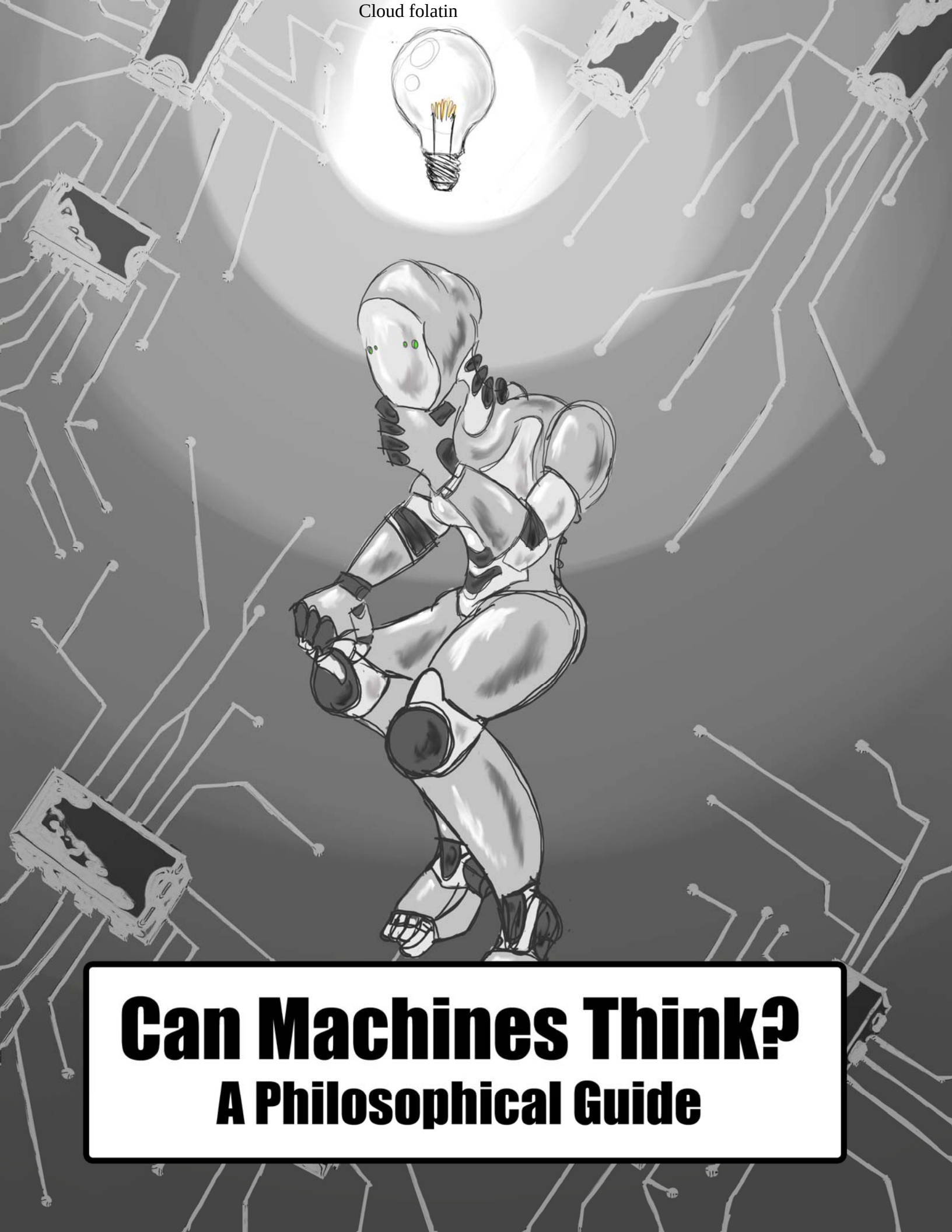
Cloud folatin



Can Machines Think?

A Philosophical Guide

This work represents my St. Mary's Project, part of the liberal arts curriculum of St. Mary's College of Maryland. It consists of a number of comics illustrating what I think are some key concepts in Philosophy of Mind generally and Philosophy of Artificial Intelligence generally.

The comics are arranged in a hybrid order based on both topic and chronology. Ideally they begin with our earliest, most commonsense views of mind and move towards our newest, more radical views. At the end of every section or subsection will be a black box with some questions to consider. These are not intended as mere summaries, but rather the questions raised are meant to spark further thinking about the topic. Each comic is done in a different style, sometimes to highlight certain thematic elements of the topic at large, other times for variety. While I use some technical jargon, I believe that these comics will be intelligible even to the philosophically-untrained reader. In addition to this document, the comics are also available online at macorrellsmp.wordpress.com, with all of the features that one expects from this whole internet business.

It is my goal that, after reading these comics, you will be prepared to tackle more dense works in modern Philosophy of Mind, armed with a familiarity with some of the key terms and positions in the field. And a good thing, too: the emerging field of AI will soon begin to pose philosophical questions that cannot be ignored. It is possible that by the close of the century advances in artificial intelligence will make questions of the moral standing and nature of artificial minds ones that are not merely idle speculation for philosophers but questions of the utmost importance on an international scale.

Too often philosophy lags far behind our scientific progress; such was the case with the harnessing of atomic power, to the loss of the world. The solution is not to artificially slow down the progress of science (if such a thing is even possible) but for philosophers to look past the horizon of current technology and so avoid being caught off guard by a shift in paradigms. A couple dozen pages of comics will not even begin to solve this problem, but hopefully this briefest of introductions will allow you to attain the background that will allow some more serious work on the problem.

This work would've been impossible without the patience of my advisor Michael Taber, the support of my family, the good humor of my friends, and the kindness of all of the above.

-Michael Correll, April 2009

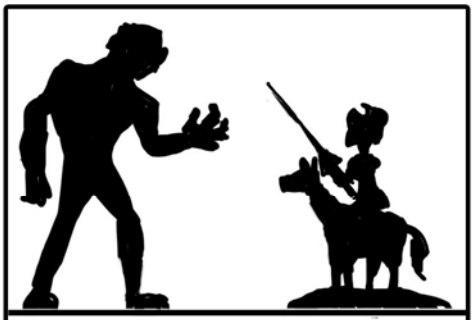
Computers already do a lot of things that some consider "thinking"; difficult math equations and game strategy beyond human ability. But will a computer ever be able to act in a play or create a great work of art? In short, can machine thought ever resemble human thought, with all of its creative and adaptive power?

That's what I think is the meaning behind the question, "Can Machines Think?" and I hope by exploring the subject the other tacit meanings of the question will become clear.

At first glance, the mind isn't very much like the body.

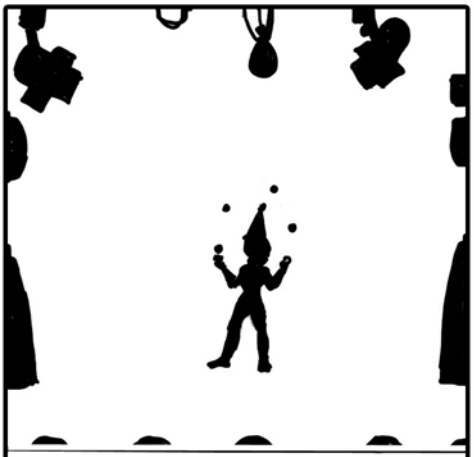
Our mind can direct the movement of our body by force of will alone, like a puppet on a string.

Our mind can take in information from our senses and use it to make a coherent picture of the world.



And while it is possible for our senses to be mistaken...

We cannot in principle be wrong about statements like "I think that something is true" or "I feel pain." Mental statements are special.



Our mind is also like a vast theater, capable of imagining impossible things in the greatest of detail.

In short, there seems to be a fundamental difference between what it is to be a mental thing or a physical thing.

One explanation is that there are two kinds of "stuff": mental stuff and physical stuff.

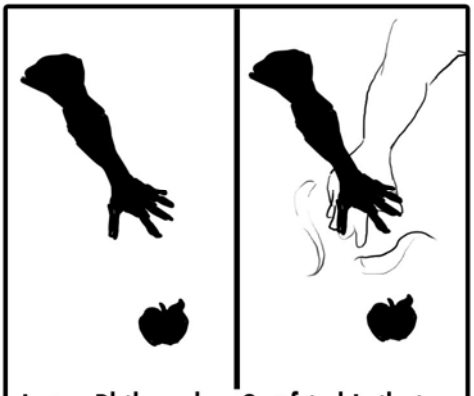
One explanation for these two kinds of stuff is that the mental and the physcial interact in an almost mystical fashion in the locus of the brain.

This view of the mind was championed by René Descartes, outlined in one of the first works of modern philosophy, *Meditations on First Philosophy.*

Earlier theories and theologies had only reinforced the notion that there is something that is like "you," and that this "youness" is somehow seperate and different from your body.

The question then becomes: how exactly does the mind interact with the body, if they both indeed are two unrelated substances?

Elisabeth of Bohemia, who kept up a correspondence with Descartes, thought the problem could not be solved while maintaining the duality of the physical and mental.



Later, Philosopher Gottfried Leibniz suggested an alternative: our minds are not actually causally connected to our bodies, but instead are linked as both mirror the mind of God.

As the scientific revolution began to pick up steam, the Cartesian answer to the problem of mind became less satisfying; formerly mystical things were shown to have naturalistic workings, and the brain was next.

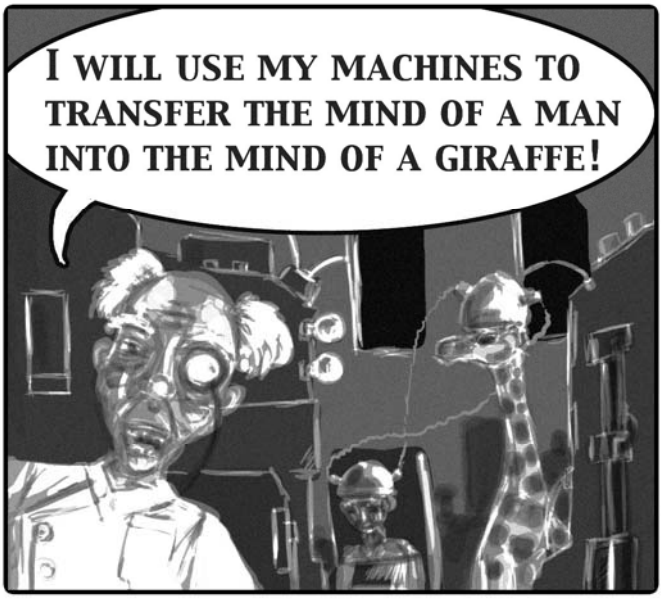
As science began to examine the brain in detail, it was found that body and behavior were linked in ways that nobody could've predicted.

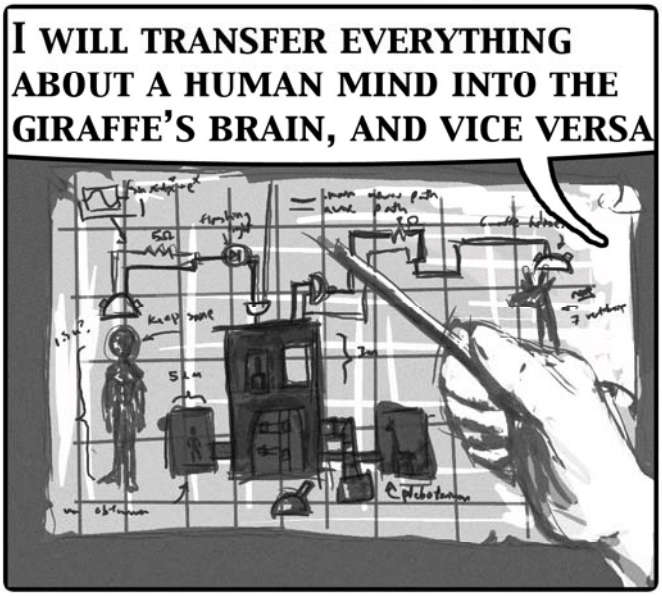
Only since the 20th century have we been able to examine the brain in any real depth, and see how the structure of the brain creates everything we identify as "ourselves."

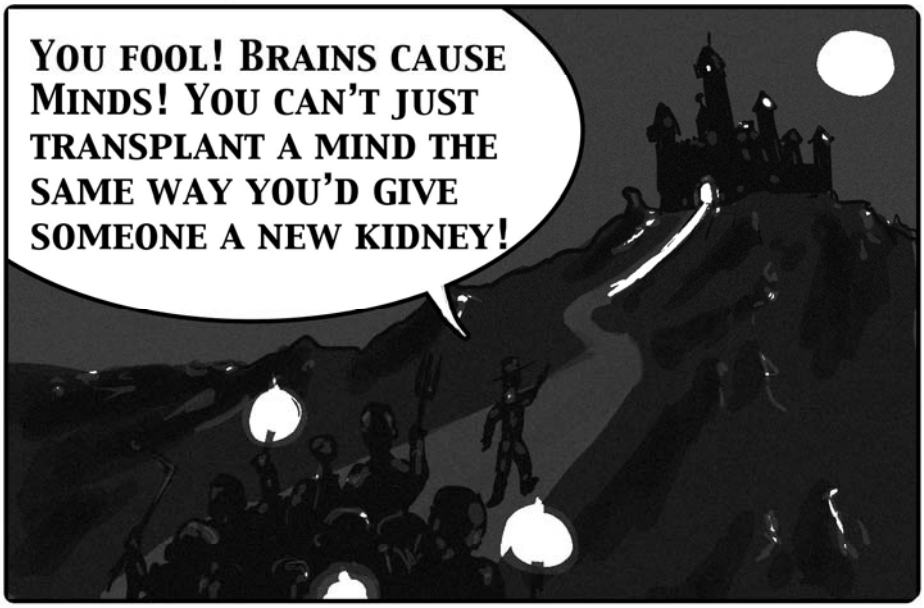
Very few philosophers call themselves "dualists" these days; by and large the universal position is that of there being only one physical substance, a view called "monism." If we accept dualism as Descartes describes it, can we still somehow engineer robots or computers that have what we'd consider minds? Does Artificial Intelligence require monism? What about science in general? Why is dualism such a common belief in human history, if it is false?

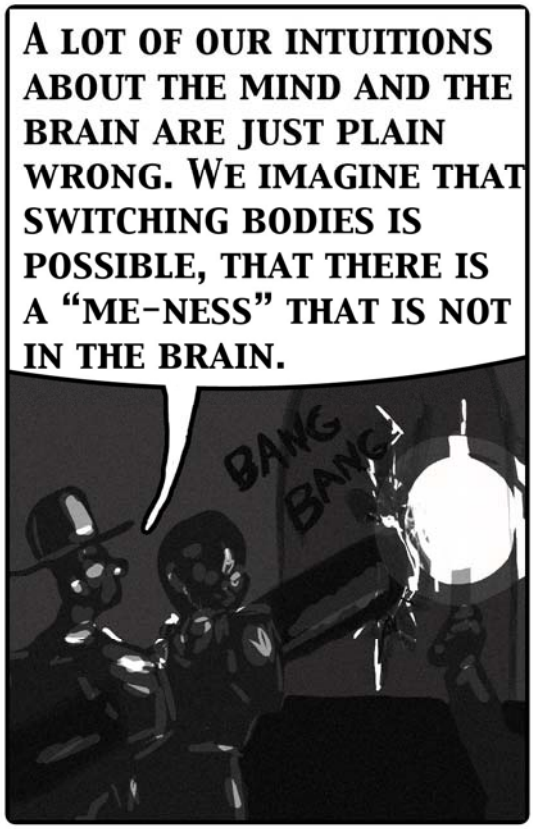

INSANE, THEY CALLED ME! MAD!


I'LL SHOW THEM!
KRAK


I WILL USE MY MACHINES TO TRANSFER THE MIND OF A MAN INTO THE MIND OF A GIRAFFE!

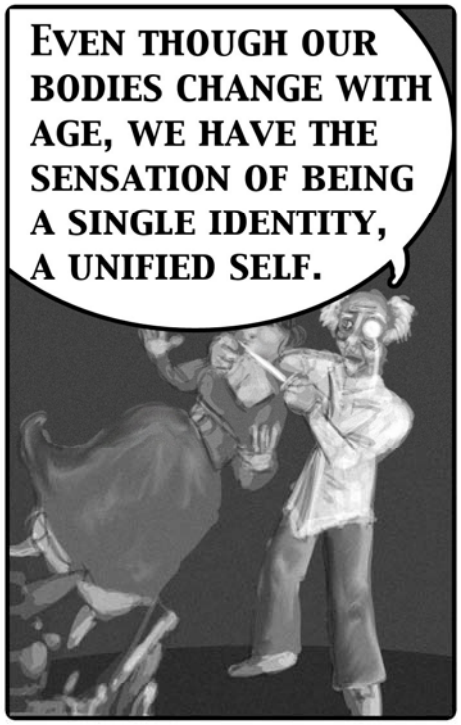

I WILL TRANSFER EVERYTHING ABOUT A HUMAN MIND INTO THE GIRAFFE'S BRAIN, AND VICE VERSA

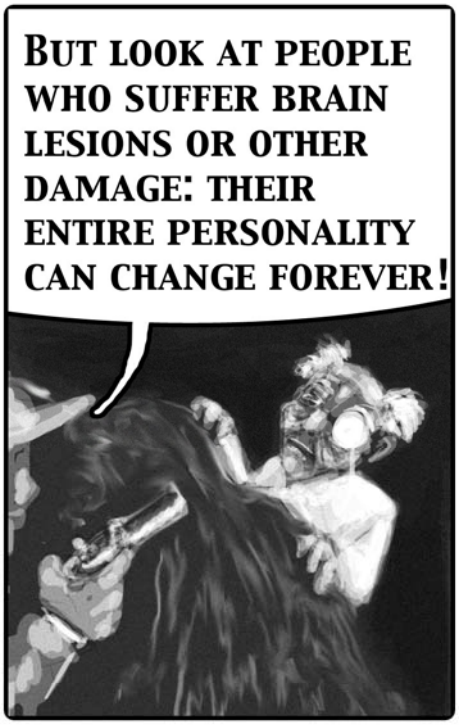

YOU FOOL! BRAINS CAUSE MINDS! YOU CAN'T JUST TRANSPLANT A MIND THE SAME WAY YOU'D GIVE SOMEONE A NEW KIDNEY!

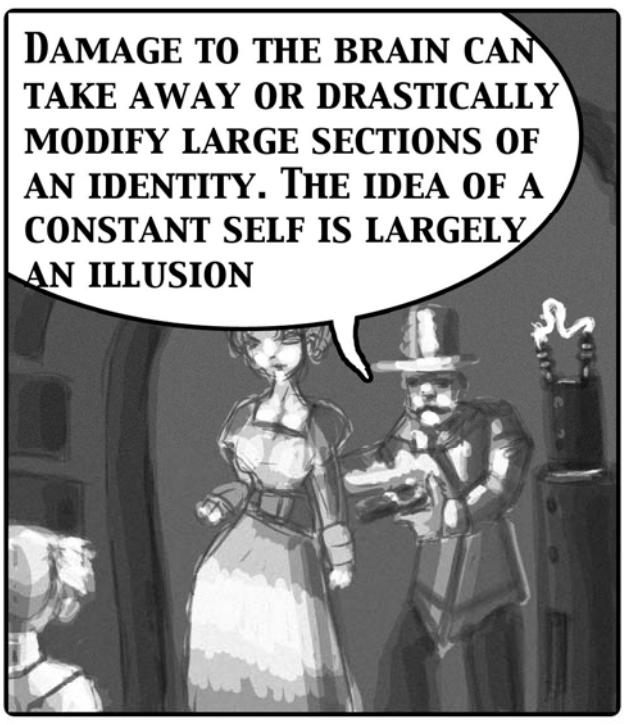

A LOT OF OUR INTUITIONS ABOUT THE MIND AND THE BRAIN ARE JUST PLAIN WRONG. WE IMAGINE THAT SWITCHING BODIES IS POSSIBLE, THAT THERE IS A "ME-NESS" THAT IS NOT IN THE BRAIN.
BANG BANG

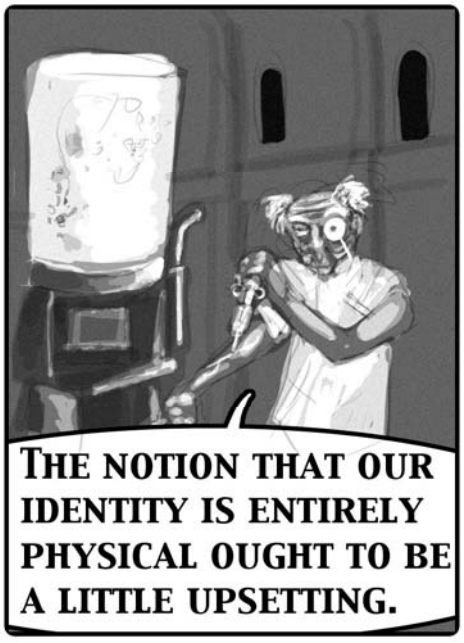

WE ARE UNNERVED BY CREATURES LIKE THE COCKROACH WHICH CAN SURVIVE FOR DAYS WITHOUT A HEAD.

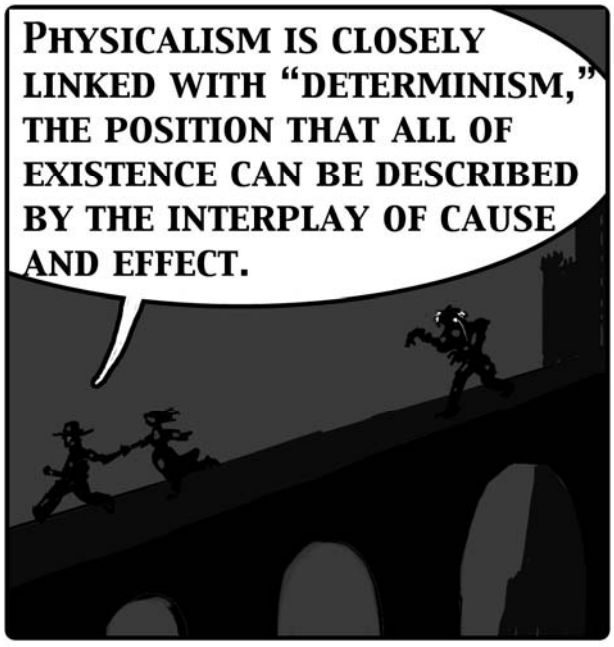

AREN'T BRAINS A LITTLE STRANGE, THOUGH? WE THINK THAT OUR BODY "BELONGS" TO US, BUT WE THINK THAT A MIND IS SOMETHING WE JUST "ARE."

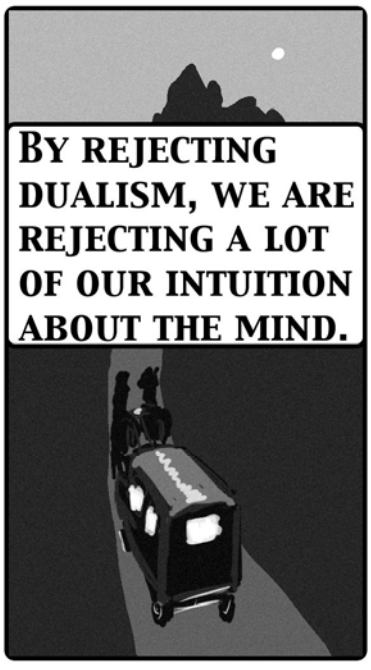


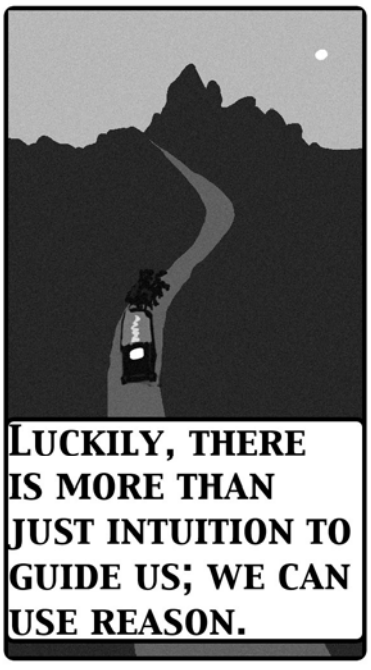


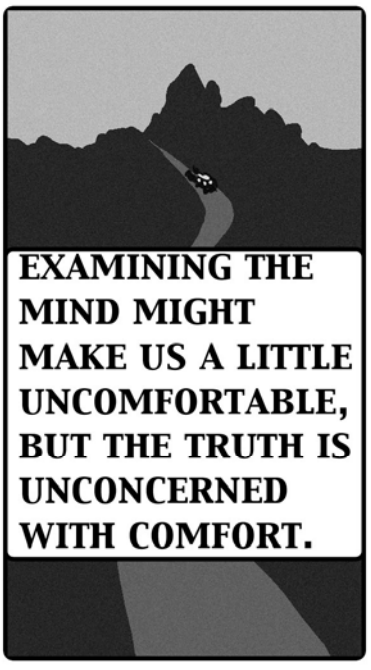




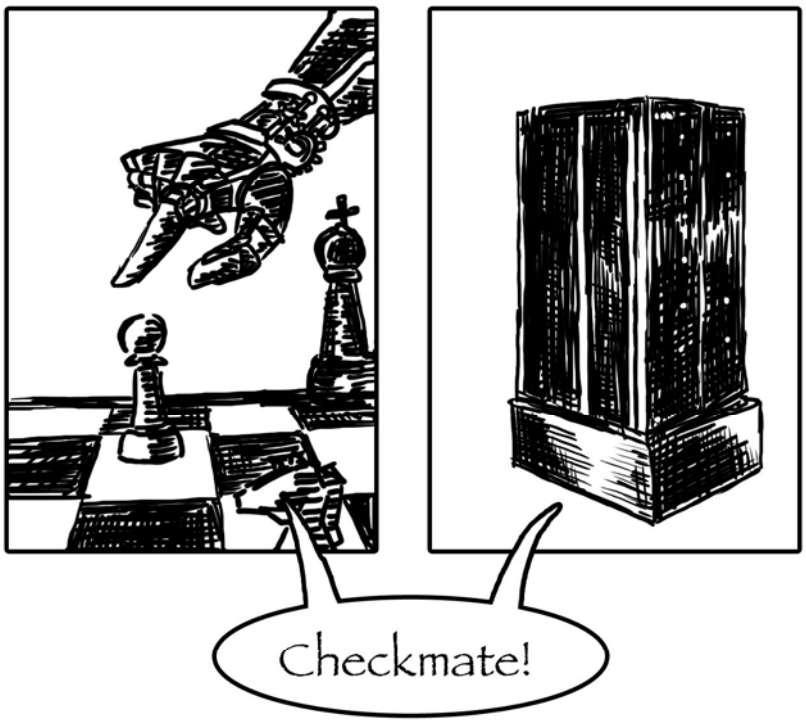




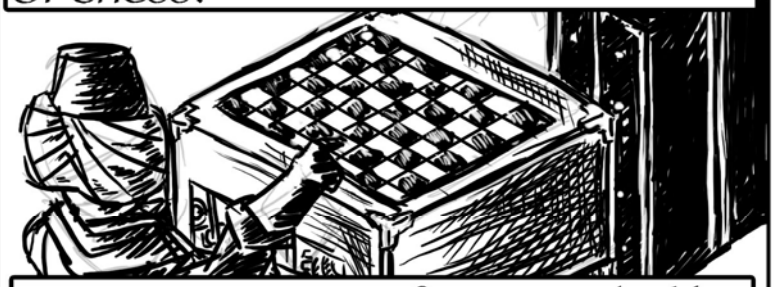




When does our intuition cease to be a useful philosophical tool when we are discussing the mind? Does it ever? Why do you think we tend to become uneasy when talking about determinism or free will?



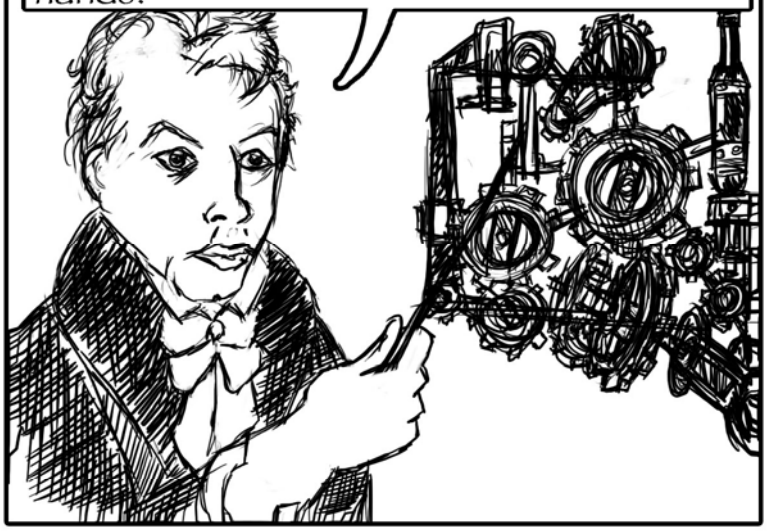
In 1770 Wolfgang von Kempelen created "The Turk," a clockwork machine that played a strong game of chess.



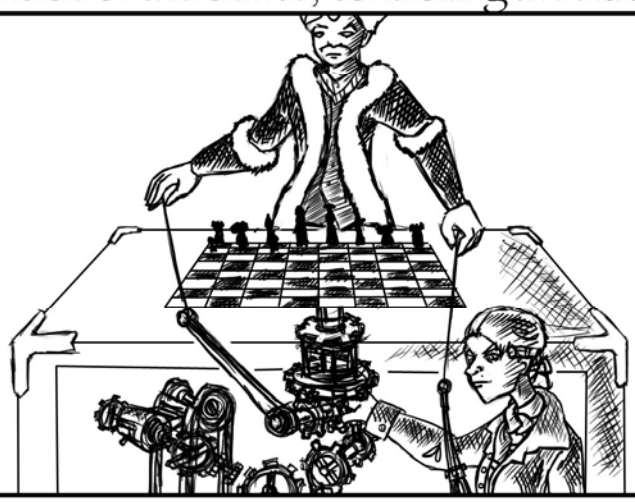
In 1997 a team from IBM led by Feng-hsiung Hsu used the "Deep Blue" computer to beat Gary Kasparov, chess grandmaster

This "mechanical Turk" defeated such opponents as Benjamin Franklin, and can even complete the "Knight's Tour" puzzle, where a knight must pass through every square of a chessboard once and only once.

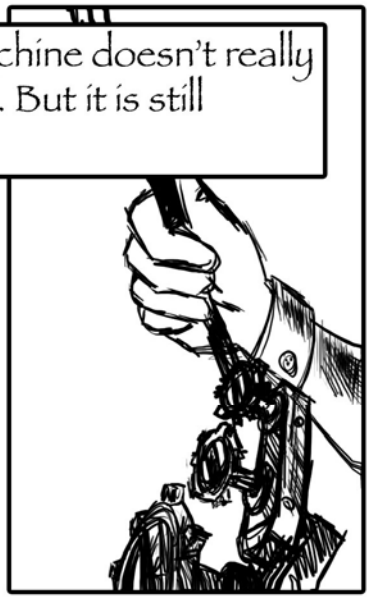
It operates via a series of complicated gears and springs that move the mechanical hands.



And, er... there is a chess player hidden inside of the device, controlling the machine.



Okay, fine. So the machine doesn't really play chess all by itself. But it is still impressive, right?

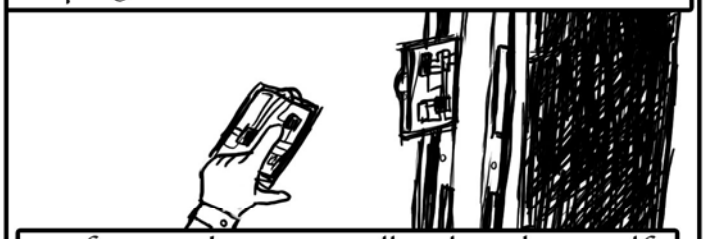


The "Deep Blue" computer, on the other hand, is quite the thinker.

The machine constructs a "tree" of possible moves from a given game position and then "trims" the moves that are not optimal

Since there are so many possible sets of moves in chess, we gave Deep Blue a list of thousands of grandmaster openings and closings.

And, um, we used our own chess masters to manually tweak the machines between games to adapt to our opponent's style of play.



So fine, it doesn't really play "by itself"

A lot of computer behavior that looks intelligent is the result of humans working behind the scenes.

Humans program their own expectations and shortcuts into machine behavior.

Computers are really good at following "algorithms," lists of instructions that always arrive at a right answer.

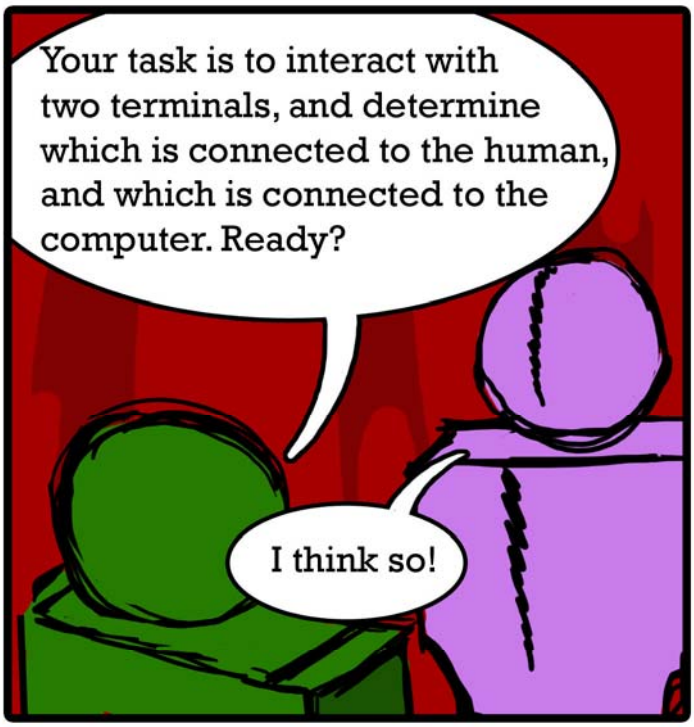
Recipes are a good example of an algorithm. Follow the steps correctly, and you'll make the recipe correctly

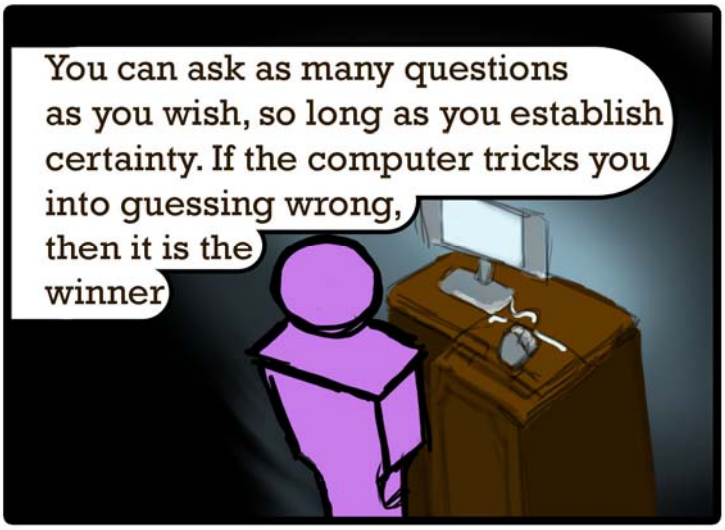
Computers aren't so good at "heuristics" sometimes known as "rules of thumb."

Saying "it looks like rain today, you should bring an umbrella" is a heuristic. It's not always accurate, but it is an important part of how we make choices.

Heuristics, in short, are patterns formed by lots and lots of experience. To make a machine that is more than just a mechanical Turk, we need to make a device capable of learning from experience, making its own rules of thumb. This type of flexibility is not easy to program into machines, but I don't think that it is impossible either.


From "Computing Machinery and Intelligence" by Alan Turing
Are you all ready to play...
IMITATION GAME
The Imitation Game!
Let's meet our two challengers!
One of our challengers is a computer program
And the other is a real live human being, typing into a terminal

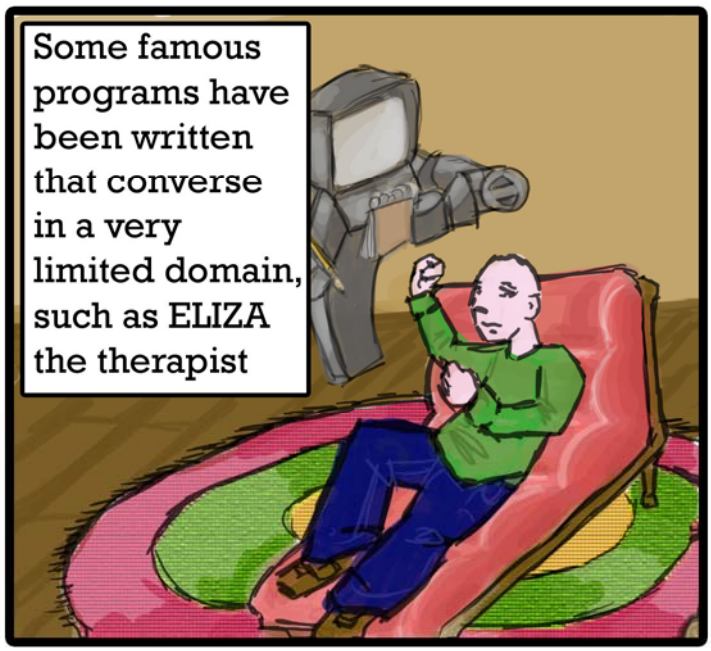


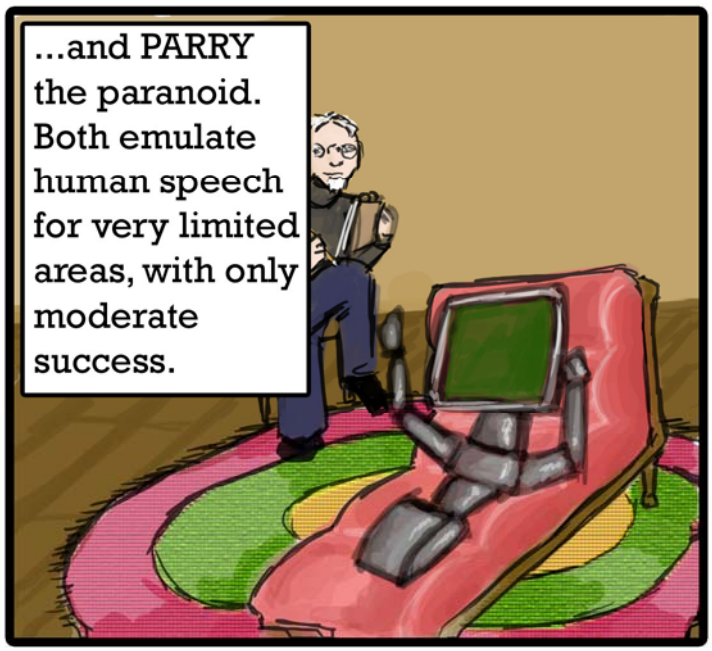


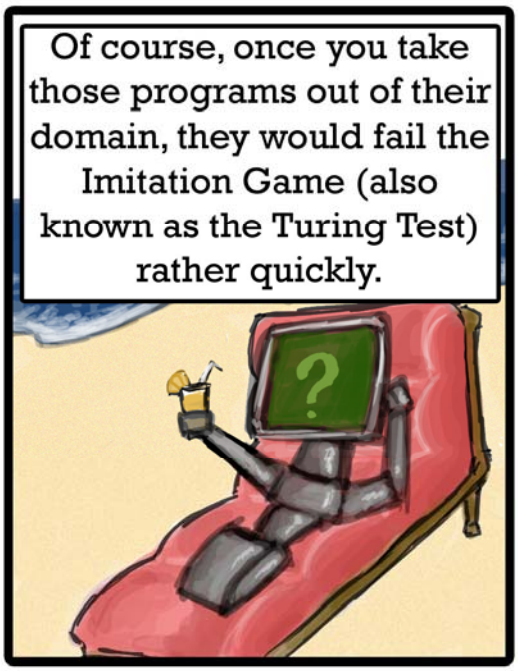
Judge: What is your favorite poem?
Player 1: I'm a fan of Coleridge, especially the Rime of the Ancient Mariner.
Judge: Why is that?

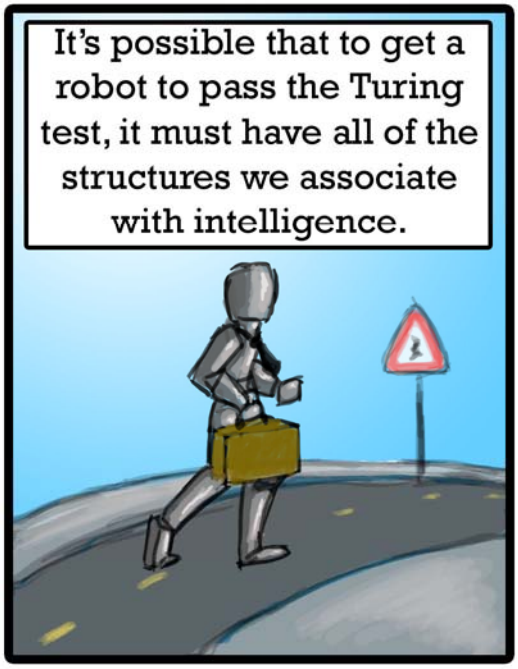
Judge: Do you ever feel sad?
Player 2: Of course. Into every life some rain must fall.
Judge: Can you give me any specifics?

If no amount of questioning provides an answer as to which is human, can't we consider the computer to functionally be the same as an intelligent human being?









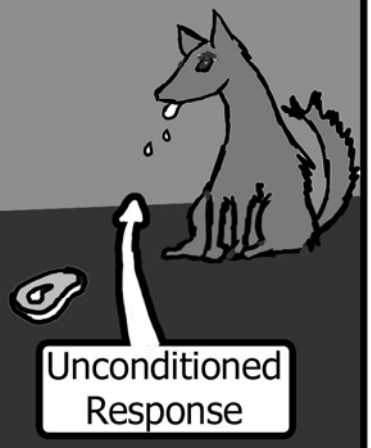
Is it possible to have a program that passes the Turing test, but isn't intelligent? Is it possible to have a program that passes the test at all, for that matter?
If it is impossible in principle for a computer to pass the test, what makes people different from computers such that we can pass with ease?

You are likely familiar with Pavlov and his experiment with dogs.

Unconditioned Stimulus

In it, a bell was rung whenever food was presented to a dog.

...causing the dog to salivate



Unconditioned Response

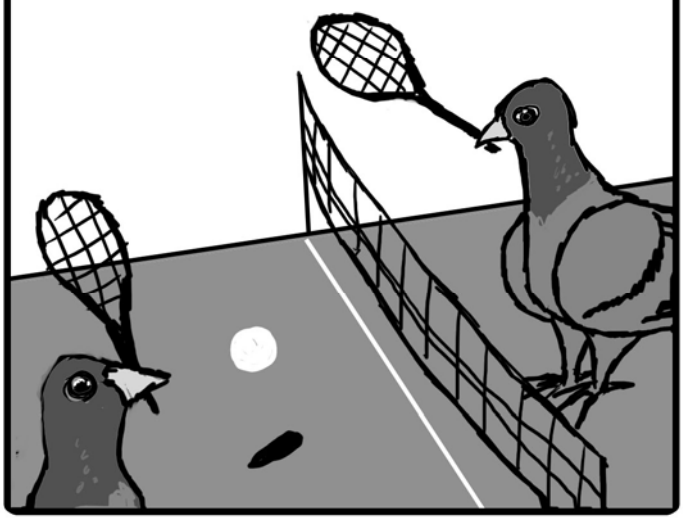
Conditioned Stimulus

Conditioned Response

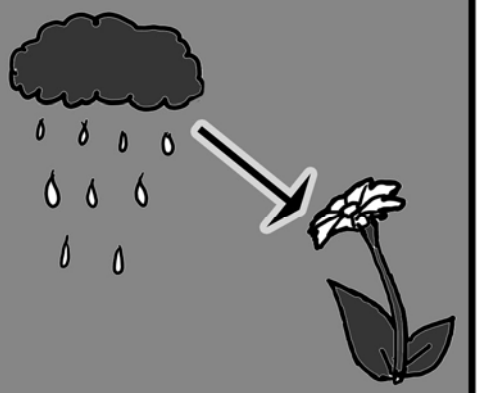
Eventually the sound of the bell alone is enough to provoke salivation.

Using similar techniques, it is possible to condition animals to do quite complicated tasks.

Noted psychologist B.F. Skinner even used the similar principle of operant conditioning to teach pigeons to play table tennis (not very well, but still).



The best part is that all of these actions are in the terms of stimulus/response, objective facts that can be measured scientifically.



If very complicated behavior can be generated using this conditioning, is it so strange to think that all behavior can be described in this way?

"Behaviorists" like B.F. Skinner think that all "intelligent" behavior can be described by physical behaviors, which are conditioned responses to one or more stimuli.

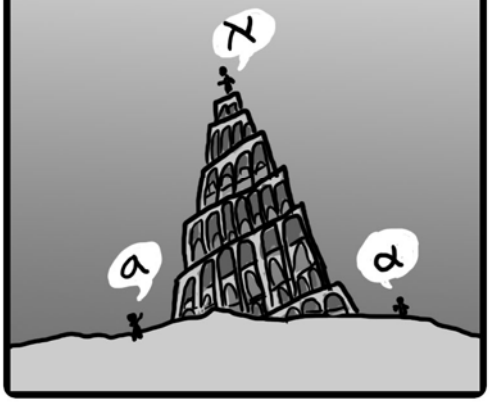
If successful, this theory of mind would give us an entirely objective way of determining facts about the mind, instead of subjective facts that are impervious to study. Behaviorism would allow us to completely ignore the internal aspects of cognition and focus instead on the external properties of the mental.

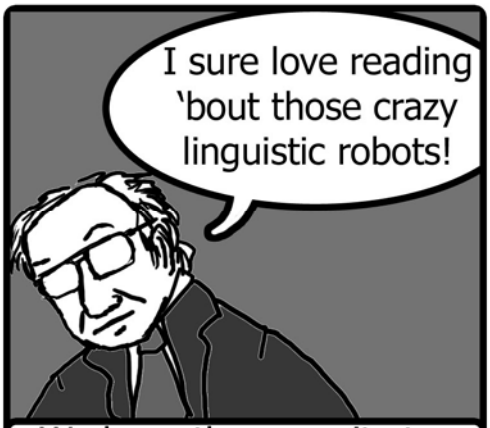
While behaviorism might explain a lot of behavior, I don't feel it holds water as a complete theory of mind.

For one, why do we appear to have subjective, internal experiences?

When we have similar input, how do we choose between the various outputs we can perform?

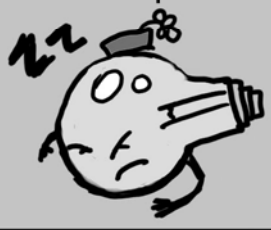
Noam Chomsky, famous linguist, thinks that our capacity for language is inherently fatal to any purely behaviorist view.



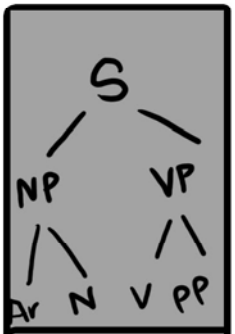


We have the capacity to utter sentences that have never before been spoken in any language, and in fact we do so frequently.

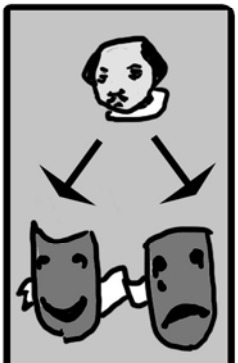
Even with a small vocabulary of words we can create millions of sentences, many of which are grammatically correct but have no semantic meaning, and so could not have originated as a result to a stimulus in the environment, eg. "Green ideas sleep furiously."



Infants receive a surprisingly small amount of linguistic input when growing up. This scarcity of input seems to suggest that some language skills are not acquired in the environment

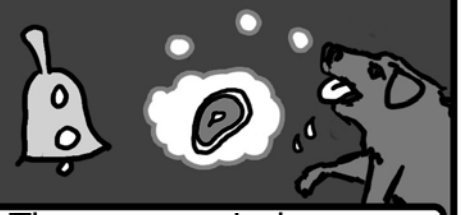


Language is learned by acquiring a few general rules, not with specific stimuli



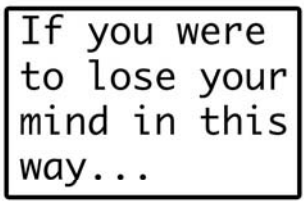
Stimulus does not completely determine linguistic output

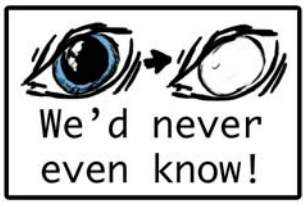
The implication is that behaviorism does not completely (or adequately) describe the actions of a conscious mind.

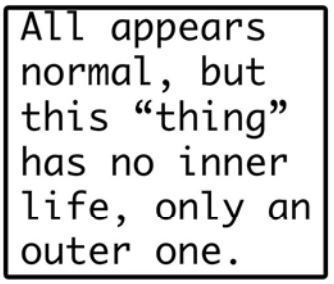


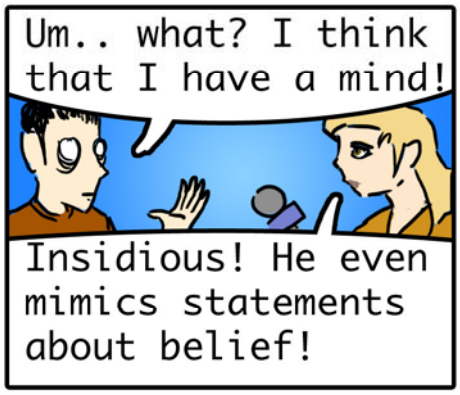
There seems to be some part of mental activity that is beyond mere response to a stimuli

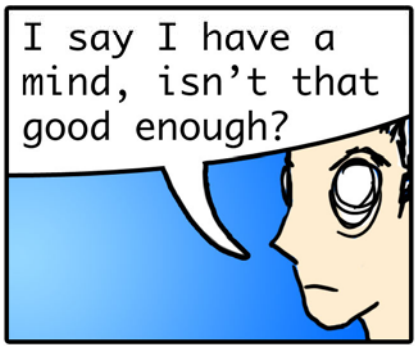
I don't mean to imply that conditioning does not occur in the brain; it almost certainly does occur. What I think is that only talking about the mind in terms of behavior is a mistake. The ideal would be a theory of mind that maintains the objectivity of behaviorism while at the same time allowing for the reality of subjective experience.







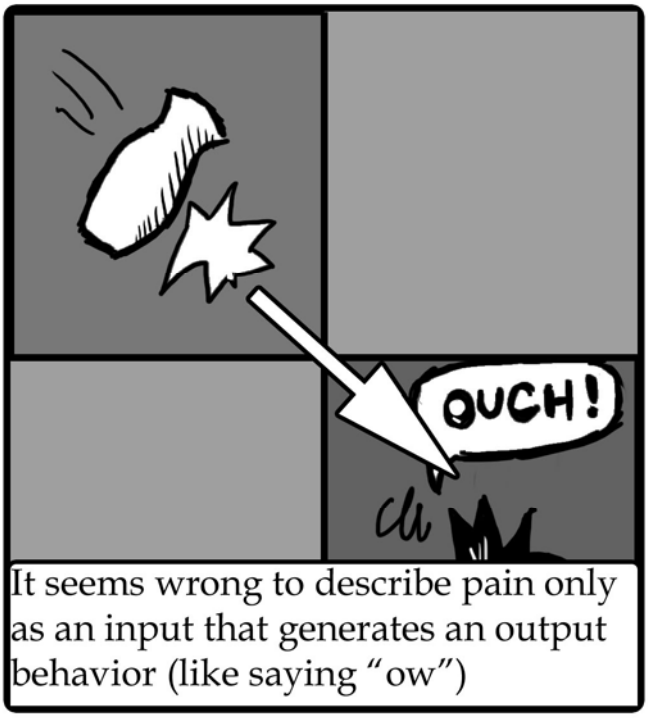




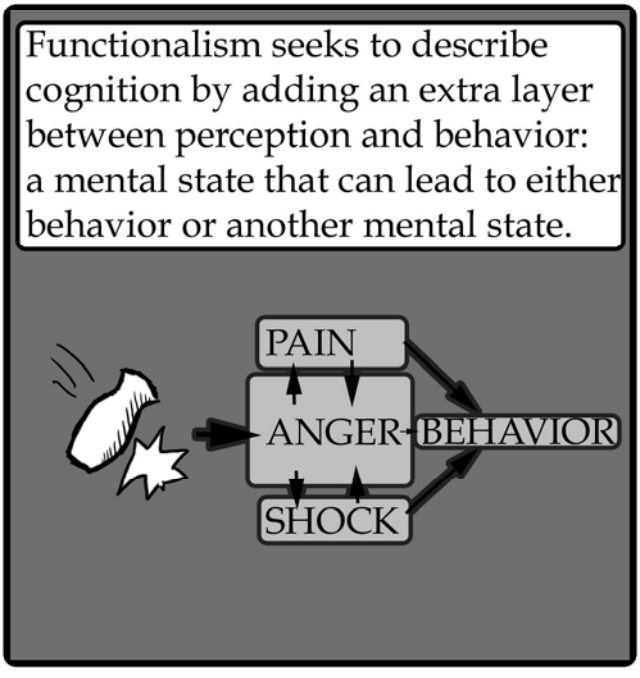
Are philosophical zombies possible in principle? Can there be a being who acts identically to a being with a mind, and yet is mindless? If so, how can we be sure that all AIs are not "zombies" of this type? Does it matter if they are?

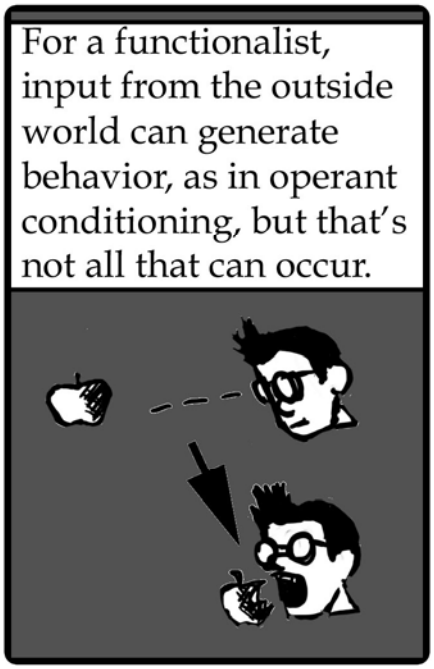

OUCH!

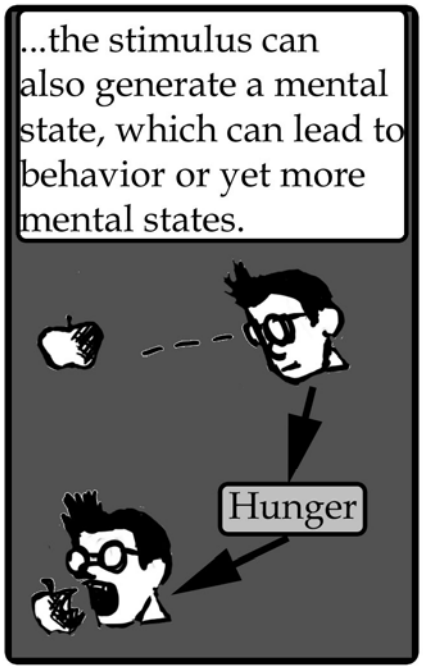

You could say that this person is in the mental state "pain"
PAIN


OUCH!
It seems wrong to describe pain only as an input that generates an output behavior (like saying "ow")


After all, there are many different behaviors that could conceivably result from the experience of pain
#@!
MY VASE!
STUPID VASE!


Functionalism seeks to describe cognition by adding an extra layer between perception and behavior: a mental state that can lead to either behavior or another mental state.
PAIN
ANGER
BEHAVIOR
SHOCK


For a functionalist, input from the outside world can generate behavior, as in operant conditioning, but that's not all that can occur.

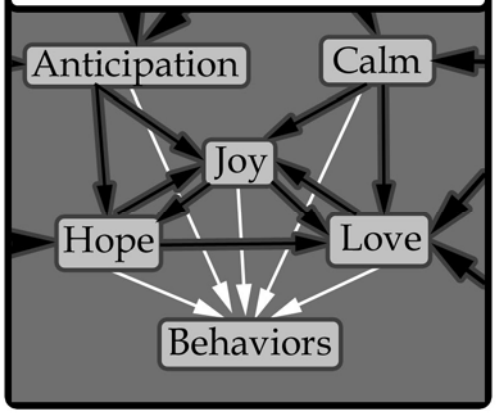

...the stimulus can also generate a mental state, which can lead to behavior or yet more mental states.
Hunger

Many AI modules in newer video games are roughly functionalist in their implementation in-game.

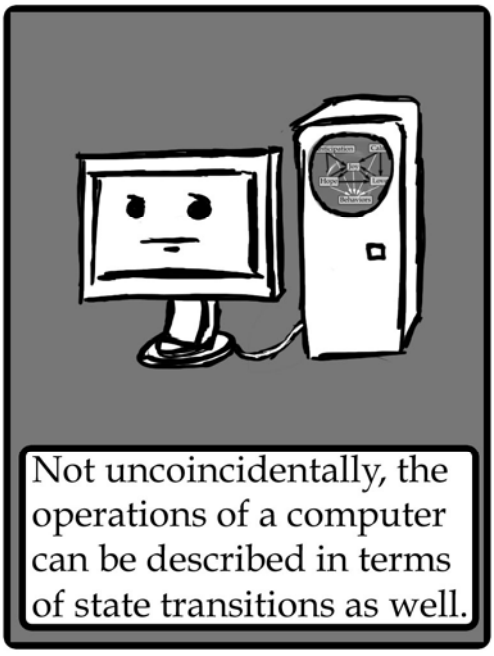
They will transition to the "alert" state upon sighting the player entering the area

They will then choose one of many actions to undertake, including flanking, charging, running away, etc. Just the act of seeing you won't completely determine the resultant action.

Functionalists want to describe the mental by what it does, not what it seems like. Its function, not its form.

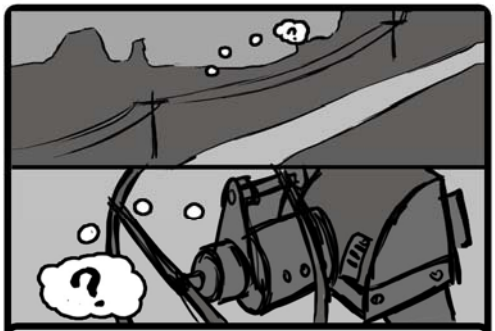


For them, a complete theory of mind will resemble a massive table of transition rules and outputs.



Not uncoincidentally, the operations of a computer can be described in terms of state transitions as well.

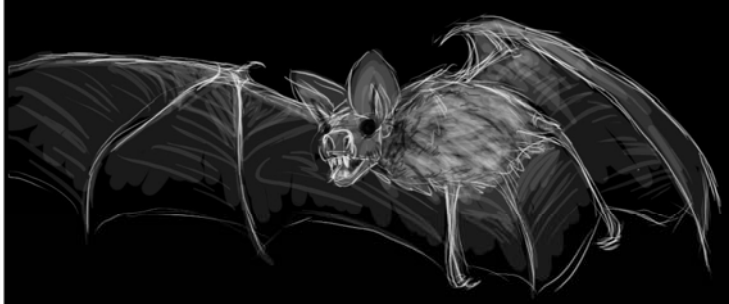
The creativity and the productivity of the mind might be a difference of degree, not a difference of kind, from the "mind" of a computer or other system.



We might be taking the computer analogy too far. Do we see the mind as computer-like only because computers are new? Minds used to be likened to telegraphs and clockwork.

Does functionalism avoid all of the problems of behaviorism? How might functionalism account for our sensations of identity, and subjectivity? For that matter, how does it account for sensations at all? What sort of discoveries in neuroscience might prove or disprove the functional theory of mental states? Is it a falsifiable theory?

From “What Is It Like to Be a Bat?” Thomas Nagel, and



“What Mary Didn’t Know” by Frank Jackson

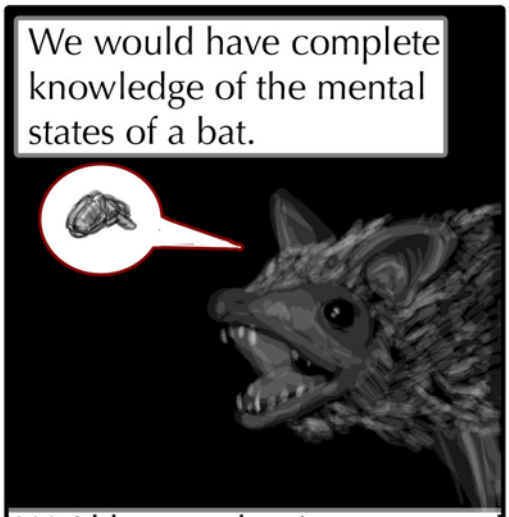
What is it like to be a bat?

A bat can use echolocation to “see” in the dark.

Instead of hands, they have leathery wings.

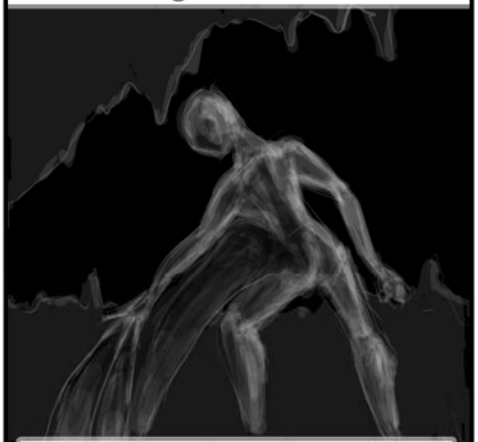
Suppose science were to completely describe a bat’s biology and brain.

We would have complete knowledge of the mental states of a bat.



We’d know what input caused what neural outputs exactly.

Would we really know what it is like to be a bat? To have wings and sonar?



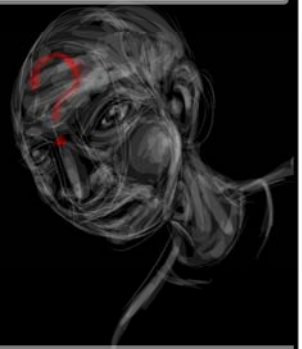
Would we know anything about a bat’s experiences?

Even if we could mimic a bat’s brain, this might not carry information about a bat’s experiences in a way we could use.

At best, we’d know what it is like for a human to try to be a bat, not what it is like for the bat itself.

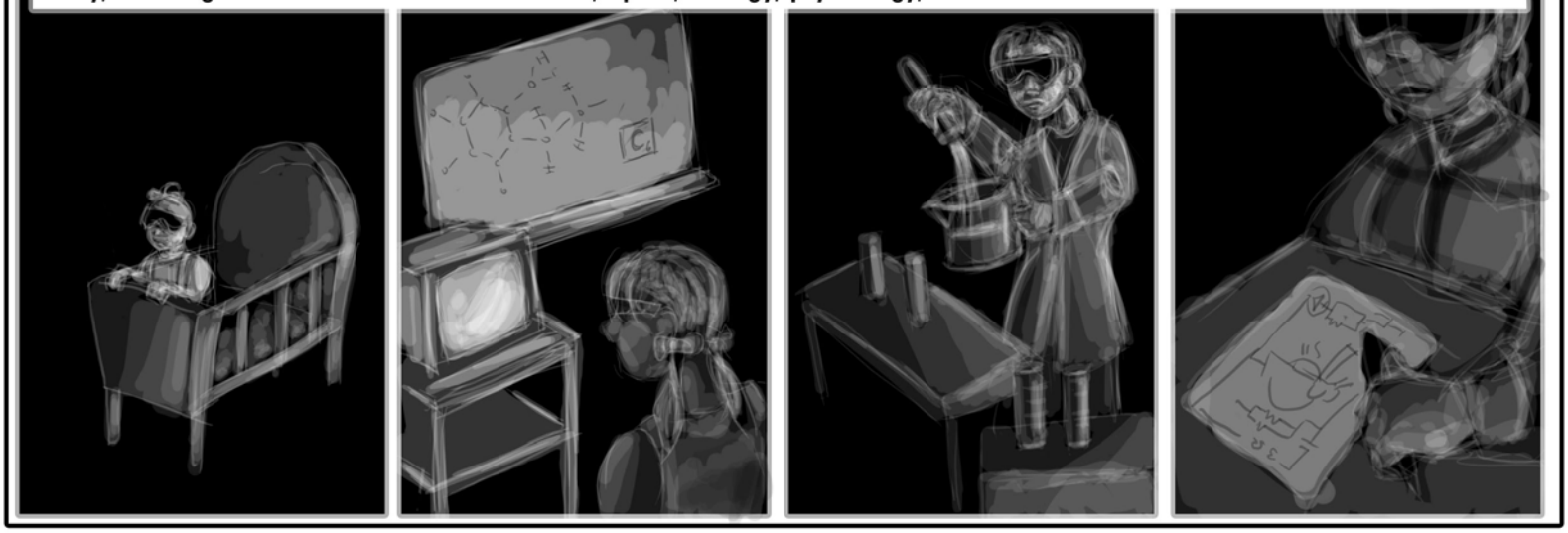
This would imply that there are, in principle, facts that science can never know about minds: what it is like to have an experience.

The problem is not exclusive to bats.

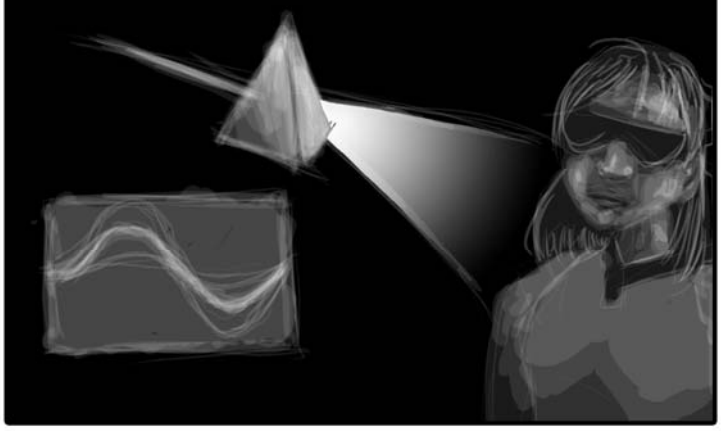


What is it like to be you, or to be me?

Imagine there was a woman raised from infancy with special goggles that prevented her from seeing color. Her life is spent in study, learning all facts about the human brain, optics, biology, psychology, and all of the other facts related to human vision.



Now this person knows everything there is to know about seeing color in principle, but her goggles prevent her from actually seeing the world in anything but black and white.



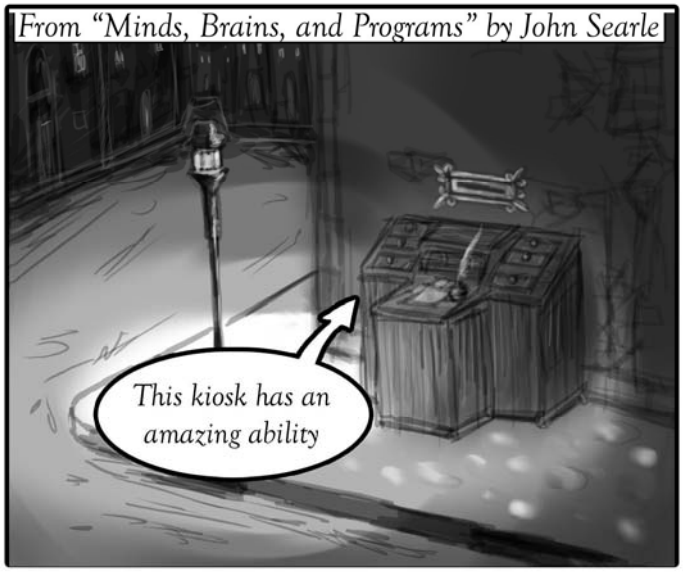
Now suppose her goggles are removed, and she is let out into a world of color.

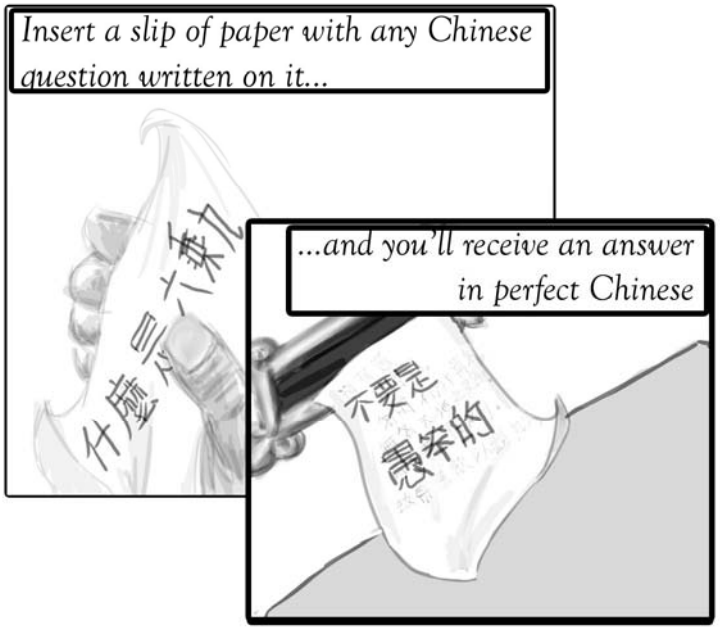
Will she learn something new the first time she sees, say, the color red? Is that a new fact?

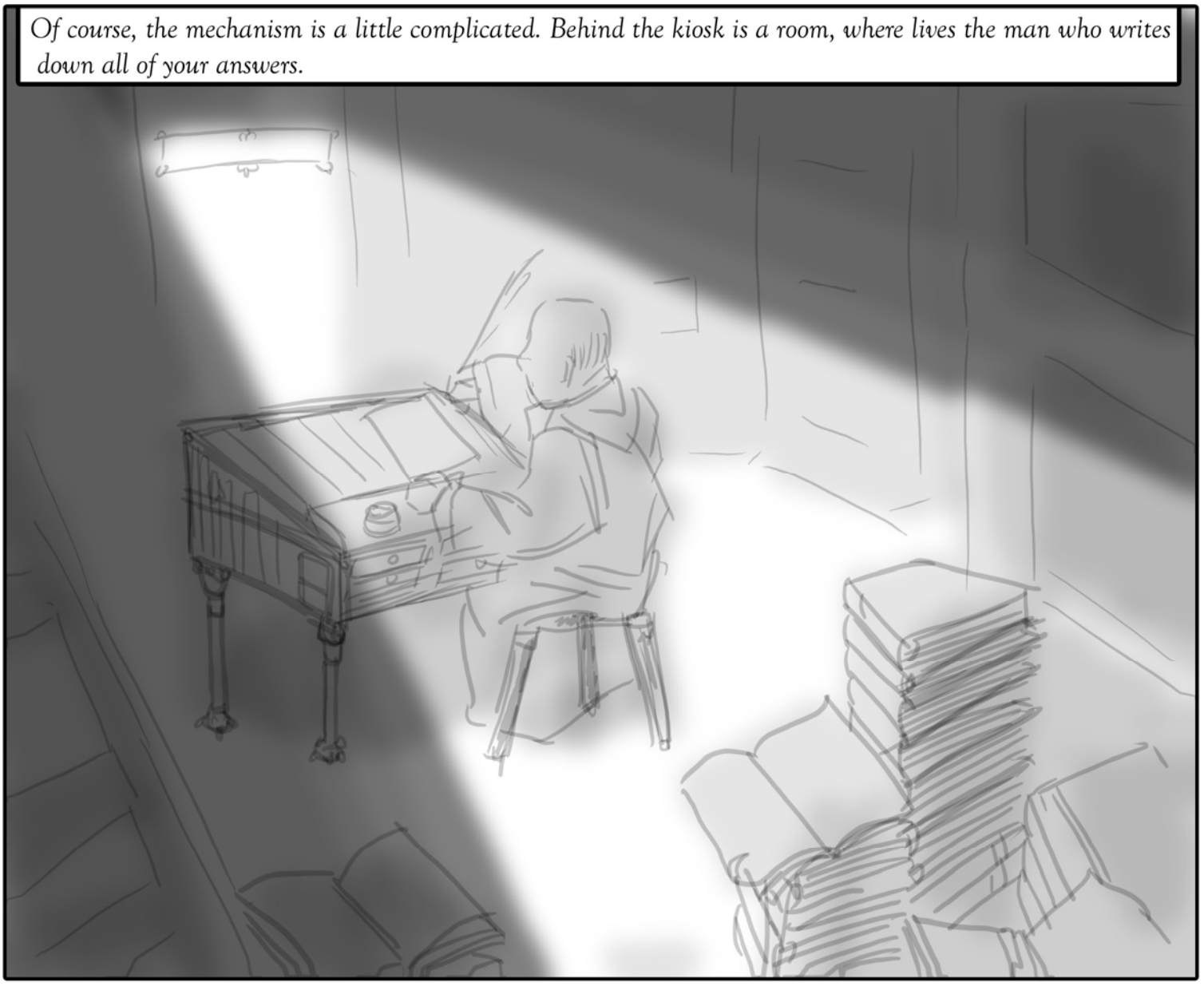
If learning all the physical facts about experience is not the same as actually experiencing, then it follows that there is something about experience that is not physical.

These subjective experiences, like seeing red or being a bat, are called "qualia." If qualia, in principle, cannot be analyzed by science, then we might have a problem making a mind.

Are qualia real? Is it possible that just knowing how the brain works is not enough to replicate a mind? Even if just knowing objective facts is not enough to find out about the subjective character of experience, can computers have qualia? What makes it clear that brains, even the brains of animals we don't normally consider "intelligent", can have qualia, but that even our "smartest" computers cannot? What properties must qualia have, in order to make them opaque to scientific enquiry?

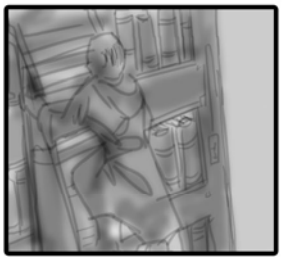

From "Minds, Brains, and Programs" by John Searle
This kiosk has an amazing ability

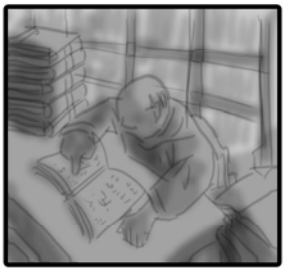

Insert a slip of paper with any Chinese question written on it...
...and you'll receive an answer in perfect Chinese


Of course, the mechanism is a little complicated. Behind the kiosk is a room, where lives the man who writes down all of your answers.

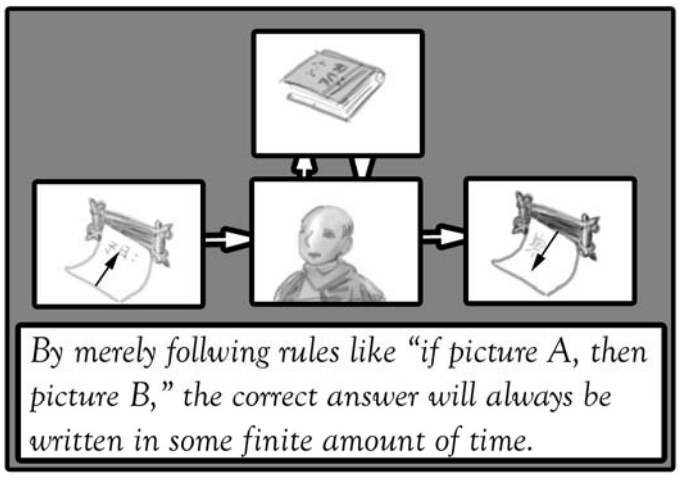
Unfortunately, he doesn't speak a word of Chinese. He has to consult a large number of books to figure out what to write, based on what symbols he sees.

Hmm... if I see this squiggle...

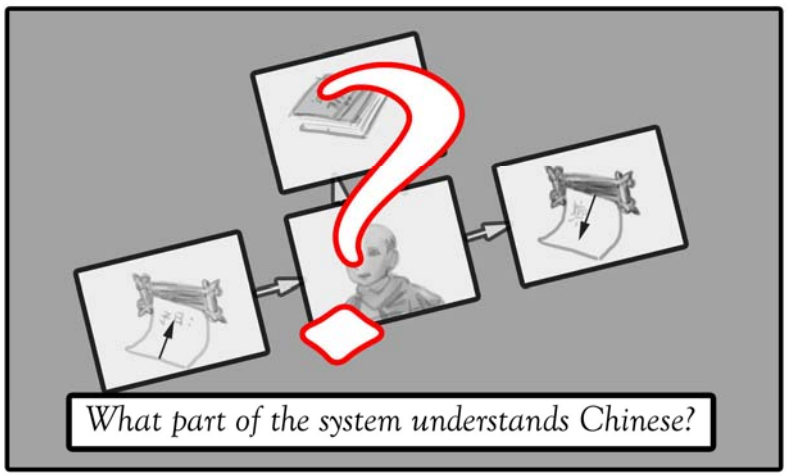




Then I should write this squiggle.

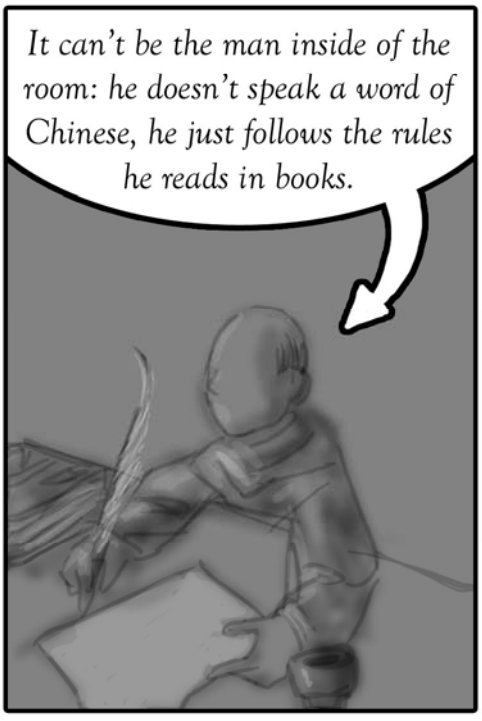


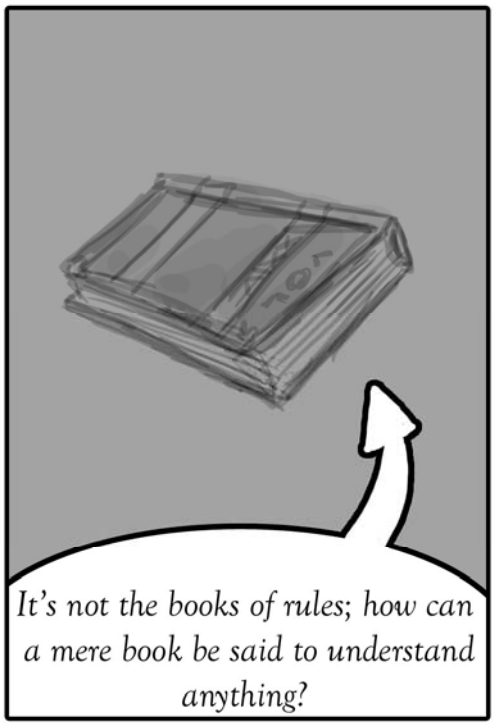
By merely follwing rules like "if picture A, then picture B," the correct answer will always be written in some finite amount of time.



What part of the system understands Chinese?

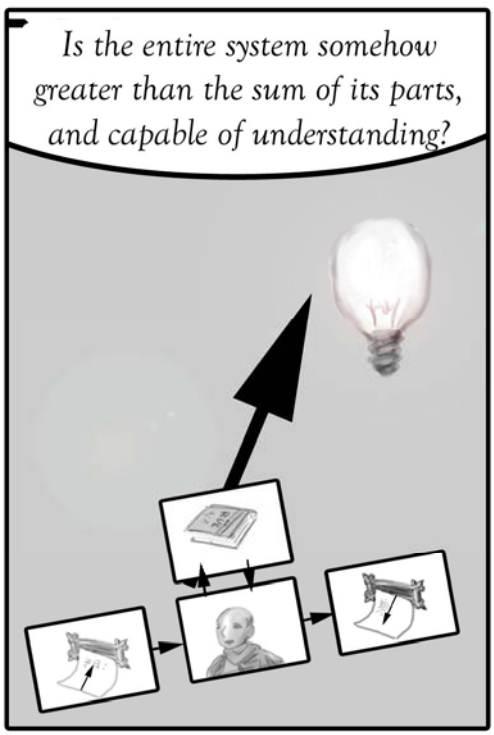
It can't be the man inside of the room: he doesn't speak a word of Chinese, he just follows the rules he reads in books.

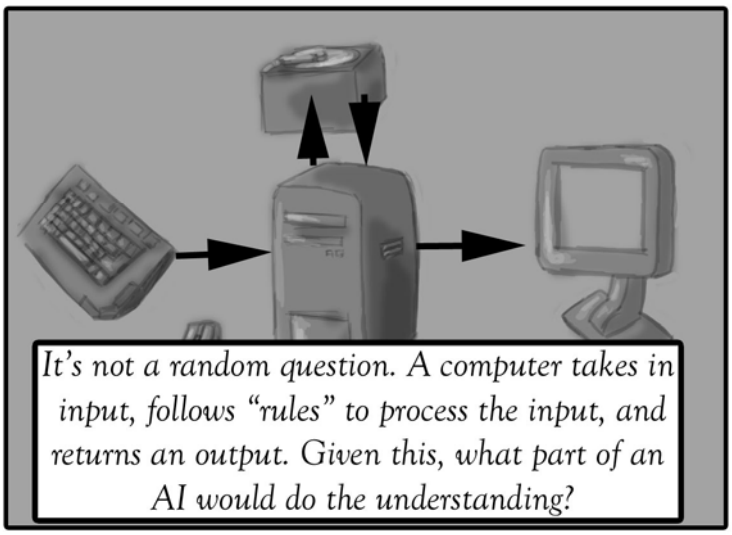




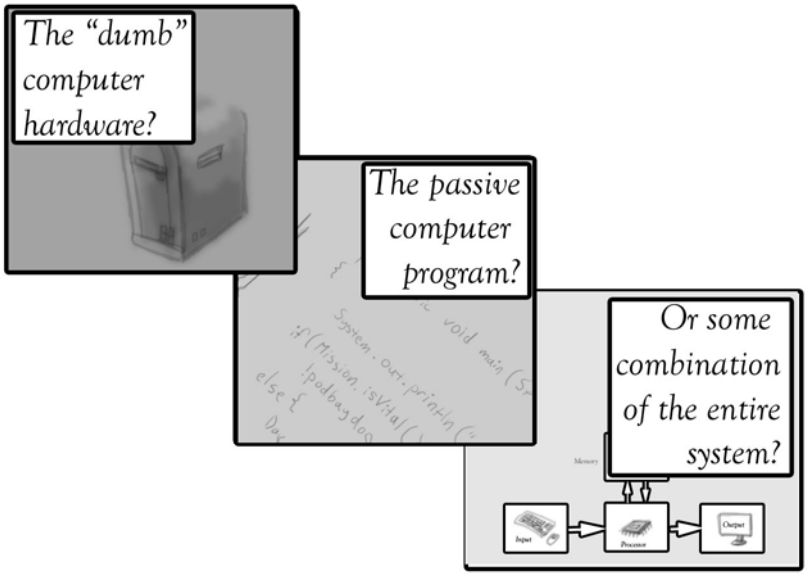
It's not the books of rules; how can a mere book be said to understand anything?

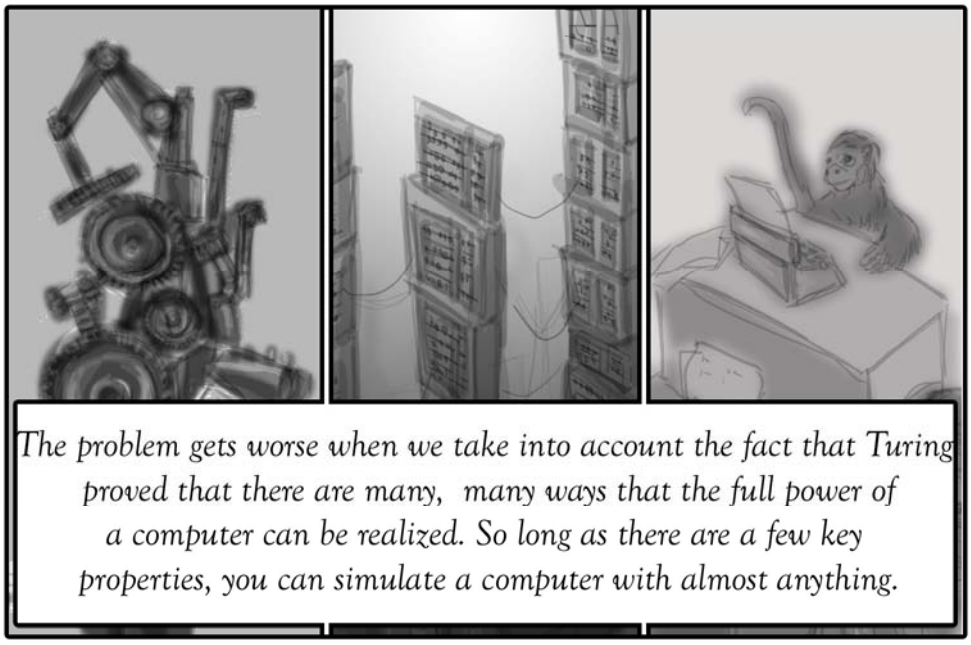
Is the entire system somehow greater than the sum of its parts, and capable of understanding?



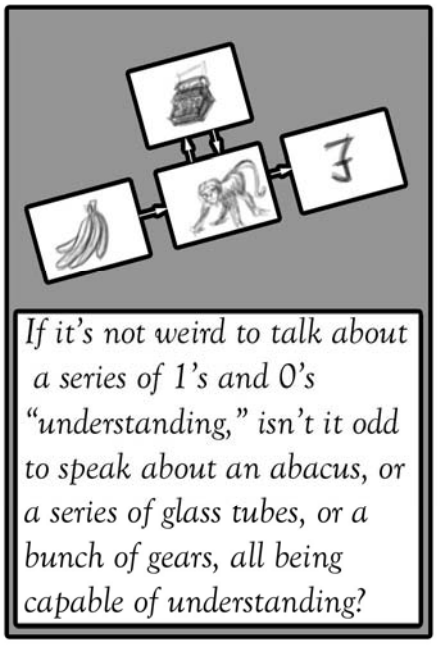


It's not a random question. A computer takes in input, follows "rules" to process the input, and returns an output. Given this, what part of an AI would do the understanding?

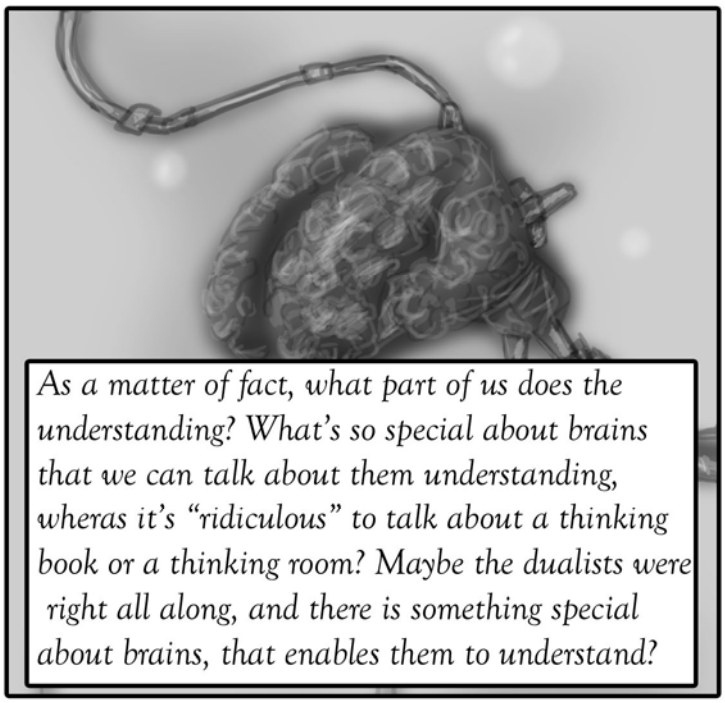




The problem gets worse when we take into account the fact that Turing proved that there are many, many ways that the full power of a computer can be realized. So long as there are a few key properties, you can simulate a computer with almost anything.



If it's not weird to talk about a series of 1's and 0's "understanding," isn't it odd to speak about an abacus, or a series of glass tubes, or a bunch of gears, all being capable of understanding?



As a matter of fact, what part of us does the understanding? What's so special about brains that we can talk about them understanding, wheras it's "ridiculous" to talk about a thinking book or a thinking room? Maybe the dualists were right all along, and there is something special about brains, that enables them to understand?

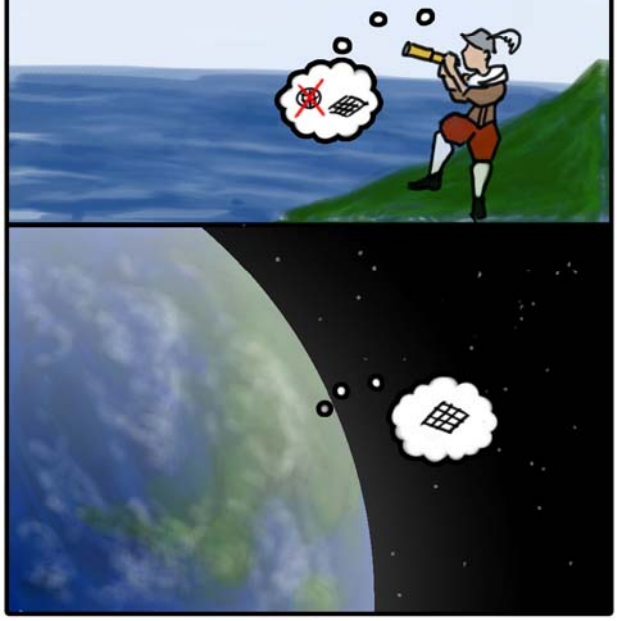
This is John Searle's "Chinese Room" argument against the idea that computers can ever, in principle, think like a human. I don't think it is very convincing, but it's very difficult to put my finger on exactly what is wrong with it.
Some questions to think about:

Is it possible to create a book of rules for Chinese?

What is special about brains that it isn't weird to talk about a brain understanding something?

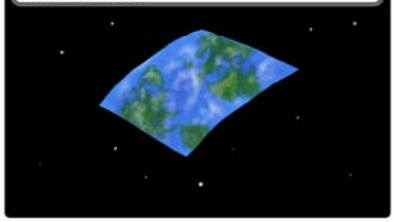
Could computers, or something like them, ever be designed to have this weird property?

It is a popular fact that not all is as it seems. We often have a wrong view of the world.



Often times our language does not catch up with the amount we learn about the world.

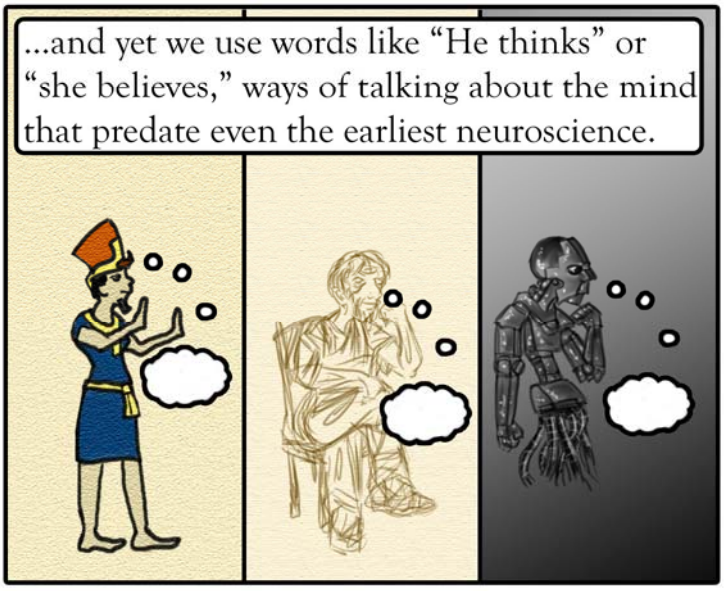
e.g. "The four corners of the earth"



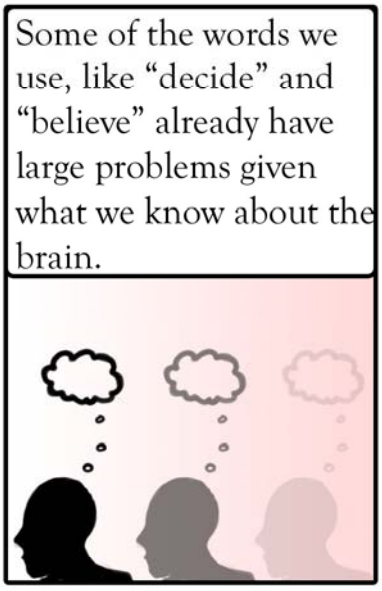
"Blind as a bat" (bats can see roughly as well as us)

"Firefly" (does not use fire to create light)

...and yet we use words like "He thinks" or "she believes," ways of talking about the mind that predate even the earliest neuroscience.



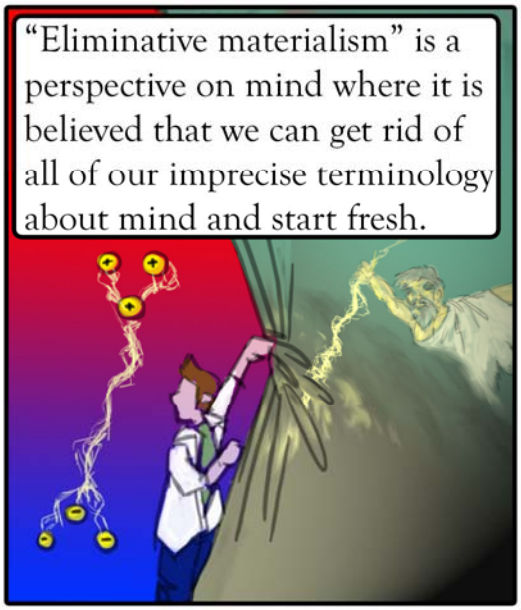
Some of the words we use, like "decide" and "believe" already have large problems given what we know about the brain.

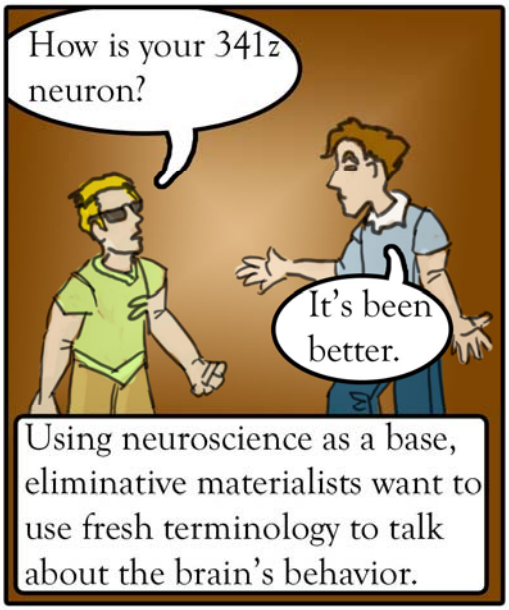


What if talking about "mind" is the same as talking about Zeus as a cause of lightning?

A relic of a linguistically archaic past?

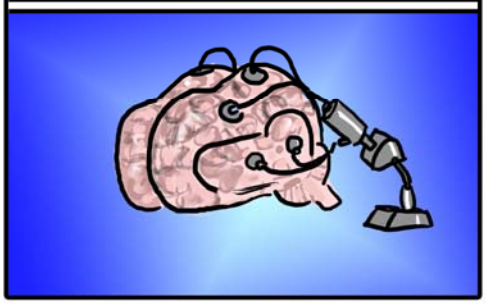
"Eliminative materialism" is a perspective on mind where it is believed that we can get rid of all of our imprecise terminology about mind and start fresh.





Using neuroscience as a base, eliminative materialists want to use fresh terminology to talk about the brain's behavior.

Once we remove outdated language, a lot of the problems with our theories of mind (intentionality, subjectivity, the role of qualia) disappear. "Mind" is no longer an issue.



This is an incorrect way of talking about how a volcano works:

Run! Vulcan, god of the earth, is angry!

Volcanoes cannot be angry. We do not need to take anger into account at all.

This is also incorrect:

Germany was worried about the progress of the Russian front.

Entire nations do not have feelings or fears.

What evidence do we have that this exchange is not also incorrect?

Bob thinks that this apple is red.

"Folk psychology" is the name given to our everyday, common-sense theory of mind.

This machine hates me!

ERROR!

That lion wants to eat me!

It has been more or less the same for centuries of human thought.

Like any theory, folk psychology makes predictions, some good, some bad.

That person is hungry!

That bear just wants to play!

Unfortunately, "good enough" is not good enough for a theory of mind.

?

Just as our common sense theory of physics is often wrong, our theory of mind might also err.

Already neuroscience and psychology are showing a large difference between what we assume and what actually turns out to occur in the brain.

A lot of the problems with our current view of mind, such as the prevalence of dualism, are a result of applying folk psychology terms where they just do not belong.

Mind

?

Body

Is eliminative materialism a viable alternative to "folk psychology?"

Why do you think folk psychology is so prevalent in everyday life, if it is inaccurate?

How does this view of the mind differ from "behaviorism," as discussed previously?

How might some of the arguments against AI be refuted, if we accept this view of mind?

The preceding comics should by no means be taken as the final word on the subject of the philosophy of Artificial Intelligence; if anything, they are meant to whet the appetite for more serious follow-up study. My goal was to distil my thinking about the subject to the simplest elements so that I could give the briefest of overviews of the breadth and depth of that strange intersection of computer science, philosophy, neurology, and psychology that is AI.

The anthologies *The Nature of Mind* (ed. Rosenthal, Oxford, 1991) and *The Mind's I* (ed. Hofstadter and Dennett, New York, 2000) were instrumental in the completion of this work, but I also consulted *The Conscious Mind* (Chalmers, New York, 1996), *Brainchildren: Essays on Designing Mind* (Dennett, Boston, 1998), and *Artificial Minds* (Franklin, Boston, 1997). If you are interested in more detail in any of the subsections of my work (and I hope you are), I have provided the following list of seminal papers and important books organized by topic.

Mind/Body Dualism:

Descartes, *Meditations on First Philosophy* (Meditations II and VI most pertinantly)
Ryle, "Descartes' Myth"
Smullyan, "An Unfortunate Dualist"

The Mechanical Turk/Deep Blue:

Hsu, *Behind Deep Blue: Building the Computer that Defeated the World Chess Champion*

The Turing Test:

Dennett, "Can Machines Think?"
Turing, "Computing Machinery and Intelligence"

Behaviorism and critiques:

Chomsky, "A Review of B.F. Skinner's *Verbal Behavior*"
Pinker, *The Language Instinct*

Functionalism and critiques:

Block, "Troubles with Functionalism"
Jackson, "What Mary Didn't Know"
Nagel, "What is it Like to Be a Bat?"
Putnam, "The Nature of Mental States"

The Chinese Room:

Searle, "Minds, Brains, and Programs"

Eliminative Materialism:

Churchland, "Eliminative Materialism and the Propositional Attitude"
Dennett, "The Unimagined Preposterousness of Zombies: Commentary on Moody, Flanagan, and Polger"
Rorty, "Persons Without Minds"

Of course this is not an exhaustive list, but if you can work past the occasionally esoteric terminology of philosophy papers this ought to provide a good background for further investigation. Noticeably absent from my discussion are the fascinating fields of embodied cognition and phenomenology in general. I won't lie; other than brief dabbling

with Heidegger my background is almost exclusively analytic rather than continental. That being said, the work of Rodney Brooks at MIT offers an interesting alternative to what is typically referred to as the "information processing" modality of AI, which is the view that the mind's primary job is to gather information from the environment and process it into useful behavior. Brooks thinks that internal representation of facts in the brain is a much smaller part of mind than is typically thought. Without using any real sort of internal representation, Brooks has been able to create robots that engage in pretty sophisticated behaviors. For a slightly less conventional view, biologist and philosopher Francisco Varela's quasi-spiritual works on the self and cognition will offer more than enough to ponder for those of you who are suspicious of the concept of the self, and science's interaction with the world in general.

Now that the bookkeeping is out of the way, I suppose I should answer the question that I posed at the outset: can machines think? Will machine thought ever resemble flexible, creative, intelligent human thought? I think already the question is moving from the realm of the philosophers and inexorably onto the laps of the engineers. Take a complex system like the Google search algorithm. It performs an enormously complex task (sorting through millions of websites to return only relevant ones), and does so with the aid of a constantly evolving semantic net that knows that when you type "cate reciped" you most likely meant "cake recipes." This is no easy task, and requires a lot of abilities we would otherwise consider exclusive to the purview of intelligent beings. Of course, Google is far from intelligent; but it is one of the closest things to an intelligent machine entity currently in use. The Amazon.com and Netflix recommendation systems are also surprisingly supple and intelligent systems, capable of giving meaningful results given widely disparate data despite the fact that they are little more than fiendishly clever simple algorithms for drawing connections between disparate data, aided by access to an enormous database of user feedback.

You will notice that I have again ducked the question. I have said that there are surprisingly intelligent systems, but I have said nothing about humanly intelligent systems. I do not think such systems are impossible, and in fact I think that there will be a lot of systems that will meet or exceed human performance in a number of key areas (or already have). The construction of flexible intelligent systems seems inevitable given both the sheer number of useful applications that would benefit from AI, and the amount of research and interest that is directed towards solving the problem. Granted, one of the many philosophical critiques of AI might cause researchers to have to drastically rethink the problem (as they did after the early overoptimistic failures of AI, and the later "AI winters" of the late 70's and late 80's), but I do not think the project is doomed in principle.

The biggest philosophical obstacle to artificial intelligence (as opposed to the very real obstacles of cost, computing power, and other engineering concerns) is in my estimation the creation, from the ground up, of the capacity for an artificial system to sensibly create heuristics that deal with complex environments. The AI systems we have today are usually brittle or of incredibly limited domains. And while this might be

because we just haven't thrown enough computing power at the problem, I think it is more likely that there is something brittle and limited about the way we currently construct and think about AI. While we still think of AI as the labeling of stimuli from the environment and the manipulation of logical propositions connected with those stimuli, we inherently limit the domain of our AI applications. I think a better approach would be one connected with the inherent modularity and multiplicity of mind (the so-called "Society of Mind" championed in the some of the works of AI researcher Marvin Minsky). This approach builds up a system that is capable of dealing with many different situations through the creation and manipulation of small agents (or demons, or capacities, or talents, or whichever term you prefer) that handle individual tasks. Through the interplay of many of these agents, a mind is created.

What then of the criticisms given by Nagel and Searle in the comics above? How can semantic meaning arise from pure syntactic interactions? How can the objective interactions of the physical create the subjective language of raw feels? They are related questions, and ones that I will deal with perhaps a little too flippantly. Firstly, how does semantics arise in the human brain? And how does the objective physical brain create subjective experience? The fact that we cannot answer these questions for our own case means that the assumption that a computer in principle cannot fulfill these same obligations is an argument *ad ignorantiam*. I see no reason in principle why a computer cannot have subjective experiences (whatever those may be) in the same way a human might, and until we have a better idea on what is meant by consciousness or subjectivity most arguments for the impossibility of AI reek of biological chauvinism, the assumption that because the biological brain is the only intelligent thing we know, it is the only possible intelligent thing. Compare the similar initial incredulity that greeted the notion that the other stars in the sky might be suns with their own solar systems and possible forms of life.

The current AI project is a little scattershot. Some (the neurologists and cognitive psychologists, mostly) are investigating the neurological bases of cognition in the human brain, and then attempting to implement those structures on a computer. Others (computer scientists *et al*.) are attempting to start by designing computer systems with certain logical architectures, and scaling them up until they can compete with human intelligence. Finally, the roboticists (both the traditional ones at the ones inspired by Brook's representation-less ideas) are hoping that using computational techniques in a complex real-world setting will generate results. I am of the opinion that these approaches, while I doubt they will all dovetail and meet in the middle somehow (I think it as practical and likely as a chemistry completely subsuming biology; many of these approaches operate on drastically different levels of description and so will not likely play nice together) I think that all of these approaches add knowledge to our understanding of intelligence. A successful artificially intelligent system would almost certainly bring to the table knowledge gained from many (if not all) of these approaches.

What if we do succeed, what then? You will notice that I have studiously avoided the ethical and political ramifications of intelligent machines. What rights ought an AI to have, given the fact that the machine would likely have radically different properties when compared to a human being? Is turning off an AI the same as murder? If we can make an AI, ought we to make one? These questions have no easy answers. Much like with the creation of the atomic bomb, the creation of a self-aware intelligent AI will have to spark large changes in the interaction between science and society.

An even more fascinating concept is the idea of the "technological singularity" championed by futurists such as Raymond Kurzweil and Vernor Vinge. They postulate that the scientific progress of the human race is increasing at an exponential rate; while thousands of years passed from the invention of the written word to the first printing press, the first powered flight was followed a mere half-century later by our first moon landing. Technology allows for faster invention, which allows for more technology, etc. Once we have an artificial intelligence capable of increasing its own intelligence in the same exponential fashion, or we can increase our own intelligence using biological innovations, then this progress will increase at an even faster rate. At some point the rate of innovation will be asymptotic as human innovation becomes increasingly guided by super-human intellects, whether these intellects are AIs, "augmented" humans, or some combination of the two. At some point speaking of "human" society will be an anachronism. An interesting idea, to be sure, but tinted by the grandiose notions and assumptions that often mar futurism as a predictive tool.

I think a self-aware AI would in large resemble a human being. After all, we have only ourselves as examples of a self-aware intelligent being from which to study, so it makes sense that the broad outline of an AI would have not a few similarities with ourselves. The capacities of a computer to be copied and the modularity of computational systems might add some interesting wrinkles, however. Imagine a personality that could be copied identically and implemented on a number of systems, or a personality that could be made twice as intelligent or twice as creative with a simple upgrade or two. And of course form will follow function; there is no reason that an AI designed to, say, prove number theory results would have anything more than a passing resemblance to an AI designed to defuse bombs in the real world, let alone the consciousness of a human being.

Of course we are engaging in the popular philosophical pastime of counting our chickens before they hatch. I think a number of the objections raised in the preceding pages will seem quaint and outdated from the vantage point of 50 years hence, much as the state-of-the-art computer of the last decade is a laughable antique today. We still have quite a bit to do before we can consider the AI problem solved, and it might turn that there are problems that are beyond our ability to solve. As unhelpful as it is to say this, time will tell which branches of AI research are fruitful and which are not.