "program" they have created is able to converse fluently in Chinese and passes the Turing test. But inside the room, the person carrying out the symbol manipulation has no actual understanding of Chinese, and the addition of the "program" does nothing to make that understanding appear.

A fundamental distinction Searle makes between programs and minds is that programs have purely *syntactic* content, meaning all of the content is a manipulation of formal symbols, and the program has no access to what is being meant by the manipulation. The symbols running across the tape of a Turing machine do not *symbolize* anything to the machine itself, but they do to us, the observers and programmers. Hence, the meaning behind the symbols is a derived characteristic that the program has no access to. By contrast, minds have *semantic* content, meaning that our mental states do carry with them the "symbolic meaning" of the neural processing of the brain, which the neurons themselves have no access to. Searle effectively claims that formal systems (i.e., Turing machines) cannot ever cross over into "semanticity", as their functions are always describable as manipulation of formal symbols according to given rules, which is *de facto* syntactic.

We will address that claim in further detail in the next chapter, but for now let's define another term that Searle uses, which we will be using a lot as well. That term is *intentionality.* Searle defines it as

> [...] that feature of certain mental states by which they are directed at or about objects and states of affairs in the world. Thus, beliefs, desires, and intentions are intentional states; undirected forms of anxiety and depression are not. (Searle 1981, p. 358)

States such as beliefs and desires are not just states in themselves, they are always *about* or *directed at* something. A belief is a belief *about* something, for instance. (see also Jacob 2003) They, alongside other intentional states, are held to be a fundamental aspect of the "semanticity" in the human mind, with Franz Brentano, the philosopher who reintroduced the concept of intentionality in modern philosophy, holding that intentionality is the "mark of the mental"—a claim that is also referred to as Brentano's Thesis (Dennett 1987, p. 67). Searle claims that without semanticity, there is no intentionality:

> No purely formal model will ever be sufficient by itself for intentionality because the formal properties are not by themselves constitutive of intentionality, and they have by themselves no causal powers except the power, when instantiated, to produce the next stage of the formalism when the machine is running (Searle 1981, p. 367).

We do have a tendency to talk about machines *as if* they had intentionality, and according to Searle we do that simply because it is an adequate way of describing their behavior, but the machines do not (and cannot) possess any intrinsic intentionality. A thermostat could be said to hold beliefs about the temperature of the room it is in, but that does not mean the thermostat actually *does* believe anything. In other words, the

intentionality of so-called artifacts is *derived* by those who possess *original* intentionality (i.e., us).

Many believe that Searle's argument simply concerns philosophy of mind, and the distinction Searle makes between weak and strong AI does not concern practitioners of AI. But I believe the two are inherently entwined. Hidden within the strong AI hypothesis is a claim we have already discussed, and shown to be fundamental—so in Hofstadter's footsteps, let's rephrase the hypothesis:

> Strong AI hypothesis, Church–Turing version: The human mind is equivalent to a Turing machine.

How could a practitioner of AI not care about that?

3. Syntax and Semantics

3.1. Einstein's Brain

TORTOISE: A book doesn't feel any way. A book just *is.* It's like a chair. It's just there.

ACHILLES: Well, this isn't just a book—it's a book plus a whole *process.* How does a book plus a process feel?

TORTOISE: How should I know? But you can ask it that question yourself.

ACHILLES: And I know what it'll say: "I'm feeling very weak and my legs ache," or some such thing. And a book, or a book-plus-process, *has* no legs!

TORTOISE: But its neural structure has incorporated a very strong *memory* of legs and leg-aching. Why don't you *tell* it that it's now no longer a person, but a book-plus-process? Maybe after you've explained that fact in about as much detail as you know it, it would start to understand that and forget about its leg-aching, or what it took for leg-aching. After all, it has no vested interest in feeling its leg, which it doesn't have, aching. It might as well ignore such things and concentrate on what it *does* have, such as the ability to communicate with you, Achilles, and to think.

ACHILLES: There is something frightfully sad about this whole process. One of the sadder things is that it would take so much *time* to get messages in and out of the brain, that before I'd completed many exchanges, I'd be an old man.

TORTOISE: Well, you could be turned into a catalogue too.

The above is an excerpt from Hofstadter's *A Conversation with Einstein's Brain,* as it was printed in Hofstadter and Daniel Dennett's co-authored book *The Mind's I* (1981, pp. 430–457). In the dialogue, Achilles and Tortoise discuss the rather absurd prospect of carrying out a conversation with a catalogue covering the exact brain state of Albert Einstein on the day of his death, down to the neuron level. The dialogue illustrates the uncomfortable idea that you *could,* given a vast amount of space and time, carry out any kind of mental operation with a purely syntactic representation of mental content.

This is very analogous to the postulations we made on the Church–Turing Thesis, as we argued that neural activity—in high likelihood—corresponds to mental activity at the highest level.

The absurdity of the dialogue stems from the fact that interacting with a syntactic representation of a human being in this fashion is actually impossible. The tome would be billions of pages long, and consulting it for even the simplest neural firing would take a very long time. Searle, in his Chinese room thought experiment, approaches a comparable level of complexity, but attempts to ignore it by supposing that the process of symbol manipulation for "understanding" Chinese could be carried out by a human being in a timeframe that is inconspicuous to outside observers.

That is exactly the point Hofstadter and Dennett make in their reflections on *Minds, Brains and Programs.* They argue that Searle is "inducing an illusion" by making readers overlook the complexity of the system in question:

> Now, for a person to hand-simulate [an AI program]—that is, to step through it at the level of detail that the computer does–would involve days, if not weeks or months, of arduous, horrendous boredom. But instead of pointing this out, Searle [...] switches the reader's image to a hypothetical program that passes the Turing test! He has jumped up many levels of competency without so much as a passing mention. The reader is again invited to put himself or herself in the shoes of the person carrying out the step-by-step simulation, and to "feel the lack of understanding" of Chinese. This is the crux of Searle's argument. (Hofstadter & Dennett 1981, pp. 373–374)

Now, Searle's objection that manipulation of formal symbols cannot "add up" to semantic content is not unreasonable—after all, we just got done building a simulation of a brain using the Church–Turing Thesis as our guide, where seemingly all we did was manipulate syntax. The effective methods at each given step of the brain simulation are calculable, and as such, effectively equivalent to the manipulation of formal symbols in the sense that Searle meant it—so where did we introduce semantic content? Nowhere, right? But hang on—we did that with the *brain,* not a program, just as Hofstadter has done with Einstein's brain. Yet Searle insists that brains *do* have this mysterious underived semanticity.

His argument rests on an "intuition pump", according to which the semanticity of the mind is self-evident due to our first-person access to our internal states, but the "exposed" syntax of the simulation reveals that no semanticity can possibly exist within it. Searle even addresses the brain simulation prospect directly in a similar thought experiment, in which the network of neurons is replaced by a network of pipes, with a person opening and closing valves to simulate synaptic firings (Searle 1981, pp. 363–364). Searle expects us to imagine a network of pipes not unlike the ones running

underneath city streets, and asks “can this system produce understanding?”, to which the obvious answer is “of course not”—but once again, in order to induce the illusion, Searle leaves out the aspect of scope. I would imagine that the network of pipes would have to be at the very least planet-sized in order to come close to the level of complexity in the brain—and yet, Searle expects that to be operable by a single person? Well, not really—he simply wants you to imagine *yourself* operating those valves and asking yourself whether those pipes are thinking anything at all, to which the answer should be self-apparent. The fact of the matter is that the theoretical worlds Searle presents *cannot* be conceived of at the level of precision that actually understanding their implications necessitates.

In the case of the Chinese room, Searle wants you to imagine yourself both inside and outside the room; from the outside, it appears that you are conversing with a system that, if not human, at least passes the Turing test with flying colors. But when you transport yourself inside the room to do the symbol manipulation, you are supposed to realize that no matter how complex the system is and how much semantics can be derived from it, it has none of the intrinsic semantics or intentionality, because obviously no meaning or representation is encoded into the formal symbols themselves.

However, if approached realistically with regard to speed, this fantasy quickly falls apart, as the system’s response time would be so slow that outside observers would believe it simply does not function (nevermind passing the Turing test), and the person inside the room would not have the ability to assess the symbols in a way that would allow him to derive conclusions about the symbols in question. For all he knows, the symbols *are* a catalogue of a brain, as well—but in this case, the brain of a native Chinese speaker.

So, since the catalogue of Einstein’s brain is so closely analogous to the Chinese scripts, what happens if we remove the factor of slow speed? Suppose that we replace the human operator for the catalogue with a supercomputer capable of computing the synaptic firings at the rate of an actual brain, and equip the system with sensors to produce all the necessary inputs (i.e., senses like sight, hearing and touch). We could also give it the ability to *rewrite* the catalogue according to new neural pathways being created, among other aspects of change within the brain. There should now be nothing stopping this system from thinking and acting like Einstein himself, but we have, according to Searle, accomplished nothing in terms of semantic content and intentionality, which is clearly seen if his argument is broken down to its propositions (courtesy of Dennett):

> Proposition 1. Programs are purely formal (i.e., syntactical).
>
> Proposition 2. Syntax is neither equivalent to nor sufficient by itself for semantics.
>
> Proposition 3. Minds have mental contents (i.e., semantic contents).
>
> Conclusion 1. Having a program—any program by itself—is neither sufficient for nor equivalent to having a mind. (Dennett 1987, p. 324)

According to Searle, semantic content is unreachable for *any kind* of purely syntactical system. We have not added anything to our Einstein system that is not syntactic, and the only thing we have done aside from that is drastically increase the speed of symbol manipulation. For Searle, speed is simply irrelevant—speed will only be a factor in derived semantics, but the system itself will not have intrinsic semantic content. We could keep improving the system so that eventually we would have a perfect physiological replica of Albert Einstein himself, and that would still not change anything. Semanticity simply cannot be reached "from below", in Searle's view.

3.2. Original and Derived Intentionality

We are left with two options. The first is that our brains somehow have acquired semantic content without it having been "built up" from syntactic content, in such a way that not even precise simulations can get to it. For a religious or spiritual person the concept of soul will be an obvious solution here, as is "mind-substance" for a Cartesian dualist, but Searle himself has criticized dualistic views of the mind and has insisted that intentionality is "a biological phenomenon, and it is as likely to be causally dependent on the specific biochemistry of its origins as lactation, photosynthesis, or any other biological phenomena" (Searle 1981, p. 372). This view is very difficult to reconcile with the stance that human minds have underived semantic content, which has led to Dennett claiming that Searle's issue is with *consciousness,* not semantics:

> Searle has apparently confused a claim about the underivability of *semantics* from syntax with a claim about the underivability of *the consciousness of semantics* from syntax. For Searle, the idea of genuine understanding, genuine "semanticity" as he often calls it, is inextricable from the idea of consciousness. He does not so much as consider the possibility of unconscious semanticity. (Dennett 1987, p. 335; Dennett's emphases)

If Searle's fundamental issue turns out to be about the deep-seated intuition that consciousness is unrealizable in machines, but regardless present in all humans (even though nobody has access to each other's first-person experience to confirm that), we will have regressed back to Cartesian dualism. It would just so happen that we have privileged access to consciousness and nothing else can have it, and hence consciousness takes on the role of a mysterious, ever-elusive substance, even if

Searle does not literally mean that to be the case. That is not a satisfying conclusion for us, nor should it be for Searle, if he is as critical of dualism as he claims he is.

The second option, espoused by Dennett, is that *derived semantics is all there is.* This view strikes against some of the most deep-seated intuitions we have, which is why it is unpopular, but it has some remarkable consequences in our search for the essence of AI. That is because if all semantics is derived, then the semanticity we suppose humans to have is derived as well, just as it is in other systems. So the as-if intentionality Searle talked about is not just a nice utility or a manner of speaking, it is also the *only* kind of intentionality there is—leaving nothing that machines cannot accomplish in terms of mental content. At the same time, it begs the question of what role exactly intentionality plays in our conception of AI, if Searle thought it made such a fundamental distinction between humans and machines (we will get to that).

Specifically, Dennett argues against what he calls the doctrine of original intentionality, which "[...] is the claim that whereas some of our artifacts may have intentionality derived from us, we have original (or intrinsic) intentionality, utterly underived" (Dennett 1987, p. 288). Artifacts in this context refer to anything of our creation with artifice, such as AI.

Dennett illustrates his argument with a thought experiment: he asks you to suppose that you wanted to experience life in the twenty-fifth century, and that the only way to accomplish this is by being placed in a hibernation device for the duration. He then asks the question, how would you go about designing such a device? (Dennett 1987, pp. 295–298)

He goes on to list some of the strategies you might use in designing the device. For one, you might want the device to be mobile, because if it were stationary, there would be a higher possibility of it being destroyed, perhaps because of its being in the way of a building project. Likewise, with mobility, the device (henceforth called a robot) would be able to seek out new energy sources if the ones at its current location were to run out. To guide its locomotion, it would need to have sensory systems and the ability to choose actions that further your goal of survival. Secondly, you might have to plan for the possibility that you will not be the only one with this mission; copycats may crop up, and your robot would have to survive in a world with other similar robots. Your robot would need to be able to calculate the benefits and risks of cooperating with other such robots, and even to form long-lasting alliances. It would have to act entirely on its own accord for hundreds of years without any input from its designer. (Ibid.)

You would end up with a robot that, in order to best pursue its ultimate goal (i.e., preserving you), would need to be able to act independently, set subsidiary goals, and have access to its internal states. It may even turn out to be wise to *not* program the robot to have the goal of protecting you, but rather simply protecting *itself,* because self-preservation strategies may turn out to be more effective than preserving you at the cost of its own survival.

This robot, although incredibly sophisticated, would still have no original intentionality. Even with what appears to be, at least, a primitive form of consciousness and the ability to direct intention, its intentionality would still be derived

> [...] from its artifactual role as your protector. Its simulacrum of mental states would be just that—not *real* deciding and seeing and wondering and planning, but only *as if* deciding and seeing and wondering and planning. (Dennett 1987, p. 297)

Up to this point in Dennett's thought experiment, proponents of the doctrine of original intentionality will be nodding along—of course this robot, impressive as it may be, only possesses derived intentionality! As it was designed by a human for a specific purpose, whatever mental states it has have no intrinsic semantics, because the actual meanings behind its actions lie in the designer's intentions, not in the intentions of the machine!

Here, however, is how the thought experiment is cashed out:

> I want to draw out the most striking implication of standing firm with our first intuition: no artifact, no matter how much AI wizardry is designed into it, has anything but derived intentionality. If we cling to this view, the conclusion forced upon us is that our own intentionality is exactly like that of the robot, for the science-fiction tale I have told is not new; it is just a variation on Dawkins's vision of us (and all other biological species) as "survival machines" designed to prolong the futures of our selfish genes. (Dennett 1987, p. 298)

Once you buy into the very commonplace idea that natural selection is real, all sorts of implications regarding semantic content and intentionality emerge. If the survival robot's case for original intentionality is dismissed on the basis of its intentionality being derived from the intentions of its designer, then it must also be the case that our *own* intentionality is derived from the interests of our *genes.*

Searle held that our mental states symbolize the manipulations of "symbols" that happen in the firings of synapses, and that is what gives us our underived semanticity, whereas in artifacts the semanticity is not built into themselves, but is instead derived by the creators who hold the keys to the meanings behind the symbols. Here I believe

that Searle makes a fundamental category error in supposing that programs are analogous to mental states and computers are analogous to brains (see Searle 1981, p. 369). The symbols running across a Turing machine symbolize nothing to the machine itself, but the symbols are what they are precisely because they *do* symbolize *something*—it just happens to be dictated by the creators of the machine and the deciders of the symbols, rather than the machine whose function it is to process the symbols.

The claim of underived semanticity, then, suggests that unlike the humble Turing machines, there is nothing "above" the human mind that would dictate the symbolization of our symbols to us. Instead, our mental states, according to the claim of underived semanticity, *in themselves* constitute this symbolization. It is no wonder why this conception of mental states fails to be analogous to programs, as that was the wrong analogy to begin with! Rather, a more accurate analogy would be that computers and programs *in conjunction* are analogous to brains—as neither one on their own is analogous to a brain—and the "symbolization" in both brains and computer-program-conjunctions is not innate to themselves.

So what is it that dictates what our symbols symbolize? Ultimately, it is our genes—and by extension, the entire process of natural selection (and by even further extension, the laws of physics). It is our genes that encode our mental states into our brains, as well as everything that comes associated with it—such as our inclination to think of ourselves as "selves", rather than collections of single-celled automata. Whatever meaning we *think* those states represent, above that meaning is another meaning that says, roughly, "this function increases this species' capacity to propagate genes". Since *that* meaning is not encoded into our mental states, those mental states do not constitute *underived* semanticity.

In Searle's view, it would follow that without underived semanticity, original intentionality is automatically off the table, but we may as well make an explicit case for it. The whole idea of *original* intentionality is that *we* (as humans, rather than mindless collections of cells) are the originators of our intentional states, which means that there should be no intention behind our using of intentional states, as the intentional states in themselves produce the "original" intention. But, seeing as we are the result of a multi-million-year evolutionary process, in which it is ultimately our genes which dictate what mental states we have at our disposal, then it follows that there *is* an intent behind our use of intentional states—the intent of our genes to *survive and propagate.* Our intentional states exist as a consequence of *having been selected for,* so their *very existence* is subservient to the intentions that precede them!

This is perfectly analogous to our case with the survival robot, where the robot's intentions are derived from our supposed original intentions. The robot exists solely for the purpose of fulfilling our intentions to survive to the twenty-fifth century—and hence the robot's intentions cannot be original—but in our intentions to survive we are also fulfilling our genes' intentions to survive. I suggested earlier that you might not want to design the robot with your survival in mind, but rather its own, and now you might realize why. We, too, strive to survive for our own sake instead of our genes' sake, and for our genes that has apparently worked out as a very effective strategy—perhaps it would for the robot as well.

But hold on, this all seems very perverse—surely saying that genes intend *anything* is in itself a derived meaning. After all, neither our genes nor Mother Nature direct any intentionality of their own origin; we merely talk *as if* they had intentionality because there does not seem to be a better way to describe how their actions manifest into vast, life-encompassing patterns. Well, yes—and *that is precisely why there is no original intentionality!* Genes have no original intentionality because they have no mental powers to represent their intentions, but neither do we because our intentionality does not originate from us. Nothing has it!

Either that is the case, or one has to drop the doctrine of our privileged access to original intentionality. Either our artifacts cannot have original intentionality, and as such we cannot either, or we do decree that we have original intentionality despite all the problems that entails, and as such it has to be the case that our artifacts can have it as well. In any case, we are survival machines of another mechanism's design—much fancier and more complex than anything we could come up with (for now, at least), but otherwise, fundamentally no different.

3.3. Dualistic Intuitions

This chapter may seem like an attack on Searle, but really it is about the intuition that there are "deeper facts" about the mind that machines can never have access to, be it due to the presence of a soul or Cartesian mind-substance or any other such thing. It is probably the most common and pervasive intuition we have as a species (hence religion). We are addressing Searle's arguments in specific because he has done the due diligence of formalizing his intuitions in such a way that we can properly address them, and because the Chinese room argument is very pertinent to AI in particular. Many other philosophers of Searle's renown share his ideas; Dennett talks about them in detail in chapter 8 of *The Intentional Stance* (see 1987, pp. 287–321).

In a weird roundabout way, it seems that Searle is at least partially right—we cannot get to underived semantics from syntax. But that seems rather meaningless if *all* semantics is derived, and thanks to natural selection we have a very good reason to suppose that is the case. So, then it is just a matter of when it makes sense to derive semantics from a program, and in what fashion.

Of course, one could simply state that “the indivisible soul that exists in all of us” is what gives us our original intentionality, and that would certainly dismiss everything we have discussed. But outside of saying “okay, you are certainly free to believe that”, there is no substantial way to address that statement. We can simply say that AI is not contingent on the existence or non-existence of souls, as otherwise there is a whole minefield of possibilities that cannot be accounted for.

What does seem to be fundamental to AI, however, is the speed at which its syntax is processed. Turing machines may be “universal” in the sense that any syntactical function may be performed by them, but they will escape our incentive to attribute intelligence to them if they do not act the part convincingly.

Given what we have discussed in this chapter, we now have a pretty good idea of what is required for a true artificial intelligence:

sufficient syntax + sufficient speed of manipulation

This is deliberately vague. The word “syntax” alone can encompass both a program by itself and the instructions for manipulating the program, if they exist separately. This is important to note because that is how most computer programs are—they are written in high-level code that sits multiple levels of abstraction above the binary instructions “understood” by the processor. You cannot “feed” a program to a processor directly and expect anything to happen. So we have to generalize a bit by lumping all necessary syntax “in one”, lest we continue down a wrong path by assuming that a program by itself is sufficient for manipulation. That was the error Searle made in drawing analogies between programs and mental states—the distinction between hardware (i.e., machine) and software (i.e., program) is not as clear-cut as he thought. Interestingly, this does mean that an “AI program” is not in itself AI, but let’s suppose that the term itself carries with it the assumption that the program is designed to be incorporated in a system that, as a whole, *is* AI.

4. The Intentional Stance

4.1. Mark III Beast

The following is an excerpt from *The Soul of Anna Klane* by Terrel Miedaner, as it was reproduced in *The Mind's I* (1981, pp. 111–113):

> Hunt opened one of several dozen cabinets and brought out something that looked like a large aluminum beetle with small, colored indicator lamps and a few mechanical protrusions about its smooth surface. He turned it over, showing Dirksen three rubber wheels on its underside. Stenciled on the flat metal base plate were the words MARK III BEAST.
>
> Hunt set the device on the tiled floor, simultaneously toggling a tiny switch on its underbelly. With a quiet humming sound the toy began to move in a searching pattern back and forth across the floor. It stopped momentarily, then headed for an electrical outlet near the base of one large chassis. It paused before the socket, extended a pair of prongs from an opening in its metallic body, probed and entered the energy source. Some of the lights on its body began to glow green, and a noise almost like the purring of a cat emanated from within.
>
> Dirksen regarded the contrivance with interest. "A mechanical animal. It's cute—but what's the point of it?"
>
> Hunt reached over to a nearby bench for a hammer and held it out to her. "I'd like you to kill it."
>
> "What are you talking about?" Dirksen said in a mild alarm. "Why should I kill... break that... that machine?" she backed away, refusing to take the weapon.
>
> "Just as an experiment," Hunt replied. "I tried it myself some years ago at Klane's behest and found it instructive."
>
> "What did you learn?"
>
> "Something about the meaning of life and death."
>
> Dirksen stood looking at Hunt suspiciously.
>
> "The 'beast' has no defenses that can hurt you," he assured her. "Just don't crash into anything while chasing it." He held out the hammer.
>
> She stepped tentatively forward, took the weapon, looked sidelong at the peculiar machine purring deeply as it sucked away at the electrical current. She walked toward it, stooped down and raised the hammer. "But... it's eating," she said, turning to Hunt.
>
> He laughed. Angrily she took the hammer in both hands, raised it, and brought it down hard.
>
> But with a shrill noise like a cry of fright the beast had pulled its mandibles from the socket and moved suddenly backwards. The hammer cracked solidly into the floor, on a section of the tile that had been obscured from view by the body of the machine. The tile was pockmarked with indentations.
>
> Dirksen looked up. Hunt was laughing. The machine had moved two meters away and stopped, eyeing her. No, she decided, it was not eyeing her. Irritated with herself, Dirksen grasped her weapon and stalked cautiously forward. The machine backed away, a pair of red lights on the front of it glowing alternately

brighter and dimmer at the approximate alphawave frequency of the human brain. Dirksen lunged, swung the hammer, and missed—

Ten minutes later she returned, flushed and gasping, to Hunt. Her body hurt in several places where she had bruised it on jutting machinery, and her head ached where she had cracked it under a workbench. "It's like trying to catch a big rat! When do its stupid batteries run down anyways?"

Hunt checked his watch. "I'd guess it has another half hour, provided you keep it busy." He pointed beneath a workbench, where the beast had found another electrical outlet. "But there is an easier way to get it."

"I'll take it."

"Put the hammer down and pick it up."

"Just... pick it up?"

"Yes. It only recognizes danger from its own kind—in this case the steel hammerhead. It's programmed to trust unarmed protoplasm."

She laid the hammer on a bench, walked slowly over to the machine. It didn't move. The purring had stopped; pale amber lights glowed softly. Dirksen reached down and touched it tentatively, felt a gentle vibration. She gingerly picked it up with both hands. Its lights changed to a clear green color, and through the comfortable warmth of its metal skin she could feel the smooth purr of motors.

"So now what do I do with the stupid thing?" she asked irritably.

"Oh, lay him on his back on the workbench. He'll be quite helpless in that position, and you can bash him at your leisure."

"I can do without the anthropomorphisms," Dirksen muttered as she followed Hunt's suggestion, determined to see this thing through.

As she inverted the machine and set it down, its lights changed back to red. Wheels spun briefly, stopped. Dirksen picked up the hammer again, quickly raised it and brought it back down in a smooth arc which struck the helpless machine off-center, damaging one of its wheels and flipping it right side up again. There was a metallic scraping sound from the damaged wheel, and the beast began spinning in a fitful circle. A snapping sound came from its underbelly; the machine stopped, lights glowing dolefully.

Dirksen pressed her lips together tightly, raised the hammer for a final blow. But as she started to bring it down there came from within the beast a sound, a soft crying wail the rose up and fell like a baby whimpering. Dirksen dropped the hammer and stepped back, her eyes on the blood-red pool of lubricating fluid forming on the table beneath the creature. She looked at Hunt, horrified. "It's... it's—“

"Just a machine," Hunt said, seriously now. "Like these, its evolutionary predecessors." His gesturing hands took in the array of machinery in the workshop around them, mute and menacing watchers. "But unlike them it can sense its own doom and cry out for succor."

"Turn it off," she said flatly.

Hunt walked to the table, tried to move its tiny power switch. "You've jammed it, I'm afraid." He picked up the hammer from the floor where it had fallen. "Care to administer the death blow?"

She stepped back, shaking her head as Hunt raised the hammer. "Couldn't you fix—“ There was a brief metallic crunch. She winced, turned her head. The wailing had stopped, and they returned upstairs in silence.

4.2. Predictive Strategies

Now that we have arrived at something resembling a formula for AI, let's undermine it completely. The formula simply reflects the "innate" constituents of an AI, with all derived meaning stripped from it. The perverse thing about it is that the same formula could be used to describe the human mind in just as much specificity! The only difference between them is what the word "sufficient" means in each context, and determining "sufficiency" from syntax alone seems like a fool's errand. What precisely should there be a sufficient amount of? Rate of synaptic firings? Number of neural pathways? And what about AI—lines of code? Number of gigahertz?

No, the only way to tell when sufficiency is accomplished in a system is by its derived semantics. The system has achieved sufficient syntax and speed of manipulation for intelligence when *the system behaves in a manner that appears intelligent.* That is effectively what Turing claimed as well in his formulation of the Turing test—he did not even begin to suggest that the syntax of a Turing machine, however complex it is, is in itself useful in determining whether the machine is intelligent.

Think of it this way: even if we were able to come up with a right number for any given aspect of the syntax of the system, how would we know that was the right number? Not by looking at the number as is—the number by itself is meaningless. We would have to look at how the system behaves, and from that deduce what the system is lacking or has a sufficient amount of. It regardless boils down to behavior; since there is no set-in-stone metric for intelligence, the only way to tell when a system is intelligent is by observing its behavior. The value ascribed to any given behavior is not inherent to the syntax itself, it is a value judgment on our part. So in order to tell what the syntactic content should be, we have to derive semantics from the syntax!

That does not mean the formula is *wrong,* but it represents what it means to be purely syntactic—which is to say, it does not mean much of anything. Programs are every bit as syntactic as human minds are. The semantic aspect is there, and it is fundamental, but it is not part of the equation itself, it is derived from the output. It is how we see and think about the end product.

It may seem strange that the arguments presented in the previous chapter come down to a simple behavioristic argument for determining intelligence, but it is largely what we sought to accomplish—we wanted to knock down the human mind from its throne of privileged access to semantic content. Demystifying the mind is necessary in order to make the case that semantic content is attributed to humans by assessing behavior,

not by assuming that humans *simply have it,* completely underived. Semanticity, however derived, is still an incredibly valuable tool, even if “deep down” everything happens syntactically, simply because we cannot even begin to predict human behavior from the millions and millions of synaptic firings happening every second. The same holds true for any other system where the syntax is complex enough to make prediction from syntax unfeasible.

Assuming the human mind has semantic content, and thus intentionality, is just a strategy we employ to avoid having to deal with the deep-down nature of synapses and neurons (or even deeper than that!). It turns out that supposing that other humans *mean* something by their behavior is an extremely effective way of predicting said behavior, and being good at prediction is essential for operating as an effective survival machine (which is what we are).

The flipside of that strategy is that, if humans have no original intentionality, it is in no sense restricted to humans. We can, and do, employ the same strategy towards other animals (and this is intuitively the case for some animals, such as dogs and cats), and even (properly instantiated) programs.

But which programs, and why? This is where we get to the crux of this thesis: I posit that *the primary distinction between AI and non-AI programs rests on the strategies used to predict their respective behaviors.* If one were to use the same strategy towards a program as they would towards a human, it is safe to say that the program can be called AI (within certain parameters, which we will get into later).

Before delving further into why that may be the case, we ought to define what those strategies are in particular—and for that, we are going to yet again consult Dennett. He calls the strategy we have alluded at the *intentional strategy:*

> To a first approximation, the intentional strategy consists of treating the object whose behavior you want to predict as a rational agent with beliefs and desires and other mental stages exhibiting what Brentano and others call *intentionality* (Dennett 1987, p. 15).

The intentional strategy, in other words, involves assuming the *intentional stance* or treating the object as an *intentional system:*

> Here is how it works: first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in many—but not all—instances yield a decision about what the agent ought to do; that is what you predict the agent *will* do. (Dennett 1987, p. 17)

The intentional strategy is in competition with other strategies: a basic and often highly unfeasible strategy is the *physical strategy,* which involves determining the physical constitution of the system, and—using knowledge of the laws of physics—making a prediction on how the system will behave given a certain input (Dennett 1987, p. 16). For something like a human mind, this is akin to predicting human behavior by looking at the interactions of neurons and synapses (or, again, even deeper than that), which we obviously cannot do. But it is the "dogma of physical sciences" that in principle, given accurate enough measuring tools, the physical stance could be used in practice for any given physical system (Ibid.).

The physical stance is localizable to the idea of predicting behavior from syntax—for instance, while a purely physical stance in the context of computer programs would involve understanding the physical constitution of the processor, a more local variant would ignore the low-level operation of the system and focus instead entirely on the program's code. It is still physical, but with an intent to separate the "signal" from the "noise" by having the relevant knowledge to realize that processor-level operation is very likely irrelevant to the problem at hand. For some people, in some cases, this particular stance might actually be beneficial—for instance, for the poor programmer whose job it is to maintain an uncommented codebase written by her predecessors—but even then another strategy might prove more fruitful.

A more broadly useful strategy is the *design strategy,* or assuming the design stance, where "one ignores the actual (possibly messy) details of the physical constitution of an object, and, on the assumption that it has a certain design, predicts that it will behave *as it is designed to behave* under various circumstances" (Dennett 1987, pp. 16–17; Dennett's emphasis). In the case of a computer program, anyone other than its programmers—and sometimes even the programmers themselves—will find this strategy to be a much more reliable way to predict the program's behavior. The users of the program, after all, usually have very limited or zero access to the program's code, which rules out the physical stance altogether (and even in the best case scenario—open source software—it would be highly unreasonable to expect the user to consult to code in order to figure out how the program works). The relationship between the design stance and the physical stance is closely analogous to the

relationship between semantics and syntax: the design stance allows one to ignore the symbols of the code (i.e., the syntax) and merely look for what the symbols *symbolize* (i.e., the semantics).

This strategy will likely prove to be the most useful for most conventional programs. For instance, predicting the outcome of clicking a box with a cartoon image of a hand showing a thumbs-up gesture placed underneath a social media post is much more effortlessly done if you assume that the box was placed there and looks the way it does *by design,* and envisioning what the designer would have had in mind when designing the box. Likewise, designers will rely on their users assuming this stance in order to get them to do anything at all on their program or service. A user could, *in theory,* throw his hands in the air and proclaim "but I don't know what the designer intended! I don't even know who they are or what they think!" and begin to reverse-engineer the program in order to figure out what clicking the box does. But while that proclamation may be true, abiding by it is a major hindrance to effective use, so we make concessions and rely instead on the reasonable assumption that what we envision to be the design intent is, in fact, the design intent, and save a lot of time and energy in the process. When those assumptions turn out to be wrong, that often turns out to be the fault of bad design—or at least we are quick to blame the designer.

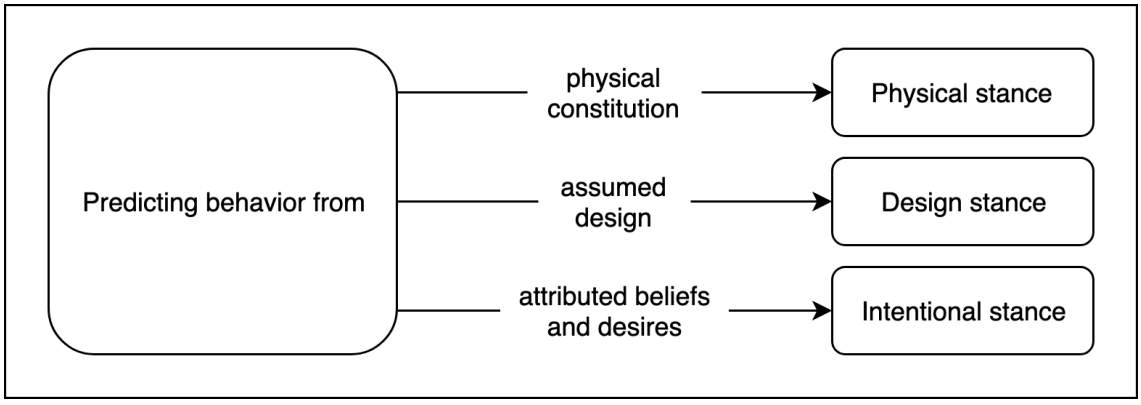

Figure 3. The three primary predictive strategies in brief.

The design stance has utility even in cases where *there is no actual designer.* When the system behavior exhibits patterns that seem to serve some particular purpose, we see that system as a designed system. For instance, the heart (and basically any other organ) is easily seen as a designed system due to its particular behavior in the larger system it inhabits. (Dennett 1987, p. 17) But it only has a designer insofar as we are willing to call natural selection a designer, which is what we are intuitively inclined to do, despite full knowledge that natural selection possesses no intentionality to design with—that is how strong the temptation to see systems as designed is!

4.3. Intentional Systems

When the design stance is practically inaccessible, the next step up is the intentional stance. In the case of computer programs, assuming the intentional stance would mean that the patterns of behavior exhibited by the program are not adequately predictable by assuming the program is designed.

This is plainly the case with something like a chess machine. Modern expert chess machines teach themselves to play the game through machine learning, which makes their behavior not a direct result of design in the first place, but even in the case of older or simpler deliberately designed chess machines, attempting to play against them using the design stance will likely prove futile. For one, a good designer will obfuscate the design in order to make the machine harder to figure out. But more importantly, chess is a deeply psychological game—attempting to find out what the opponent is thinking is a major part of the game. In order to play well, one has to adopt the mindset that *there is something to figure out* about the opponent, and that involves attributing the beliefs a good player holds regarding the game of chess, the desire to win and to play as a good player should, and the intent to act in accordance of those beliefs and desires.

How else would one gauge which game strategy would work against the opponent? You have to assume there is a possibility that the opponent “realizes” what you are attempting and counters your tactics, and that the opponent may “know” something you do not—otherwise there is no substantial difference between different tactics. Those assumptions require intentional attribution, so unless you want to challenge an opponent in chess without any strategy at all, you would find the intentional strategy invaluable—even if the opponent on the other end is a simple machine.

The beauty of the intentional stance is that it is largely agnostic towards technological competence. It places machines like Deep Blue and AlphaGo in the same category, whereas Kaplan and Haenlein would put them squarely in “expert system” and “AI”, respectively. The problem with defining AI with the kind of specificity that relies on some particular technological advancement or a particular interpretation on what constitutes intelligence is that such a definition is likely to become outdated as new technological advancements are made or when we come to understand the human mind better, and as a result major accomplishments in the field of AI will retroactively not be considered AI—which, as we have discussed, has apparently already happened with Deep Blue and other programs of its kind.

Using the intentional stance as a benchmark for AI solves that problem—it assumes nothing about what technology the program utilizes, or even what intelligence actually is. It merely asserts that *if one is inclined to use the same predictive strategy towards a program as one would towards an intelligent creature, the program is exhibiting, at the very least, a strong resemblance of intelligence.* Without granting any specifics about the constitution of intelligence, this is all that we should expect AI to be—an artificial *resemblance of* intelligence. It offers a well-defined threshold for AI programs to cross, too: if the design stance is a better or more intuitive choice, then the program is highly unlikely to seem intelligent, and thus should not be considered AI.

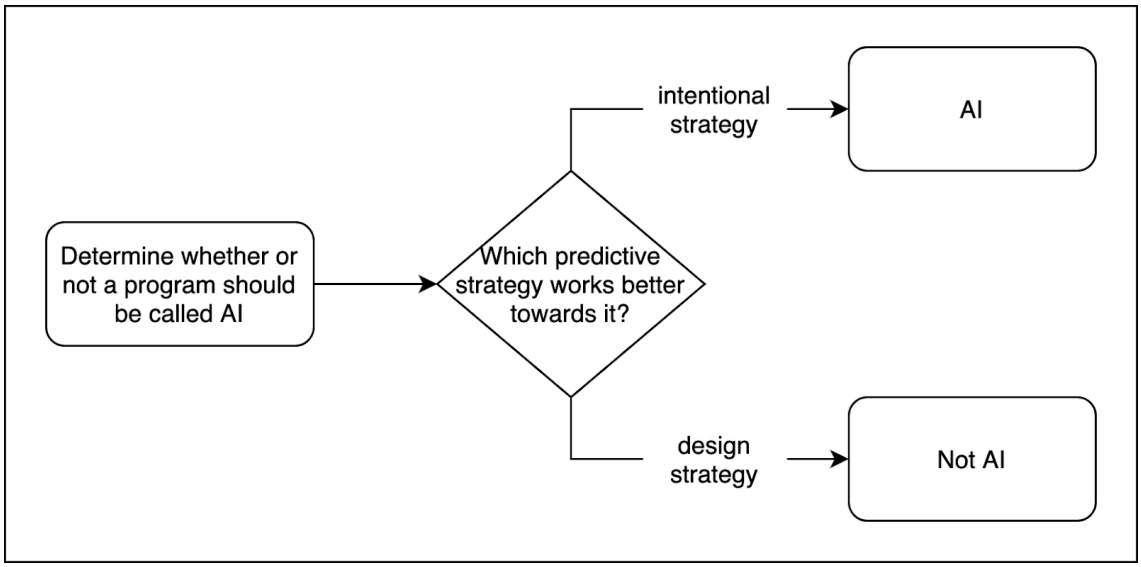

Figure 4. A simplified model for defining AI through the intentional stance.

Of course, this definition has its problems. For one, we are inclined to assume the intentional stance towards systems that, given conscious deliberation, we can see plainly as designed or physical systems and would not deem to be intelligent in the way that humans are intelligent. The Mark III Beast is a good example of this. Despite Dirksen's insistence that she "can do without the anthropomorphisms", she cannot help but see the machine as an intentional system with clear beliefs and desires. The inclination to assume the intentional stance overrides the rationalization that the machine is a designed system, because the machine appears to act for its own sake and not for the sake of a designed purpose. The Mark III Beast is an edge-case where the stance assumed towards it is in flux between intentional and design, but in such cases typically the intentional stance follows intuition, while the design stance has to be "wrestled away" from intuition with rational reasoning. This is more plainly the case if we equip the Mark III Beast with weapons to defend itself with. In a split-second survival scenario, a question such as "what does it want from me?" is a lot more

accessible than “what did the designer have in mind when programming this behavior?”, even if both questions are valid.

Once you get to know the system well enough, the design stance will likely become the more clear and accessible choice. After all, in the story, Hunt is no longer phased by the Beast’s behavior, even though at some point he might have been. As long as you are mostly ignorant of the system’s capabilities, the intentional stance seems superior, as you would have to work under the assumption that it is capable of more than it has shown you. When encountering an animalistic object such as the Mark III Beast, the first instinct is to compare its behavior to other objects you are more familiar with, and choose the stance accordingly—and since no better comparison than animals exists, you choose the intentional stance. Once you do get to know the object intimately, you learn its limitations and flaws, and the “wariness” of your stance is no longer necessary, leaving the design stance as the best option—unless the system really is so complex that no such limitation can be exposed, or the limitations are not sufficient in making the design stance a better option.

AI systems are often inconspicuous enough that their limited outputs to the user (as well as inputs from the user) do not allow the user to expose the limitations easily, if at all, making it so the intentional stance might always remain the best option. The Beast, being a physical object, can be tinkered with freely, but the same is not true with programs. Of course, the Beast does have a program component as well, but the system as a whole can be subjected to tests that an ordinary piece of software cannot. It is through these tests that one would arrive at the conjecture “it only recognizes danger from its own kind”—it is derived from scenarios which show what the machine is not programmed to respond to. Programs, by contrast, often do not allow inputs that give non-response, and if they do, it tends to be by mistake.

This is why the Turing test is such a high standard; it presupposes a capacity to be subjected to tests where the inputs are general enough to induce non-response or errors, meaning a system where such errors are not encountered is believed to demonstrate general intelligence. Most, outside of such a test scenario, would simply limit inputs so the illusion of the system’s capability never falters. Conversational programs tend to not have this luxury, so they are likely to be among the first to cause a falling back to the design stance—and even they can be credited for being *initially* accessible through the intentional stance, as long as they are competent enough.

The story of the Mark III Beast is of course fictional, but it elucidates on an important aspect of our nature—we anthropomorphize (which is just another way of saying

“treating as an intentional system”) things all the time, especially animals. With animals it is especially the case that we tend to judge their level of intelligence based on how well our anthropomorphizing works as a way to explain how they act. If we find that an animal’s behavior is explained by attributing to it a belief or desire that is “human-like”, we tend to interpret that as the animal being intelligent in some fashion. A common example of this is the class of beliefs that take the form of “I believe that if I do x, then y will happen”, which presuppose that a system holding such beliefs has an awareness of a causal relationship between x and y, as well as an awareness of its ability to enact change.

The concepts of intentionality and intelligence tend to get conflated a lot, which is not at all surprising. For one, the term “intentionality” is not one most people are familiar with, but beside that, the way we determine intelligence in a given system is by treating the system as an intentional system, since the behavior we look for in the system is how it acts according to the beliefs and desires we attribute to it. We could say that while the Mark III Beast is not “human intelligent”, it is at least “animal intelligent”, since it appears to act according to beliefs and desires we would expect an animal with intelligence to have.

Besides animals, who we can say to have intelligence, we do also anthropomorphize entirely static or mechanistic systems, with variable success. Dennett even makes the case for assuming the intentional stance towards thermostats (Dennett 1987, pp. 29–33), which, if you recall, is the example Searle used as well—but his doctrine of original intentionality made it so that he could not take it as seriously as Dennett does. While even in those cases the intentional strategy might prove to be effective, there is a distinction to be made between this kind of “utilitarian” assumption of the intentional stance and the assumption of the intentional stance that comes with the inclination that the system in question is intelligent. Even if one were to utilize the intentional strategy towards a thermostat to great effect, that does not come with the inclination that one should think of the thermostat as an intelligent agent; the strategy simply shows itself to be an effective one, and that is why it is used.

4.4. The Precession of Intentionality

The crux of the argument is this: the intentional stance begets the appearance of intelligence, but objects can be treated as intentional systems without also being treated as intelligent systems. When plants, for instance, grow in accordance to maximum absorption of sunlight, treating that behavior as intentional is a perfectly valid and effective strategy, without supposing that plants are intelligent.

However, treating a system as intelligent without treating it as intentional seems to never be a valid strategy. One can have the most powerful supercomputer calculating digits of pi at staggering speed, but since its function is plainly accessible from the design stance, there is little reason to attribute beliefs and desires to it in order to explain its behavior. Very few, if any, people would see this system as intelligent, but if we instead had the computer running a highly sophisticated conversational program, it is very likely to appear intelligent. Yet nothing at all changed about the processing power and (potentially) number of operations per second, and more importantly, it is still “just doing calculations”.

The key distinction, I argue, is the strategy used to predict its behavior. The design stance is no longer accessible because we cannot discern a design from the “hidden layers” we have no access to, so assuming the intentional stance becomes necessary in order to make useful predictions about the system’s behavior. It is very difficult, if not impossible, to conceive of an intelligence (artificial or otherwise) that does not necessitate one to assume the intentional stance towards it. If the design stance is perfectly sufficient, then the intelligence one looks for is that of the designer, not of the system itself—hence, it could be argued that the design stance is necessarily accompanied by the assumption of the intentional stance towards the envisioned designer. This is why it is so alluring to think of natural selection as an intentional system—we expect to find “intelligence” in its designs, and when we fail to find it, we treat those anomalies as if we ask Mother Nature herself, “now why would you make this the way it is?” It raises some fascinating questions about how we came to believe in the supernatural in the first place, but that is a Pandora’s box we ought to not open in this text.

One might want to say that intelligence is a kind of intentional state, though less obviously so than beliefs and desires. It can be characterized as such virtually by definition, as you can be intelligent *about* something, or *direct* intelligence *at* something. But that does not really reflect well on the holistic view we take on intelligence—beliefs and desires are individual “units”, so to speak, but intelligence cannot be split up like that. You either are or are not intelligent, and you might have more or less of it, but you do not have individual “units” of intelligence that the whole thing is made out of.

Likewise, intelligence is not a state you “enter” into. It is pre-existing, ready to be utilized wherever necessary. By contrast, a belief about something begins to exist only when you begin to believe that thing. So intelligence does not really fit precisely as an

intentional state *per se,* but it appears to be remarkably close—or at the very least, intrinsically entwined.

Rather, the appeal to intentionality seems to stem from the fact that intelligent behavior presupposes other intentional states and powers of representation. For instance, for us, solving a problem requires holding representations of the objects in play in mind, and having associated beliefs about how those objects function and interact with each other. When we perceive intelligence in artifacts, we compare that behavior to what is required in us to manifest that same kind of behavior, and in us the answer is always "intentional states"—so, the easy way to comprehend how artifacts can be intelligent is by attributing intentionality to them.

5. Fear the Philosophical Zombie

Now that we have laid out the main thesis of this work, let's take a moment to discuss what may seem like a major oversight, which is the fact that we have barely talked about consciousness. We mentioned Dennett accusing Searle of claiming that unconscious semanticity is impossible, but aside from that, we have not really addressed consciousness at all, and that may seem like we are missing something important. So allow me to briefly address that hypothetical concern.

In philosophy of mind, there is a thought experiment known as the *philosophical zombie argument,* which states that it is possible to conceive of a "zombie" that in physiology and behavior is exactly identical to a human, but lacks all conscious experience, qualia and sentience. This zombie would respond to stimuli exactly like a human would, and when asked if she was truly conscious, she would insist that she is. But the responses are simply outputs to given inputs—there need not be any conscious deliberation of those responses. It then follows from this conception of a zombie that a purely physicalistic conception of mental activity cannot account for consciousness, and that poses a problem for the claim that a mind, conscious and all, can be simulated by a program. (Kirk 2009) A machine can, too, insist that it is conscious, but that, according to the philosophical zombie thought experiment, does nothing to show that the machine is conscious, and it is in fact *impossible* to show that a machine does have consciousness.

This is strikingly similar to Searle's view that programs can never match the mind because of the problem of underivability of semantic content. Searle does not contest that a program could respond in exactly the same way as a human would, but he does insist that such a program cannot possess intentional states. In Dennett's assertion that

Searle's issue is actually about consciousness, Dennett essentially states that Searle is proposing the philosophical zombie argument, but from a different angle—which, again, makes Searle's claim that he is not a dualist highly questionable.

So why am I not afraid of philosophical zombies? Well, my problem is with how the argument relates to natural selection. As Hofstadter put it, the argument makes the case that consciousness adds no value to our survival if zombie-ness is a viable alternative, since both zombies and non-zombies would act identically and thus are just as competent at survival. Hence, it follows that the argument conceives of consciousness as an "added bonus" that is "tacked on" onto a structure that would behave in exactly the same way without consciousness. (Hofstadter 2007, pp. 324–325) I believe, alongside Hofstadter and Dennett, that this is fundamentally wrong. Consciousness, whatever it precisely is, is built so deep and is interlocked with so many of our other mental facilities, that separating it does not leave us with a mind to begin with. To quote Hofstadter (with his explicit permission):

> [The philosophical zombie argument] assumes that consciousness is some kind of orderable "extra feature" that some models, even the fanciest ones, might or might not have, much as a fancy car can be ordered with or without a DVD player or a power moonroof. But consciousness is not a power moonroof (you can quote me on that). Consciousness is not an optional feature that one can order independently of how the brain is built. (Ibid.)

Therefore, Hofstadter and Dennett, among others, argue that what the proponents of the argument are conceiving of is not, in fact, what they claim they are conceiving of, and it does not follow that we can build programs that handle inputs and outputs in exactly the same way that humans do without also incorporating consciousness (Ibid.). We might get close enough that in everyday interactions the differences are unnoticeable, but that is not the same as *identical.* If consciousness plays any role at all in our processing of inputs and outputs, then there *will* be differences, and there will very likely be tests that we can use to tell where those differences lie.

In my mind, any good theory of consciousness must either explain how it could have come to be through the process of natural selection, or be in direct opposition of the theory of natural selection. I dismiss the latter on the basis that I believe that natural selection is very real and fully accounts for our existence, consciousness and all. Accepting that as the case, it follows that the question "what does consciousness add to our capacity to survive?" is presented from a flawed premise; consciousness must have added something, *or else we would not have it.* The question is presented in a way that leaves "nothing" as a viable answer, which is nonsensical. If the zombie conception were correct, specimens with "more consciousness" would not have had a

benefit over other specimens, so there would not have been a reason why they fared better. This view cannot be accounted for in our current understanding of natural selection. The nature of random mutations also dictates that consciousness is not something that is “on” or “off”, and it must have been built up piece by piece over countless generations, leaving little, if any, room for non-physicalistic conceptions of consciousness.

Consciousness is, obviously, still very mysterious. We do not know what role it plays in natural selection, but we can be very sure, by the fact of our very existence, that it does play *some* role. Furthermore, there is no reason to suppose that a product of natural selection cannot be replicated in a Turing machine. In the worst case scenario, you would have to simulate the process of natural selection itself, but that still does not make it impossible. That is why I am not concerned with the problems consciousness (or any other facility of the mind, for that matter) may pose. Of course, they *are* problems, but not unsolvable ones.

Besides, consciousness has little to do with attribution of intentional states or intelligence anyway, for precisely the reason that proponents of the zombie argument state—we cannot tell for sure whether a system is conscious, so attribution of intentionality cannot be contingent on consciousness. As such, the problem of consciousness does not really affect my main thesis at all—but maybe that needed to be stated specifically. You may also consider this a defense of general AI, I suppose.

On a related note, an interesting consequence of the philosophical zombie argument is that, seemingly, zombies can possess intentional states and powers of representation *without consciousness.* Since zombies are supposed to be identical to humans except for consciousness, then their *beliefs that they are conscious* are actual, sincere beliefs that they hold, and they will also believe each other on those claims. It is fascinating that, somehow, intentionality is not off-limits for physicality, but consciousness is. Why draw the line there?

6. Representation and Interpretation

6.1. This Is Not a Pipe

Figure 5. René Magritte's *The Treachery of Images* (1929). The caption reads *"Ceci n'est pas une pipe."*, French for "This is not a pipe."

In chapter 4 I alluded to a peculiar conclusion about intelligence—that intelligence is not a thing in itself, as it is not an intrinsic property of the system in which it is perceived to exist. Rather, intelligence is purely a representation of something which does not exist in reality—a *simulacrum.* Philosopher Jean Baudrillard (1994) held that simulacra are "copies without originals" which exist in a state of *hyperreality,* where reality and simulation of reality become indistinguishable from each other. Intelligence exists as an intrinsic property in the representations of objects we form in the mind, but the property cannot be traced back to an intrinsic property within the object which is being represented. Yet, we cannot intuitively make this distinction—representations in our minds are *de facto* held to be reflections of reality, because otherwise we would not make sense of the world we inhabit.

This tendency of the mind to treat representation and reality as the same is beautifully illustrated in René Magritte's *The Treachery of Images* (pictured above). "This is not a pipe", it says—a representation of a pipe is not a pipe in itself. We point to

representations and say they are what is being represented—or even simply what we *think* is being represented. There is no pipe in Magritte's painting, but we also do not know what it represents. It could represent an actual pipe that Magritte painted a still life of, or it could represent no actual pipe in existence, and the representation only exists in relation to *other* representations of pipes (or representations of representations of pipes… and so on). Simply looking at the painting does not reveal the fact of the matter. Magritte further illustrated this disharmony between reality and representation in *The Two Mysteries* (pictured below).

Intelligence is no different than the pipe in the painting. It is a representation of something we are wholly ignorant of. There could be some particular constitution that flawlessly produces intelligence, but surely such a platonic ideal is not what we actually represent. Consider that figuring out what produces intelligence is the fundamental question in AI, and after a massive multi-disciplinary undertaking, we only have a vague idea of what the answer is. Despite that, we can instinctively tell intelligence from non-intelligence, all without knowing the slightest bit about the inner workings of the thing we are perceiving. It has to be the case that the representation has no original, and yet, we say that objects are, in themselves, intelligent. I do that too, and have done so throughout this text.

In 1966, Joseph Weizenbaum created a chatbot called ELIZA, intended as a parody of the way a psychotherapist speaks in an initial interview with a patient. Weizenbaum meant for the program to demonstrate how superficial communication between humans and machines is, but to his amazement, people conversing with ELIZA actually grew emotionally invested in it and found the "person" on the other end to be genuinely understanding of their concerns. (Weizenbaum 1984, pp. 2–7)

ELIZA worked better than most chatbots likely because of the psychotherapeutic style of reflecting what the patient says back to them and asking simple follow-up questions, which are fairly easy to simulate effectively. Nonetheless, in modern day people are exposed to chatbots often enough that an encounter with ELIZA would likely find them falling back to the design stance, as people have become accustomed to finding chatbots getting stuck in loops or misunderstanding what was being said, thus forcing users to choose their words with the question "how do I most effectively get what I want from this bot?" in mind—hence, thinking of the design first. But in 1966 it was certainly the case that the simulacrum of intelligence was found in ELIZA, even though Weizenbaum was sure that there is nothing to ELIZA's constitution that should produce "actual" intelligence—thus demonstrating that it is the stance towards the object, not the constitution of the object, that prescribes intelligence.

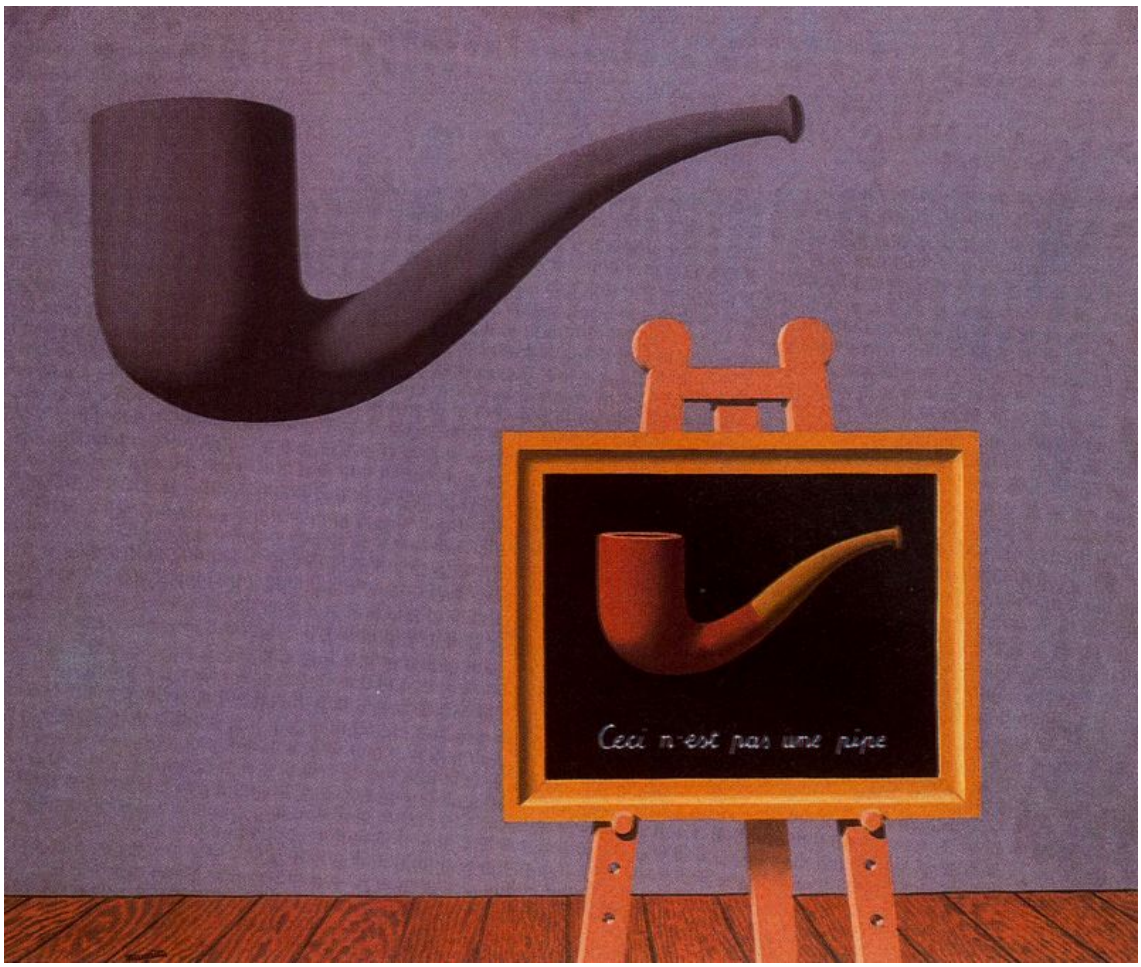

Figure 6. René Magritte's *The Two Mysteries* (1966). This painting seems to suggest that the pipe in *The Treachery of Images* is a representation of another representation. Or does it? What does this painting as a whole represent?

6.2. Inflated Interpretations

Like previously mentioned, we conflate representation and reality because it is necessary in order to make sense of the world around us. We tend to inflate the interpretations we make to convey just the right amount of "meaning" so that we do not miss the "point". Dennett illustrates this with an example of an evolutionary trait: a wide spectrum of creatures from fish to humans are

> [...] equipped with special-purpose hardware that is wonderfully sensitive to visual patterns exhibiting symmetry around a vertical axis. There can be little doubt about what the deflationary description is of the content of these intricate transducers: they signal "instance of symmetry around vertical axis on the retina." But why? What is this for? The provision is so common that it must have a very general utility. (Dennett 1987, p. 303)

The thought experiment then follows with the observation that basically the only things in the natural world that present vertical symmetry are "other animals, *but only when*

they are facing the viewer!" (Ibid.; Dennett's emphasis). Hence, an inflated, yet very satisfying, interpretation of the trait is that it allows creatures to instinctively tell when someone is looking at them. It is very easy to see how that would provide a massive benefit for survival, as it allows one to assess potential threats quicker. Likewise, it is easy to see why we would find this interpretation in particular so appealing: it explains a phenomenon in terms of a function and a purpose, which is valuable information that, in turn, helps us assess potential threats better. If we know how things function, we know better how to react to them. See, I just created an inflated interpretation for the evolutionary benefit of creating inflated interpretations!

It ought to be stressed just how inflationary these interpretations are—as we have previously discussed, Mother Nature does not infuse any purpose into her designs by herself. All these traits are functions that evolved over the course of millions of years through random mutations that fared better at propagation than others. But this granular view of natural selection offers us much smaller benefits than the inflated interpretations we make of purpose and function.

Thus, we have arrived back at the intentional strategy. It, too, centers around our inclination to create functional interpretations for phenomena, and the benefits of using it are immediately apparent. Dennett notes how in some fields there is an almost inescapable utility in attributing function by intentional vocabulary. He quotes the following passages from Alexander Rosenberg's *Intention and Action Among the Macromolecules,* which Rosenberg in turn quoted from L. Stryer's *Biochemistry:*

> A much more demanding *task* for these enzymes is to *discriminate* between similar amino acids. . . . However, the observed *error* frequency in vivo is only 1 in 3000, indicating that there must be subsequent *editing* steps to enhance fidelity. In fact the synthetase *corrects* its own *errors*. . . . How does the synthetase *avoid* hydrolyzing isoleucine-AMP, the *desired* intermediate? (Dennett 1987, p. 314; Rosenberg's emphases)

Endowing macromolecules with intentionality is practically necessary in order to make a useful functional interpretation of their behavior. The same occurs whenever genes and natural selection are discussed. Likewise—and here we cash out this metaphor we have been building up—endowing AI programs with intentionality is practically necessary in order to see them as intelligent! *This is why the representation of intelligence manifested by system behavior is preceded by the intentional stance.* Intentional vocabulary is what enables us to make functional interpretations of intelligent behavior.

Recall how I claimed that plant behavior is typically seen as intentional without being seen as intelligent. Well, that was not entirely true. Plants actually *are* often seen as intelligent by botanists, despite the fact that plants do not possess the mechanisms conventionally believed to be required for intelligence (Worrall 2016). Seeing plant behavior as intelligent is simply a highly useful functional interpretation, and it should follow that this conception of intelligence as a functional interpretation also applies to other systems.

Practitioners of AI create AI programs with the express purpose of manifesting some sort of intelligent behavior. Given that this is the case, the functions of those programs should be described using intentional vocabulary. This offers us a very basic test for "AI-ness": check if the creators used intentional vocabulary to describe their programs, and if they did, then there very likely was not a good alternative to using it—hence, they would have used the intentional strategy, and thus satisfied the main condition for the representation of intelligence. It is not a perfect test, obviously, but it captures what I think is the "essence" of AI. AI's purpose is to create systems whose behavior is explained by a functional interpretation of intelligence, and that explanation is put into words using intentional vocabulary.

Placing the onus on the practitioners of AI to define success in AI by their narration and use of intentional vocabulary is a stance that was also taken by Philip E. Agre in *Computation and Human Experience.* Agre makes the argument that

> [...] an AI system is only truly regarded as "working" when its operation can be narrated in intentional vocabulary, using words whose meanings go beyond the mathematical structures. When an AI system "works" in this broader sense, it is clearly a discursive construction, not just a mathematical fact, and the discursive construction succeeds only if the community assents. (Agre 1997, p. 14)

Jichen Zhu and D. Fox Harrell in their paper *System Intentionality and the Artificial Intelligence Hermeneutic Network: the Role of Intentional Vocabulary* (2009) also follow the works of Dennett and Agre, and touch on many of the same themes as I have. In their analysis of Hofstadter and Melanie Mitchell's AI project *Copycat,* they found that its descriptions involved two languages: one technical and the other intentional. The technical narration describes a "stochastic local search program", whereas the intentional narration describes a "fluid analogy maker". Zhu and Harrell argue that "intentional vocabulary serves as the joint between the two languages and gives rise to system intentionality", which is fundamental in the conception of AI as a discursive practice. (Ibid.)

A similar juxtaposition between intentional and technical language can be found in the descriptions of SHRDLU, a conversational program developed by Terry Winograd in 1968. In an elaboration of a conversation between a test user and SHRDLU, the program's functions are described using intentional terms like "knows", "understands", and "assumes", interspersed with technical descriptions such as "heuristics" and "semantic rules". (see Hofstadter 1999, pp. 586–596)

Given that Kaplan and Haenlein expect AI systems to "correctly interpret" (which sounds an awful lot like "understand") and learn from external data, modern systems that fulfill those criteria are describable by intentional vocabulary virtually by definition. And yet, systems from decades earlier can be described in much the same way. It is this capacity for intentional narration that I believe fundamentally makes AI what it is, from its roots to current day.

6.3. Minds, Intentional States, and Intelligence

We have come a pretty long way from arguing that the mind can be simulated, to now claiming that intelligence does not require simulating the mind to begin with. So why the effort? Well, while it is not so much the case that a mind *in itself* is required, the intentional states that Searle was actually concerned with (namely, "understanding") were, in his mind—and in the minds of AI practitioners—products of the mind. Reference to the mind cannot be avoided, as I am concerned with those intentional states as well. AI has chosen to approach the mission to reproduce those intentional states with simulation of the mind, and to legitimize that approach, I have argued against Searle's conception of the mind, because not only do I believe that mind simulation can be done, but exposing Searle's views as analogous to a soulist's views opens the door to functionalist views. Recall that Searle himself pointed out the utility of attribution of intentionality, but in his view the attribution we do to each other is *not* for the sake of that very same utility, but rather an acknowledgement of something we innately have. Arguing for intelligence as a functional interpretation of a wide variety of behaviors becomes valid *only if* the Searle-esque presuppositions regarding the mind —and its supposed privileged access to intentional states—are dealt with first, and that is why the approach taken in this work is what it is.

7. Discussion

In this work I have argued that the idea of intelligence presupposes that there is something that is intelligent, and we observe it as such. That "something" which has intelligence, and which acts as the ultimate point of reference for all other things

intelligent, is us—humans. Intelligence, in our case, is simply something we have decreed to have, and for something else to have intelligence means its behavior compares favorably to our own.

AI asks the question, “what are the fundamental constituents of intelligence and how do we simulate them?”, which is an important question if one wishes to replicate the kind of behavior observed as intelligent, but it is often confused with its much more grandiose counterpart, “can the human mind be simulated?”, due to the nature by which we tend to view intelligence as a *product of the mind,* rather than simply as a kind of a behavior. Hence, successes and failures in AI are judged based on how close they come to simulating the mind.

I have argued that intelligence is not really a thing unto itself, but rather it is an interpretation we make of certain given behaviors. We make those interpretations based on the strategies we employ to predict behavior, which in turn we have developed as a result of their utility in propagating our genes (i.e., natural selection). Intelligent objects call for different treatment than non-intelligent ones; that is why the distinction exists. Making an error of judgment in this distinction could result in a premature death, which makes it such an important evolutionary tool.

Intelligence, in other words, does not have any particular constitution, but certain constitutions *can* certainly create the type of behavior we interpret as intelligent. However, the revisionist nature of AI has had the effect that when we come to understand the constitution of the mind better, we dismiss constitutions that are too “simple” to have really caused intelligence, even if at the time of their creation their behavior was deemed intelligent. AI is in the unfortunate position of having to revise itself because our perceptions change as we come to understand the nature of the mind better. At some point, AI shifted from replicating intelligence, to being simply an application of certain methods and models.

I have also argued that the perception of intelligence is preceded by an assumption of a particular predictive strategy: the intentional stance, in which you treat the system you perceive as intelligent as intentional, i.e., possessing beliefs and desires by which it acts. The main thesis of this text follows from that argument, which is that the updating and diverging definitions of AI are reconciled if AI, as a whole, is characterized as a pursuit to create intentional systems. While AI’s successes in terms of intelligence can be dismissed with regard to constitution, they regardless were made with the intent of creating intentional systems first and foremost, and AI programs across the spectrum from early proofs of concept to invisibly acting AIs in consumer hardware are

characterized by their appeal to assume the intentional stance towards them. In short, I believe that a reasonable argument can be made that “AI program” can be defined as a program created with the intention of creating an intentional system, whose behavior is best predicted by assuming the intentional stance towards it.

So where does that leave us? Did I solve the problem of precisely defining AI? Of course not—fundamentally, what I believe I have accomplished is a new perspective on AI, not a replacement of old definitions. I believe that in our current climate of AI, the definition ultimately follows market forces, which I have no problem with. But, maybe the next time you hear the term “AI” used in a very buzzwordy sense, you might find that there is something deeper to it that warrants usage of the term.

A very easy objection that can be made of this work is that most of the major conjectures rest on the theories of Daniel Dennett, who, as is, is a fairly contentious philosopher. I do not personally find that a significant problem—I believe that the ideas of individual thinkers ought to be tested in different applications and see how they hold up. I think they do hold up quite well, and if you disagree, then at least I will have given you cause to think about the subject in a new way.

There is, obviously, the problem of scope. This work could be more comprehensive in basically every way, but the allocated time and resources are what they are. I would like to say that an easy suggestion for future work is testing the thesis against more existing theories and thoughts from different philosophers and people working in AI, but the limited scope makes it so that I am not even sure if the thesis merits that kind of work, since it might turn out that the tunnel vision a limited scope inevitably creates makes it so that the thesis has very limited applicability. I will likely have to come back to this myself once I have more comprehensive knowledge in order to make that assessment.

On a personal note, I am very happy with this work overall. It is the first extensive work of research I have produced, and it motivated me to take a deep dive into a subject that I actually had very limited knowledge of prior to starting this project. I leave the project not with feeling sick of it and never wanting to do this kind of thing again, which is a common sentiment I hear from other undergraduate students, but rather with an ever-increased interest to learn more, and branch out into new academic pursuits.

Like I said, my goal with this work was to offer a new perspective, presented in a way that is entertaining to read. If that perspective is something you do not agree with, that is fine by me, as long as you at least found it interesting and worth engaging with.

References

Agre, PE 1997, *Computation and Human Experience,* Cambridge University Press, New York.

Baudrillard, J 1994, *Simulacra and Simulation,* University of Michigan Press, United States.

Biography.com Editors 2019, *Alan Turing Biography*, Biography.com, viewed 20 May 2019, <https://www.biography.com/scientist/alan-turing>.

Copeland, J 2012, *Alan Turing: The codebreaker who saved 'millions of lives',* BBC, viewed 20 May 2019, <https://www.bbc.com/news/technology-18419691>.

Copeland, J 2017, *The Church–Turing Thesis,* Stanford Encyclopedia of Philosophy, viewed 20 May 2019, <https://plato.stanford.edu/entries/church-turing/>.

De Mol, L 2018, *Turing Machines*, Stanford Encyclopedia of Philosophy, viewed 20 May 2019, <https://plato.stanford.edu/entries/turing-machine/>.

Dennett, DC 1987, *The Intentional Stance,* The Massachusetts Institute of Technology, United States.

Elliott & Fry 1951, *Alan Turing,* photograph, National Portrait Gallery, NPG x82217. CC BY-NC-ND 3.0.

Gwern 2019, *Surprisingly Turing-Complete*, gwern.net, viewed 20 May 2019, <https://www.gwern.net/Turing-complete>.

Hodges A 2013, *Alan Turing,* Stanford Encyclopedia of Philosophy, viewed 20 May 2019, <https://plato.stanford.edu/entries/turing/>.

Hofstadter, DR 1999, *Gödel, Escher, Bach: an Eternal Golden Braid,* Basic Books Inc., United States.

Hofstadter, DR 2007, *I Am a Strange Loop,* Basic Books Inc., United States.

Hofstadter, DR, Dennett, DC 1981, *The Mind's I,* Basic Books Inc., United States.

Jacob P 2003, *Intentionality,* Stanford Encyclopedia of Philosophy, viewed 20 May 2019, <https://plato.stanford.edu/archives/sum2009/entries/intentionality/>.

Kahn, J 2002, *It's Alive!,* Wired, viewed 20 May 2019, <https://www.wired.com/2002/03/everywhere/>.

Kaplan, A, Haenlein, M 2018, *Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence,* ScienceDirect, viewed 20 May 2019, <https://www.sciencedirect.com/science/article/pii/S0007681318301393>.

Kirk, R 2006, *Zombies,* Stanford Encyclopedia of Philosophy, viewed 20 May 2019, <https://plato.stanford.edu/archives/sum2009/entries/zombies/>.

Magritte, R 1929, *The Treachery of Images,* painting, WikiArt. Free use.

Magritte, R 1966, *The Two Mysteries,* painting, WikiArt. Free use.

McCarthy, J, Minsky, M, Rochester, N, Shannon, C.E 1955, *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence,* Stanford, viewed 20 May 2019, <http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html>.

Miedaner, T 1981, 'The Soul of the Mark III Beast', in DR Hofstadter & DC Dennett (eds), *The Mind's I,* Basic Books, Inc., United States, pp. 109–113.

Minsky, M 1988, *The Society of Mind,* Simon & Schuster, Inc., United States.

Oppy G, Dowe D 2016, *The Turing Test,* Stanford Encyclopedia of Philosophy, viewed 20 May 2019, <https://plato.stanford.edu/entries/turing-test/>.

Searle, J 1981, 'Minds, Brains, and Programs', in DR Hofstadter & DC Dennett (eds), *The Mind's I,* Basic Books, Inc., United States, pp. 353–373.

Silver, D, Hassabis, D 2016, *AlphaGo: Mastering the ancient game of Go with Machine Learning,* Google, viewed 20 May 2019, <https://ai.googleblog.com/2016/01/alphago-mastering-ancient-game-of-go.html>.

Sipser, M 2006, *Introduction to the Theory of Computation,* Second Edition, Thomson Course Technology, United States.

Swaine, M 2007, *AI: It's OK Again!,* Dr. Dobb's, viewed 20 May 2019, <http://www.drdobbs.com/architecture-and-design/ai-its-ok-again/201804174?pgno=2>.

Turing, A 1950, 'Computing Machines and Intelligence', *Mind,* vol. 49, no. 236, pp. 433–460.

Weber, B 1996, *Mean Chess-Playing Computer Tears at Meaning of Thought,* NYU, viewed 20 May 2019, <http://besser.tsoa.nyu.edu/impact/w96/News/News7/0219weber.html>.

Weizenbaum, J 1984, *Computer Power and Human Reason,* Penguin Books Ltd., Great Britain.

Worrall S 2016, *There Is Such a Thing as Plant Intelligence,* National Geographic, viewed 20 May 2019, <https://news.nationalgeographic.com/2016/02/160221-plant-science-botany-evolution-mabey-ngbooktalk/>.

Zhu J, Harrell DF 2009, *System Intentionality and the Artificial Intelligence Hermeneutic Network: the Role of Intentional Vocabulary,* eScholarship, viewed 20 May 2019, <https://escholarship.org/uc/item/3rd2s695>.