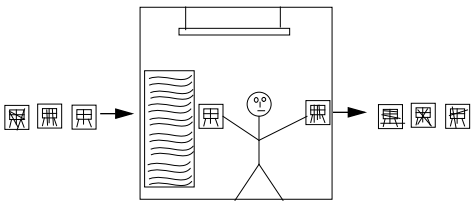
Philosophical Arguments Against "Strong" AI

Strong vs. Weak AI

- "Weak" AI just claims the digital computer is a useful tool for studying intelligence and developing useful technology. A running AI program is at most a simulation of a cognitive process but is not itself a cognitive process. Analogously, a meteorlogical computer simulation of a hurricane is not a hurricane.
- "Strong" AI claims that a digital computer can in principle be programmed to actually BE a mind, to be intelligent, to understand, perceive, have beliefs, and exhibit other cognitive states normally ascribed to human beings.

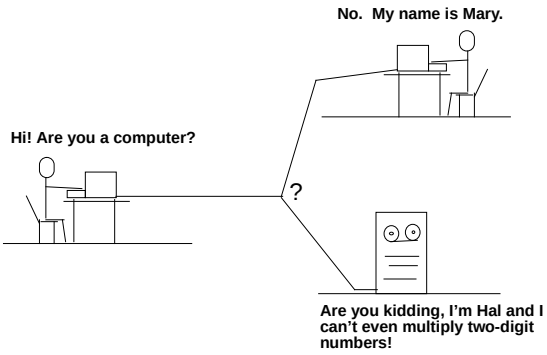
Searle's Chinese Room

- Imagine an English speaking human-being who knows no Chinese is put in a room and asked to simulate the execution of a computer program operating on Chinese characters which he/she does not understand.
- Imagine the program the person is executing is an AI program which is receiving natural language stories and questions in Chinese and responds appropriately with written Chinese sentences.
- The claim is that even if reasonable natural language responses are being generated that are indistinguishable from ones a native Chinese speaker would generate, there is no "understanding" since only meaningless symbols are being manipulated.



The Turing Test

- If the response of a computer to an unrestricted textual natural-language conversation cannot be distinguished from that of a human being then it can be said to be intelligent.



- Searle doesn't directly question whether a computer could pass the Turing test. Rather, he claims that even if it did, it would not exhibit "understanding."

Responses to Searle

- The Systems Reply: The person doesn't understand Chinese but the whole system of the program, room, plus person understands Chinese.
- The Robot Reply: If you give the computer a robotic body and sensors through which it interacts with the world in the same way as a person then it would understand.
- The Brain Simulator Reply: If the program was actually simulating the firing of all the neurons in a human brain then it would understand.
- The Combination Reply: If the program was simulating a human brain AND had a robotic body and sensors then it would understand.
- The Other Minds Reply: If there is no understanding in the room then how do ever know that anyone ever understands anything.
- The Many Mansions Reply: Maybe a digital computer won't work but you can build an artificial intelligence using different devices more like neurons that will understand.

Systems Reply

- Searle's response is to let the person internalize the entire system memorizing the program and all intermediate results.
- Assuming this were somehow actually possible, then the person would arguably contain two minds, one which understood English and one that understood Chinese. The fact that the English mind doesn't "understand" the Chinese mind seems to obscure the understanding of the Chinese mind itself.
- According to Searle, the Chinese room lacks the "causal powers of the brain" and therefore cannot understand. Why doesn't the room or silicon chips have such "causal powers." How would we know whether the "brains" of an intelligent alien species have such "causal powers." Searle claims this is an "empirical question" but gives no experimental procedure for determining it.

Robot Reply

- Searle's response is that even if the symbols entering the room come from television cameras and other sensors and the outputs control motors, the basic lack of "understanding" doesn't change.
- Some AI researchers still believe it is important to have symbols "grounded" in actual experience with the physical world in order for them to have "meaning."
- In any case, it would probably be extremely difficult to write a program with all the knowledge of the physical world necessary to past the Turing test without having learned it from actual interaction with the world.

Brain Simulator Reply

- Searle's repsonse is that even a formal simulation of all the properties of the brain wouldn't have the "causal properties" of the brain that allow for intentionality and "understanding."
- Therefore, if each of your neurons were incrementally replaced with silicon circuits that replicated their I/O behavior, your observable behavior would not change but, according to Searle, at some point you would stop actually "understanding" anything.

Other Minds Reply

- Searle's response is that of course anyone can be fooled into attributing "understanding" when there actually is none, but that does not change the fact that no real understanding is taking place..
- However, there then seems to be no empirical test that could actually decide whether "understanding" is taking place or not and solipsism is the only truly reliable recourse.

Many Mansions Reply

- Searle's response is that strong AI is committed to the use of digital computers and that he has no argument against intelligence based on potential alternative physical systems that possess "causal processes."
- Searle is not a dualist in the traditional sense and grants that the mind is based on physical processes, just that a computer program does not possess the proper physical processes.
- He claims that, if anything, proponents of strong AI believe in a kind of dualism since they believe the critical aspect of mind is in non-physical software rather than in physical hardware.

The Emperor's New Mind (and other fables)

- Roger Penrose, a distinguished Oxford mathematician and physicist, has recently published a couple of books critical of strong AI (*The Emporer's New Mind, Shadows of the Mind*)
- His basic argument is that Gödel's Incompleteness Theorem provides strong evidence against strong AI.
- Unlike Searle, he is unwilling to grant the possibility that a computer could actually ever pass the Turing test since he believes this would require abilities that are uncomputable.
- However he is also not a dualist and believes that the behavior of the brain is actually physically determined.
- Since current theory in physics is either computable or non-deterministic (truly random) he believes that a new physics needs to be developed that unifies quantum mechanics and general relativity (quantum gravity theory) that is deterministic but noncomputable.

Gödel's Theorem

- Godel's theorem states that any consistent axiomization of arithmetic (e.g. Peano's axioms) is necessarily incomplete in that there will always be true statements of arithmetic that cannot be proven from these axioms.
- This is proved by constructing,for any given set of axioms, A, a "Gödel sentence" which is unprovable from the axioms A and effectively states "This statement cannot be proven from the axioms A."
- Corollary: Any computer program that judges the truth of mathematical statements is equivalent to some formal set of axioms and is therefore necessarily incomplete.
- Corollary: No formal system (program) powerful enough to capture arithmetic is powerful enough to prove its own consistency.

Claimed Implications of Gödel's Theorem

- People seem to simply be able to "see" the truth of some mathematical statements (e.g. Gödel sentences) through "intuition" and "insight" independent of proof in any formal system and therefore an algorithm must not underly human mathematical reasoning.

- Penrose makes the more precise claim:

  Human mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truth.

  If they were, it would constitute an algorithm which can assert it's own soundness which Gödel's theorem proves is impossible.

- Penrose's mathematical philosophy is Platonism, which claims mathematical statements are true or false independent of any particular formal system (set of axioms) and that humans can "see" the truth of some statements through direct contact with the "Platonic world of mathematical ideas."

Two Interpretations

- Gödel himself was also a Platonist who believed that his theorem implied that human mathematical insight was not based on any algorithm.

- Turing on the other hand simply believed it implied that human mathematical reasoning is potentially unsound.

  In other words then, if a machine is expected to be infallible, it cannot also be intelligent. There are several theorems which say almost exactly that. But these theorems say nothing about how much intelligence may be displayed if a machine makes no pretence at infallibility.

Alternatives Consistent with Strong AI

- An unsound (but presumably usually correct) algorithm underlies human mathematical intuition and we may or may not be able to eventually know it.

- A sound algorithm underlies human mathematical intuition but it is horrendously complicated and we can never know it completely (and it of course could vary some from individual to individual).

Penrose's Arguments Against These Alternatives

- Penrose believes human mathematical reasoning is knowably sound, saying to assume otherwise:

  ...seems to be totally at variance with what mathematicians seem actually *to be doing when they express their arguments in terms that can (at least in principle) be broken down into assertions that are 'obvious' and agreed by all. I would regard it as far-fetched in the extreme to believe that it is* really *the horrendous unkowable X, rather than these simple and obvious ingredients that lies lurking behind all our mathematical understanding.*

- But is it far-fetched to assume that what appears "obvious" to a human being might actually rely on a very complex neurobiological process that could even make a mistake? Aren't the incorrect conclusions prompted by various visual illusions "obvious" to people. Clearly this alternative is exactly what most proponents of strong AI actually believe.