Intelligent Machines

Raphaël Cherney

Microengineering Section

Professor: Christian Sachse

Supervisor: Pietro Snider

SHS Project 1st year Master

Report accepted on May 18, 2012

Lausanne, 2011-2012 academic year

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Table of Contents

Introduction

Since the advent of powerful digital computers, the field of artificial intelligence has been searching to understand and replicate the intelligence of the human mind. But what is intelligence? And can a machine think like a human? For years, philosophers, scientists, and laymen have been trying to answer these questions. This paper discusses the history of artificial intelligence and its philosophical underpinnings before presenting a new framework for intelligence articulated by Jeff Hawkins, inventor of the Palm Pilot and smartphone. Only with a better understanding of what intelligence is and how the brain does it can we ever hope to achieve truly intelligent machines.

Historical Artificial Intelligence

The Beginnings

The idea of a digital computer is a relatively old one. In the 1830s, Charles Babbage, Lucasian Professor of Mathematics at Cambridge, designed his own Analytical Engine. This programmable, digital computer was entirely mechanical, constructed using an intricate system of wheels and cards. Although it was never completed, it contained all of the parts and ideas central to a modern computer. As technology advanced, so did the computational power of these machines. In particular, the transition to electronic components such as electrical switches, vacuum tubes, and ultimately transistors greatly increased the speed, reliability, and manufacturability of computers. In the aftermath of World War II, electronic digital computers finally began to become available for broader applications. This is where the history of AI as we know it truly began.

In the summer of 1956, a workshop held at Dartmouth College defined the emerging field of AI. "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 simulated it." (McCarthy, Minsky, Rochester, & Shannon, 1955) In this way, the field of AI was created under the assertion that machines can at the very least simulate the complete workings of an intelligent brain. But can a machine be intelligent in the same was as humans? Can a machine *truly* think? Or does it just act *as if* it were thinking?

Alan Turing

Widely regarded as the father of modern computer science, the English mathematician Alan Turing also had a strong influence on the field of artificial intelligence and the philosophy surrounding it. During the early days of the digital computer, he demonstrated the concept of universal computation. That is to say, he showed that all computers are fundamentally equivalent regardless of how they are built. With only a memory and something to adjust that memory based on a set of rules (Turing called these the Store, Executive unit, and Control), he proved that one could perform any definable set of operations in the universe. Furthermore, all of these Universal Turing Machines could be programmed to do the exact same functions. This, to many, implied that the brain is a Universal Turing Machine. After all, we can perform computations just like a computer, and according to Turing, all processing machines are equivalent. With this mindset, AI researchers saw the goal in intelligence as imminently achievable.

In their influential 1943 paper, neurophysiologist Warren McCulloch and mathematician Walter Pitts described how neurons could perform digital functions. While they did not explicitly say that it was how the brain functioned, they implied that the brain could be structured using living, neuronal logic gates. The idea that nerve cell could replicate the formal logic of computers bolstered Turing's model of universal computation. If a computer and a brain could be made up using the same kinds of components, it seemed almost unimaginable that we would not be able to replicate the intelligence of a neural network on a computer.

Turing believed that computers could one day be intelligent, but rather than define intelligence, he reduced the problem to a simple question about conversation. He proposed what came to be known as the Turing Test: if a computer can have a conversation with a human and fool the human into thinking it was also a person, then the machine must be intelligent. After all, why should we have a higher standard for machines than we do humans? We typically have no direct evidence of the internal mental states of other humans.

There are several criticisms and things to note about the Turing Test. For starters, it is a purely behavioral metric (possibly due to the popularity of behaviorism in the first half of the 20th century). That is to say, it measures only the output of the system and is blind to the internals; if a computer *acts* as intelligently as a human, it is assumed to have human intelligence. This logic has several obvious flaws. Furthermore, it tends to limit intelligence to human-level thought. If the goal or artificial intelligence is to create the most intelligent machines, why would we set our metric as the ability to mimic people? As Russell and Norvig note, “aeronautical engineering texts do not define the goal of their field as ‘making machines that fly so exactly like pigeons that they can fool other pigeons’” (Russell & Norvig, 2010).

Regardless of the flaws, Turing's ideas proved extremely fruitful, and many AI proponents began their quest to create intelligent machines armed with the Turing Test as a goal and Universal Turing Machines as a way to get there. There are many parallels between computation and thinking. After all, some of the most impressive feats of human intelligence involve the manipulation of abstract symbols. Whether these symbols are words, physical objects, or bits stored in a computer, it doesn’t matter how they are implemented or manipulated because whether it is with a mechanical device, computer, or network of neurons in the brain, we can create the functional equivalent with any Universal Turing Machine. Or so it was believed. This was a powerful and attractive message. The field of AI researchers quickly grew and many honestly believed that it would be possible to create intelligent programs – ones that could certainly pass the Turing Test.

Limitations

There were numerous examples of AI experiments that showed promise in very specialized environments (for which they were specifically designed). A simulated room called Box World answered questions about its environment, a program called Eliza came close to passing the Turing Test by mimicking a psychoanalyst and rephrasing questions back at the interrogator, and in 1997, IBM’s Deep Blue famously defeated the world chess champion Garry Kasparov. However none of these examples showed any generalized intelligence. They had no flexibility, and even their creators admitted they did not think like humans. Deep Blue beat Kasparov not by being smarter, but by being millions of times faster than a human. While Kasparov could look at the board and tell what was a risky or secure position, Deep Blue naively computed countless possible moves. There was no intuition. Deep Blue played chess; it did not understand chess.

After years of failed experiments, funding started to disappear and by 1974 was difficult to come by. Programming even basic behaviors human began to seem impossible. There are still people who work on AI and believe that we can reach human-level intelligence with more complete algorithms and better machines. However, most have moved on and believe that the entire endeavor was flawed. We are far away from the AI we were promised.

As American engineer Jeff Hawkins puts it “Even today, no computer can understand language as well as a three-year-old or see as well as a mouse” (Hawkins, 2004).

Weak and Strong AI

Philosophers have been around far longer than computers, and they have been struggling with many questions related to AI. Much of it boils down to the following question: *Can machines think?* Can we make computers behave in intelligent ways? And if we can, would they have real, conscious minds? The American philosopher John Searle made the distinction between *simulating* thinking and *truly* thinking. In his seminal 1980 paper *Minds, Brains, and Programs* he proposes the following:

> “*I find it useful to distinguish what I will call ‘strong’ AI from ‘weak’ or ‘cautious’ AI (Artificial Intelligence). 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. In strong AI, because the programmed computer has cognitive states, the programs are not mere tools that enable us to test psychological explanations; rather, the programs are themselves the explanations.*” (Searle, 1980)

In other words, the idea that machines act simply *as if* they are intelligent is termed weak AI, while the idea that machines can have mental states like humans (i.e. have a mind) and are *actually* thinking is known as strong AI. Russell and Norvig point out, “Most AI researchers take the weak AI hypothesis for granted, and don’t care about the strong AI hypothesis — as long as their program works, they don’t care whether you call it a simulation of intelligence or real intelligence” (Russell & Norvig, 2010). Nevertheless, there is a very important distinction between these two interpretations of artificial thought processes.

In many ways, the weak AI assertion that machines can only aspire to simulate intelligence is extremely limiting. Computers can never hope to truly achieve human-level intelligence. Even if they acted in the same way as an intelligent creature, they would only be viewed as simulating – essentially *pretending* – to be intelligent. They will never be considered intelligent like a person is. Furthermore, since weak AI relies on simulating intelligent processes, we are limited by what the model is constructed on. In many ways, we can never hope to exceed human capacities for intelligence. On the other hand, strong AI, sometimes referred to as artificial general intelligence (AGI), allows for machines to match or exceed human intelligence and understanding. Because within strong AI machines can think and represent as we do, there is no real bound to their cognitive power. They can truly learn.

Common Arguments

There have been several arguments to the proposition that machines can act intelligently on any level (weak AI). One argument, the *argument from disability*, claims that there are things that computers will simply never be able to do. Turing himself proposed the following list of things that machines will never be able to do:

> "*Be kind, resourceful, beautiful, friendly, have initiative, have a sense of humour, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, make some one fall in love with it, learn from experience, use words properly, be the subject of its own thought, have as much diversity of behaviour as a man, do something really new*." (Turing, 1950)

Half a century after the fact, we know that computers can indeed begin to do some of these things. Many have seen computers make mistakes, computers can play games, they can learn from experience, and several have even made small but very real contributions to astronomy, mathematics, chemistry, biology, computer science, and many other fields. Today, there are many tasks that a computer can do better than a human (albeit without insight or understanding), but there are far more things at which computers do not excel.

Another influential argument is the *argument from informality of behavior*. This suggests that human behavior is too complex to be understood and distilled into a set of rules. There is simply too much going on, too many factors at play. Philosopher Hubert Dreyfus pushed this viewpoint in his books. He suggests that much of human intelligence is unconscious as well and therefore cannot be captured by formal rules. The "qualification problem" describes this inability to capture all aspects of decisions. Russel and Norvig are quick to point out that that many of the issues he describes with AI are specific to the implementation (in particular the first wave of AI research). They go on to state that, "many of the issues Dreyfus has focused on…have been incorporated into standard intelligent agent design. In our view, this is evidence of AI's progress, not its impossibility" (Russell & Norvig, 2010).

Chinese Room

In 1980, John Searle, a philosophy professor at the University of California at Berkeley, proposed the following thought experiment to show that computers were not and could not be intelligent:

> *"Suppose that I'm locked in a room and given a large batch of Chinese writing. Suppose furthermore (as is indeed the case) that I know no Chinese, either written or spoken, and that I'm not even confident that I could recognize Chinese writing as Chinese writing distinct from, say, Japanese writing or meaningless squiggles. To me, Chinese writing is just so many meaningless squiggles.*
>
> *Now suppose further that after this first batch of Chinese writing I am given a second batch of Chinese script together with a set of rules for correlating the second batch with the first batch. The rules are in English, and I understand these rules as well as any other native speaker of English. They enable me to correlate one set of formal symbols with another set of formal symbols, and all that 'formal' means here is that I can identify the symbols entirely by their shapes. Now suppose also that I am given a third batch of Chinese symbols together with some instructions, again in English, that enable me to correlate elements of this third batch with the first two batches, and these rules instruct me how to give back certain Chinese symbols with certain sorts of shapes in response to certain sorts of shapes given me in the third batch. Unknown to me, the people who are giving me all of these symbols 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."*
>
> *Now just to complicate the story a little, imagine that these people also give me stories in English, which I understand, and they then ask me questions in English about these stories, and I give them back answers in English. Suppose also that after a while I get so good at following the instructions for manipulating the Chinese symbols and the programmers get so good at writing the programs that from the external point of view that is, from the point of view of somebody outside the room in which I am locked -- my answers to the questions are absolutely indistinguishable from those of native Chinese speakers. Nobody just looking at my answers can tell that I don't speak a word of Chinese.*
>
> *Let us also suppose that my answers to the English questions are, as they no doubt would be, indistinguishable from those of other native English speakers, for the simple reason that I am a native English speaker. From the external point of view -- from the point of view of someone reading my "answers" -- the answers to the Chinese questions and the English questions are equally good. But in the Chinese case, unlike the English case, I produce the answers by manipulating uninterpreted formal symbols. As far as the Chinese is concerned, I simply behave like a computer; I perform computational operations on formally specified elements. For the purposes of the Chinese, I am simply an instantiation of the computer program."* (Searle, 1980)

The idea presented here is a strong argument that traditional computers and programs cannot be truly intelligent (strong AI). The person is mindlessly executing a series of instructions. Though the responses may appear intelligent, there is no understanding that takes place. No matter how well a computer can simulate intelligence, it will never have intelligence in the same way as the human with his English responses. In this way, Searle asserts that, while weak AI may be possible, strong AI is not; formal computations on symbols cannot produce thought. More generally, the thought experiment implies that one cannot get meaning (semantics) from simple symbol manipulation (syntax).

This argument created quite a stir with philosophers and AI pundits. There were many responses, the most common argument being what is known as the *systems reply*. This response concedes that the man does not understand Chinese, but that that room as a whole does. The person is only one part of the understanding system. Much like the person does not know Chinese, a CPU cannot compute the cube root of a number. It is true for both cases that processor alone cannot succeed but that the system as a whole does have the capacity. Searle's response to this is to suggest that the person internalize the entire system (memorize the instructions and calculate in his head). In this way, he contains the entire system but still does not understand Chinese. He essentially relies on the distinction that computer programs are formal while human minds have mental contents (Searle differentiates between what he calls the syntax and semantics). If we accept that these two entities are different, then the argument holds. Another response is the *virtual mind reply*. The idea here is that while the person does not directly understand Chinese, the running system creates a distinct mind that does understand Chinese. Much like a virtual character in a video game, the system creates a completely separate entity – in this case, one that knows Chinese. The last well-known response is the *robot reply*. This argument holds that the room needs some sensory input and the ability to interact with the world in order to develop an understanding of words. Essentially, without experience, it is all meaningless. The *other minds reply* argues that just as we attribute understanding to other people based on behavior without knowing what is in their head, and we should attribute understanding to the Chinese room by that same virtue. Finally, the *intuition reply* espouses that we should not undervalue the role of intuition in understanding and that we may be discounting the person's understanding too quickly.

Searle's ideas have broadly developed into the concept of *biological naturalism*. It argues that, "mental states are high-level emergent features that are caused by low-level physical processes *in the neurons*, and it is the (unspecified) properties of the neurons that matter. Thus, mental states cannot be duplicated just on the basis of some program having the same functional structure with the same input-output behavior; we would required that the program be running on an architecture with the same causal power as neurons" (Russell & Norvig, 2010). In other words, the connections and architecture of the brain are key to its ability to understand. He lacks a strong proof for this, seems to imply something unique (almost magical) about neurons that allow understanding and consciousness. Hawkins provides a rather compelling solution with his memory-prediction framework (discussed later). With such a system, the architecture and training (rather than programming) are key to understanding, somewhat fitting this view.

Functionalism

The strong AI hypothesis (i.e. that a machine can have a mind) has its roots in functionalism. Hawkins summarizes the idea nicely:

> "*According to functionalism, being intelligent or having a mind is purely a property of organization and has nothing inherently to do with what you're organized out of. A mind exists in any system whose constituent parts have the right causal relationship with each other, but those parts can just as validly be neurons, silicon chips, or something else. Clearly, this view is standard issue to any would-be builder of intelligent machines*." (Hawkins, 2004)

This idea implies that machines and humans can have the same mental states (given the same causal processes) and that machines can be intelligent in the same way as a human. As a thought experiment, consider if a chess game would be any less real if it were played with a salt shaker standing in for a lost knight piece? Clearly not, because it is functionally equivalent by virtue of how it acts (how it moves on the board and interacts with other pieces). Similarly, every few years, most the atoms that make up your body are replaced by new ones. In spite of this, you remain yourself in all of the ways that matter to you. Atoms are functionally equivalent. The same should be true for the brain. Consider if we replace each each neuron in your brain with a functionally equivalent micro-machine. Would anything be different? Functionalism suggests that you would feel exactly the same.

If we accept this principle, then an artificial system that uses the same functional architecture as an intelligent, living brain should be equally intelligent – and not just a simulated intelligence, but truly intelligent. On one extreme, we can imagine simulating the physics every single neuron and connection in the brain. The cognitive scientist Marvin Minsky writes that, "if the nervous system obeys the laws of physics and chemistry, which we have every reason to suppose it does, then…we…ought to be able to reproduce the behavior of the nervous system with some physical device" (Crevier, 1993). Few disagree that such a simulation of the brain is possible, and if we accept functionalism, then this simulation is intelligent as well. Nevertheless, John Searle points out that, "on the assumptions of strong AI, the mind is to the brain as the program is to the hardware, and thus we can understand the mind without doing neurophysiology. If we had to know how the brain worked to do AI, we wouldn't bother with AI" (Searle, 1980). In other words, if strong AI is true, then the essence of intelligence should not be fundamental to the brain or even its particular wiring, and therefore we should be able to create it in another way. To simply copy every component of the brain is to give up on trying to understand how intelligence works.

Hawkins warns that there is also a danger in going too far in the other direction (toward simple behavioral metrics). Some propose that we can generate efficiencies by finding an engineering solution to the problem, much like human-designed airplanes fly with fixed wings and jet engines instead of biologically modeled flapping wings. Hawkins contends that much like with Searle's Chinese Room, "behavioral equivalence is not enough. Since intelligence is an internal property of a brain, we have to look inside the brain to understand what intelligence is" (Hawkins, 2004). He goes on to admit that there are certainly unnecessary elements from our evolutionary past, but we must recognize that there are important features mixed in with the superfluous. Hawkins argues that studying and understanding the mechanisms of the brain is the only way to achieve truly intelligent machines. At first glance, it may appear as though Hawkins simply wants to remain as near as possible to a complete simulated brain – something that all functionalists can get behind. However, Hawkins proposes something far more radical.

The Flaw

As a longtime admirer of the neocortex, the portion of the brain responsible for almost everything we think of as intelligence (perception, language, imagination, mathematics, art, music, etc), Hawkins sees a serious flaw in the historical approach to AI. In particular, he cites the focus on *behavior* as defining intelligence as an intuitive but incorrect view. Much like in Searle's Chinese Room, understanding is very different from intelligent behavior. While reading the words on this paper, you are showing an incredible intelligence – despite the fact that no real behavior associated with this fact.

> "*I believe that the quest for intelligent machines has…been burdened by an intuitive assumption that's hampering our progress. When you ask yourself, What does an intelligent system do?, it is intuitively obvious to think in terms of behavior. We demonstrate human intelligence though our speech, writing, and actions, right? Yes, but only to a point. Intelligence is something that is happening in your head. Behavior is an optional ingredient. This is not intuitively obvious, but it's not hart to understand either…*
>
> *It's not difficult to understand why people – laymen and experts alike – have thought that behavior defines intelligence…All machines, whether made by humans or*

> *imagined by humans, are designed to do something. We don't have machines that think, we have machines that do. Even as we observe our fellow humans, we focus on their behavior and not on their hidden thoughts. Therefore, it seems intuitively obvious that intelligent behavior should be the metric of an intelligent system… However, looking across the history of science, we see our intuition is often the biggest obstacle to discovering the truth.*" (Hawkins, 2004)

In other words, what we lack in the field of AI is more than just powerful computers, but a proper framework of intelligence. We need to understand what understanding is in order to implement it, and it is clear that our current models are quite wrong. Hawkins proposes that "understanding cannot be measured by external behavior…it is instead an internal metric of how the brain remembers things and uses its memories to make predictions" (Hawkins, 2004).

A New Framework of Intelligence

Foundations

In 1978, Vernon Mountcastle, a neuroscientist at Johns Hopkins University published a paper titled *An Organizing Principle for Cerebral Function*. In this paper he proposes a common structure to the cortex. That is to say, the same kind of processing is taking place on auditory signals, visual signals, proprioceptive signals, and essentially all of the signals that your body receives; the cortex uses the same computational tool to accomplish everything it does. For years neuroscientists had been looking at the differences between parts of the brain while ignoring the important similarities. Much like Darwin noticing the significance of the biological similarities of birds in the Galapagos Islands, this discovery revolutionized thinking of the brain. Hawkins calls it "the Rosetta stone of neuroscience" (Hawkins, 2004), and in his Nobel Prize acceptance speech, David Hubel declared Mountcastle's discovery "the single most important contribution to the understanding of cerebral cortex since Ramón y Cajal"[1] (Hubel, 1981). His idea also suggests how the previous attempts at AI were misguided. While programmers tried to make computers understand language through

[1] Santiago Ramón y Cajal one of the first to study the microscopic structure of the brain and is considered the

grammar, syntax, and semantics, Mountcastle proposed that the brain (essentially the only thing we accept as intelligent) used a single algorithm, independent of the function or sense.

Despite the fact that it is widely known, an important and interesting aspect of Mountcastle's proposal is that all of the inputs to our cortex are basically the same. We think of our senses as very separate entities: vision is vision; hearing is hearing; touch is touch. However, the signals the brain receives are essentially the same, whether coming from the optic fiber, auditory nerve, or any other sense: they are all simply spatial-temporal sequences of neuron firings. In this way, it wouldn't matter to the brain if you perceived the world through sonar, radar, magnetic fields, or some other exotic sense. In fact, several animals make use of these unique senses (some birds and fish can sense the magnetic field of the earth). This brings up some interesting philosophical questions. If all our knowledge of the world is based on patterns (albeit predictable ones), how can we know what is real? After all, we can never know the world directly; we only have a model in our head that we update with our senses. Several science fiction works have played with this theme, but in reality, for the most part, our model of the world is based on a physical consistent reality. Yes, we can only interact with it in certain ways, but it is there.

A growing body of evidence supports Mountcastle's proposal. Every year, more research demonstrates the flexibility and generality of the cortical tissue. There are examples animals and people learning new skills in unique parts of the brain or repurposing old ones. In brief, "the cortex is not rigidly designed to perform different functions using different algorithms any more that the earth's surface was predestined to end up with its modern arrangement of nations" (Hawkins, 2004). If this is true, then what is it that the cortex does?

Memory-Prediction Framework

Hawkins proposes that the brain is not a computational processing machine, but rather a specialized *memory system*. He believes that the brain (the neocortex in particular) stores and plays back spatial-temporal patterns. Due to its hierarchical structure, the neocortex is able to generate *invariant representations* – consistent, internal representations of the world based on noisy sensory information. These abstractions, in turn, enable greater understanding. When we experience a pattern through our senses, our brain is constantly making predictions of what to expect next based on our memories. In fact, Hawkins argues that prediction "is the *primary function* of the neocortex, and the foundation of intelligence" (Hawkins, 2004). This may initially seem like a lot to accept, but there is a solid biological foundation to his claims. His theory also explains many of the difficulties of traditional AI, which thought of the brain as a computational organ. This new "real intelligence" is based on the interaction between senses and an auto-associative, highly parallel, hierarchical memory.

To prove his point on the role of prediction, Hawkins proposes the following altered door thought experiment:

> "*When you come home each day, you usually take a few second to go through your front door or whichever door you use. You reach out, turn the knob, walk in, and shut it behind you. It's a firmly established habit, something you do all the time and pay little attention to. Suppose while you are out, I sneak over to your home and change something about your door. It could be almost anything. I could move the knob over by an inch, change a round know into a thumb latch, or turn it from brass to chrome. I could change the door's weight, substituting solid oak for a hollow door, or vice versa. I could make the hinges squeaky and stiff, or make them glide frictionlessly. I could widen or narrow the door and its frame. I could change its color, add a knocker where the peephole used to be, or add a window. I can imagine a thousand changes that could be made to your door, unbeknownst to you. When you come home that day and attempt to open the door, you will quickly detect that something is wrong. It might take you a few seconds' reflection to realize exactly what is wrong, but you will notice the change very quickly. As your hand reaches for the moved knob, you will realize that it is not in the correct location. Or when you see the door's new window, something will appear odd. Or if the door's weight has been changed, you will push with the wrong amount of force and be surprised. The point is that you will notice any of a thousand changes in a very short period of time.*" (Hawkins, 2004)

He goes on to discredit the traditional AI approach of creating a database of properties for everything we interact with (it is not practical or possible for the brain to specify all the attributes of the world you experience and neurons are simply too slow to implement such a system). Instead, he offers the following:

> "*There is only one way to interpret your reaction to the altered door: your brain makes low-level sensory prediction about what it expects to see, hear, and feel at every given moment, and it does so in parallel. Al regions of your neocortex are simultaneously trying to predict what their next experience will be. Visual areas make predictions about edges, shapes, objects, locations, and motions. Auditory areas make predictions about tones, direction to source, and patterns of sound. Somatosensory areas make predictions about touch, texture, contour, and temperature.*
>
> *'Prediction' means that the neurons involved in sensing your door become active in advance of them actually receiving sensory input. When the sensory input does arrive, it is compared with what was expected. As you approach, your cortex is forming a slew of predictions based on past experience. As you reach out, it predicts what you will feel on you fingers, when you will feel the door, and at what angle your joints will be when you actually touch the door. As you start to push the door open, your cortex predicts how much resistance the door will offer and how it will sound. When your predictions are all met, you'll walk through the door without consciously*

> *knowing these prediction were verified. But if your expectations about the door are violated, the error will cause you to take notice. Correct predictions result in understanding. The door is normal. Incorrect predictions result in confusion and prompt you to pay attention. The door latch is not when it's supposed to be. The door is too light. The door is off center. The texture of the knob is wrong. We are making continuous low-level predictions in parallel across all our senses.*" (Hawkins, 2004)

There are countless examples and experiences that support Hawkins's view on prediction. Consider listening to a familiar song. You hear the next note in your head before it is even played. The neurons in your head are firing based on your memory of the song; your brain is predicting what to expect next. If someone alters the recording and changes a note, you will quickly realize that there was a mistake from what you remembered. It is also interesting to note that we are much more sensitive to intervals between pitches than the absolute note (very few people have perfect pitch). That it why you can recognize songs regardless of the key they are in. This is a hint of the invariance of the representation within your mind. This fact would also suggest that songs are stored in your memory as sequences of steps with less regard to any particular note (which similarly explains why you can understand speech from people with very different voices). Even with new songs, we have expectations. If you are familiar with Western music, then you will likely presume a regular beat, repeated rhythm, and expect songs to end on the tonic pitch. Even though you may not be aware of it, your brain is automatically predicting these beats and rhythms whenever you listen to new songs. It is largely unconscious, and your brain does not know exactly what will come next, but it makes predictions about what patterns are more or less likely to happen.

We all know of the incredible human capacity for memory, and the ability for an experience to "jog" our memory. Hawkins suggests that this this is more than a capacity to remember people, places, or events, but rather the capacity for intelligence itself:

> "*The human cortex is particularly large and therefore has a massive memory capacity. It is constantly predicting what you will see, hear, and feel, mostly in ways you are unconscious of. These predictions are our thoughts, and, when combined with sensory input, they are our perceptions. I call this view of the brain the* memory-prediction framework *of intelligence.*
>
> *If Searle's Chinese Room contained a similar memory system that could make predictions about what Chinese characters would appear next and what would happen next in the story, we could say with confidence that the room understood Chinese and understood the story. We can now see where Alan Turing went wrong. Prediction, not behavior, is the proof of intelligence.*" (Hawkins, 2004)

The idea boils down to the following: *to be able predict is to understand*. This is actually a surprisingly logical conclusion if we accept that there is nothing magical to the mind. Instead, intelligence is simply the ability to recall and abstract patterns. This is in contrast to the traditional AI mindset. It is the difference between *computing* a solution to a problem and *remembering* it. When you catch a ball, you do not calculate its precise trajectory through the air, instead you your brain recalls a temporal sequence of muscle commands and adjusts it to the particular situation.

Epochs of Intelligence

Defining Intelligence

I purposefully have pushed off defining intelligence until this point. For years, philosophers, linguists, and just about everyone else have proposed different meanings for the term *intelligence*. The dictionary definition of "the ability to acquire and apply knowledge and skills" (The New Oxford American Dictionary, 2009) is a reasonable, albeit broad, interpretation that accounts for many aspects that AI researchers also find important. In particular, the AI community believes that intelligent systems should be able to do the following (Russell & Norvig, 2010):

- Reason (use strategy, make judgments under uncertainty, etc.)
- Represent knowledge
- Plan
- Learn
- Communicate in natural language
- Generate goals

These are ambitious and high-level goals. Nevertheless, such a set of objectives still does not define intelligence. Russell and Norvig argue that there is a distinction between definitions based on *thought processes* and *reasoning* and those that address *behavior*. Similarly, we could classify intelligence based on *human* metrics or a more *ideal* measure, that they call rationality. If we want to consider machines that *act humanly*, then the Turing Test or other similar metrics would be logical. If, however, we are more interested in *thinking humanly*, then we may want to take a cognitive modeling approach. The *thinking rationally* approach to AI posits a "laws of thought" approach in which we must somehow understand and codify the aspects of intelligent thought. Finally, we have the *acting* *rationally* viewpoint that Russell and Norvig favor. They argue that “a rational agent is one that acts so as to achieve the best outcome, or, when there is uncertainty, the best expected outcome” (Russell & Norvig, 2010). This is a nice definition for computer scientists, because it is a behavioral metric (which we can somehow test with a machine) and is based on a standard, mathematical understanding of rationality. Furthermore, like many questions in computer science, it becomes a question of optimization – finding the “best” (however that is defined) outcome. They admit, however, that based on their definition of intelligence, achieving perfect rationality is essentially impossible (due to the complicated nature of environments). We know this to be true: assuming behavior shows intelligence, if we are given a problem that does not have a clear right or wrong choice, how can we say if the decision of an artificial agent was intelligent, regardless of what it decides. Nevertheless, this is the approach that many AI researchers use; they simply use some “performance measure” that defines what is successful or not for their particular artificial system.

As we have seen, however, there is a problem with the current framework for studying intelligence. Behavior is not intelligence, only a byproduct of it. If we accept the idea of the memory-prediction framework, then intelligence is based on the ability to predict and exploit consistencies within an environment. If our senses send us a consistent message that the molecules of a pencil tend to say together, then we may begin to understand and recognize it as an object. Despite that fact that it might move in our field of view or feel slightly differently the next time we touch it, there is a consistency to our senses around this object. We exploit this consistency to develop understanding. Over time, we can build a mental model or invariant memory of the object based on how it and similar objects behave or interact. We can then predict all kinds of things about the object: that it won’t magically disappear; that it will keep the same shape; how heavy it will feel; the sound it makes when hitting a table. This kind of understanding happens at all levels of mind (we build our higher-level understanding based on lower-level understanding). When we learn language, we begin by understanding sounds, then words, then sentences, then paragraphs, and finally ideas. If we didn’t first have a good representation of sounds, we would never get to the final ideas. Humans have an incredibly complex model (and therefore understanding) of the world. That is why have succeeded in being such a successful species: we can take very complex patterns and exploit them. We are incredible prediction machines.

I propose the following definition of intelligence: *intelligence is the ability to recognize and exploit patterns*. This pattern exploitation can come in the form of behavior, or it can simply be an altered mental state – the generation of a memory. In such a way, intelligent beings extract structure from the world and use it to their advantage. By this definition, in order to be intelligent, a system needs two important things. First, it needs an input – a collection of signals such as the ones living organisms receive from their senses. Second, we need a memory in which to build our model. With these two things and the proper structure, intelligence can occur. Note that this is a purposefully loose definition of intelligence, one that computers could easily fulfill. The question of artificial intelligence then becomes an issue of matching or surpassing human-level intelligence. I want to emphasize the fact that intelligence is a continuum, not some arbitrary threshold. It is not some test like Turing suggested. Intelligence can come in many forms. Humans are intelligent, no doubt, but so are many animals and even things we don't always think of as being intelligent. Hawkins puts it nicely:

> "*Everything I have written so far…depends on a very basic premise – that the world has structure and is therefore predictable. There are patterns in the world: faces have eyes, eyes have pupils, fires are hot, gravity makes objects fall, doors open and shut, and so forth. The world is not random, nor is it homogeneous. Memory, prediction, and behavior would be meaningless if the world was without structure. All behavior, whether it is the behavior of a human, a snail, a single-cell organism, or a tree, is a means of exploiting the structure of the world for the benefit of reproduction.*" (Hawkins, 2004)

With this understanding of intelligence, Hawkins sees three, distinct *epochs of intelligence* in the history of our world: evolutionary, experiential, and human.

Evolutionary

Consider a single celled organism in a pond. The cell has molecules on its surface that respond to the presence of nutrients. It also has a flagellum that allows it to swim around the pond. As it swims, the cell reacts to differences in the concentration of its food source and adjusts its swimming accordingly. In this way, the cell is able to exploit its chemical awareness of the environment. Is this cell intelligent? By our definition, yes. The organism is making a prediction that by swimming along the gradient, it will find the nutrients it needs. The cell also has a memory in its DNA. In a sense, this behavior was simply learned over years of evolution and stored in a DNA sequence. The organism receives patterns from the world and exploits them. It is intelligent, in a very basic way.

Similarly, plants exploit different aspects of the world. They take advantage of the pattern that water tends to be absorbed in the ground and sunlight is best received by growing upward. Every organism has its own ways of predicting and using the world for its survival, some simple, some more advanced. It is amazing what some animals have learned to do simply through millions of years of evolution. In this way, with mutation and natural selection as a learning strategy, we build up an intelligent structure to the world.

It is interesting to note that artificial evolution through genetic algorithms is a widely used learning method in engineering, robotics, and computer science. By creating an artificial genotype and fitness function, populations are evolved, leading to learning being stored in the artificial DNA. Using such methods, we have been able to optimize difficult problems and create intelligent behavior in robots (Floreano & Mattiussi, 2008).

Through the course of evolution, nature discovered the ability to communicate through chemical signals. If a tree is harmed in a particular location, it will secrete chemicals through its vascular system, which will trigger a defensive system to increase activity in another part of the tree. Over time, animals also developed similar systems. Neurons likely evolved to solve a similar problem as plants' vascular systems, but at a much faster rate. In time, the connections between these neurons became changeable, enabling faster learning. This leads us to the second epoch: experiential learning.

Experiential

Evolution can be a horrendously slow process. Initially, if an environment changed, many organisms would simply die. They had no way to adapt because evolutionary learning happens at an evolutionary pace. Consequently, the prospect of being able to learn from experience became incredibly attractive. In many ways, it seems almost logical that animals evolved nervous systems the way that they did. With the (evolutionary) development of neurons, animals were able to learn about the structure of the world within their own lifetimes. Then, if a new environmental threat arrived, the animal would no longer have to rely on genetically preprogrammed behaviors but could learn ones. This provides an incredible evolutionary advantage.

All mammals have a neocortex, or new brain, around their more primitive brain that allows this advanced learning. In this way, there are many animals that can learn as we do – animals that everyone would agree are intelligent. However, different animals' ability to make sense of the world is not equal. In large part, much of this is directly correlated with the size of the neocortex. The more cortical tissue, the more the animal can synthesize and make sense of the incoming patterns. We can imagine how evolution would favor animals with increasingly intelligence and complex behavior – eventually leading to animals like humans. Over time, animals evolved a powerful ability to understand the world, but humans have taken this to another level through advanced language.

Human

Two features distinguish human intelligence from that of animals. The first is that humans have a very large neocortex. A larger neocortex allows deeper analogies; it allows us to find structure within structure; and it helps us make more complex predictions. Secondly, humans developed advanced language. Many animals can communicate, but none has been able to transmit knowledge like humans have. Through communication we can learn more about structure of our world without having to experience it ourselves; we can build on the knowledge of others who we have never met. Whether it is spoken, written, or simply wrapped into cultural traditions, we are able to pass on more and more knowledge to younger generations. In this way, our understanding of the world has increased in leaps and bounds relative to the intellectual progression of animals.

In short, intelligence began as a slow, evolutionary process where understanding of the world had to be hardcoded in our genes. Eventually, we developed more flexible memories that allowed us to learn within our lifetime. Based on our own experiences (and those of others we could empathize with) we could adjust our model of the world and alter our behavior. Finally, with the development of a larger neocortex and advanced language, humans were able to distinguish themselves. Human beings have been able to extract the structure and exploit the patterns of our world unlike any other species. The next question is whether machines will join and surpass us.

Building Intelligent Machines

The Prospect

Hawkins argues that the development of intelligent machines is likely inevitable, but will likely look very different than what we have imagined or read about in science fiction novels. As Hawkins puts it, “There is no reason that an intelligent machine should look, act, or sense like a human. What makes it intelligent is that it can understand and interact with its world via a hierarchical memory model and can think about its world in a way analogous to how you and I think about our world" (Hawkins, 2004). These machines, while having mental states like humans, will lack the incredibly complex machinery of our bodies or the more primitive brain functions that lead to emotion. In this way, they will never understand or behave in the exact same ways as humans (and Hawkins argues that may be for the better). Nevertheless, they will be truly intelligent. These machines will develop a new level of understanding about the world. They may also take advantage of different and unique sensing capabilities. For example, one can imagine an intelligent machine that uses information about the weather from around the world to predict weather patterns down the road. This machine could sense, understand, and predict storms just as we know the sequence of notes in our favorite song.

Advanced intelligent systems will still require years of development, and even once they are built, they will need much specialized training. It takes a years for a child to learn how speak properly. Similarly, machines will need to experience enough of their world to be able to understand and predict it. Fortunately the training only has to happen once. It can then be replicated to other systems (and prevented from being altered). In this way, we can develop a smart car, for example, train it once, and replicate the memory in all of the other cars on the assembly line.

Hawkins doesn't presume to know what the future holds, but he believes that intelligent machines based on the memory-prediction framework can do great good for the world. The most significant changes we cannot even imagine (as is true with all revolutionary technologies). After all, artificial intelligent machines can run at a higher speed, can have a larger capacity, can use novel sensing systems, and can have perfect replicability. With these features, we can imagine developing systems far beyond the capacity of the human mind. The possibilities are nearly endless. As Hawkins says, "Our intelligent machines will be amazing tools and will dramatically expand out knowledge of the universe" (Hawkins, 2004).

Conclusion

The potential of intelligent systems is great, but is tied to a proper understanding of intelligence. After examining the pitfalls of historical AI research, we have presented a new framework for intelligence that shows promise for understanding and replicating the capacity of human mind. This paper only touches on the possibilities and implications of developing such artificially intelligent systems. There is still a lot of work that needs to be done, but with a proper framework for the problem, it may be achievable. We may be able to reach an all new epoch of intelligence were we can use a powerful distributed system to understand even more about our universe. As Alan Turing once said,

> "*We can only see a short distance ahead, but we can see that much remains to be done.*" (Turing, 1950)

Bibliography

Cole, D. (2009). The Chinese Room Argument. *The Stanford Encyclopedia of Philosophy* .

Crevier, D. (1993). *AI: The Tumultuous Search for Artificial Intelligence.* New York, New York: BasicBooks.

Dreyfus, H. (1972). *What Computers Can't Do.* HarperCollins.

Floreano, D., & Mattiussi, C. (2008). *Bio-Inspired Artificial Intelligence.* Cambridge, Massachusetts: The MIT Press.

Hawkins, J. (2004). *On Intelligence.* New York: St. Martin's Griffin.

Horst, S. (2009). The Computational Theory of Mind. *The Stanford Encyclopedia of Philosophy* .

Hubel, D. (1981). On the Primary Visual Cortex, 1955-1978: A Biased Historical Account. *Nobel lecture*.

McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. Dartmouth College.

McCulloch, W., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics , 5*, 115-133.

Russell, S., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach.* Boston, Massachusetts: Pearson Education Inc.

Searle, J. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences , 3* (3), 417-457.

The New Oxford American Dictionary (2nd ed.). (2009). New York, New York: Oxford University Press.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind , 59*, 433-460.

Van Gulick, R. (2004). Consciousness. *The Stanford Encyclopedia of Philosophy* .

Voss, P. (2007). *Artificial General Intelligence.* (B. Goertzel, & C. Pennachin, Eds.) Springer.