The Philosophy and Ethics of Artificial Intelligence

Intelligence

Dr. Sanjay Modgil

Office hours: 10am – 12pm

(Email me for an appointment)

Overview of Course

- Introduction
- Intelligence
- Consciousness
- Reasoning and Communication
- Ethics and Morality
- Algorithms
- For good or bad ? AI in Medicine, AI in War
- Superintelligence and the Value Loading Problem
- AI and Human Society
- Review/Revision

Intelligence

Overview of Today's Lecture

- What is Intelligence ?
- What is *Artificial* Intelligence ?
- *Narrow*, *General* and *Super* Intelligence
- How can we tell that a machine is intelligent ?
- Arguments against the possibility of Artificial Intelligence

Sources and Reading

- *Life 3.0*. : *Being Human in the Age of Artificial Intelligence* Max Tegmark
- *Superintelligence*. Nick Bostrom
- *Artificial Intelligence: A Modern Approach*. Stuart J. Russell and Peter Norvig
- Stanford Encyclopedia of Philosophy (https://plato.stanford.edu/index.html)

What is Intelligence ?

Dictionary definitions e.g.

- ability to acquire and apply knowledge and skills,
- the faculty of understanding; intellect. a mental manifestation of this faculty, a capacity to understand.

"*Intelligence is the ability to accomplish complex goals*"

(Max Tegmark, Life 3.0)

Subsumes (implies) above and other definitions e.g. capacity for logic, "understanding", self-awareness, learning,

What is Intelligence ?

"*Intelligence is the ability to accomplish complex goals*"

- According to this definition is it possible to measure intelligence; to say that one being is more intelligent than another ?
- Which is more intelligent *Go* playing or *Chess* playing computer ?
- X is more intelligent than Y if all goals that can be accomplished by Y can be accomplished as least as well by X, and X can accomplish at least one goal better than Y.

  What kind of intelligence ordering would one obtain ?

What is Intelligence ?

- *Anthropomorphic* (human centred) perspective rates certain tasks much more intelligent than others .e.g. , 5005.658 x 13.31566 versus recognizing a friend in a photo
- Fact that low level sensorimotor tasks seem easy to humans despite requiring huge computational resources is known as **Moravec's paradox**
- Explained by fact that our brains make such tasks **feel** easy by dedicating massive amounts of dedicated 'hardware' to them (quarter of our brains)
- Later we will briefly mention a more principled (mathematical) framework for characterising intelligence

What is Artificial Intelligence ?

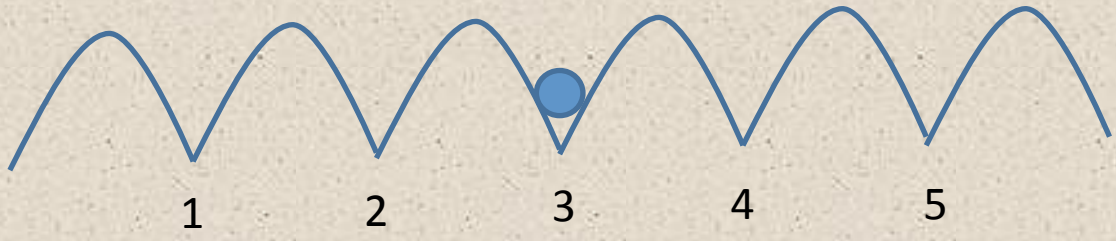
- From an AI perspective intelligence is ultimately about *information* and *computation* and not dependent on the underlying 'hardware' being carbon based (like human brains)

- Intelligence is *substrate neutral* and so no reason why machines cannot one day be as (or even more) intelligent than humans

- But how can information and computation be embodied in "dumb" physical matter ?

Information

- Information storage devices have components arranged in a way related to the state that the information is about
- Fundamental physical property of information storage devices is that they can be in *many different long lasting (stable) states (takes more energy to displace ball than that provided by random disturbances)*



- Simplest memory device has two stable states encoding a binary digit 1/0
- Can be physically embodied in different ways (on hard drive point on surface magnetized in 2 diff. ways, positions of electrons in working memory, fiber optics bits = strong or weak laser beams)

Information

=> Information is substrate neutral/independent

- In biological world Info storage in DNA and much more in brains ! 10 gig electrically (which neurons firing) and 100 tera chemically (strength of synaptic connections).

- Best computers can out-remember any biological system

Computation

- Computation transforms one information state to another, by implementing a *function* and highly complex functions imply accomplishment of highly complex goals

- Connect battery to electromagnet E with intervening 2 switches S1 & S2 so that E is on iff S1 and S2 are on. When E is on pulls open switch S3. Hence S3 off (0) only if S1 and S2 on (1 and 1), else S3 is on (1)

- We have a NAND gate and *any well defined function can be implemented by connecting NAND gates !*

- That is, NAND gates are *Turing complete* – they are equivalent to a mathematical model known as a Turing machine that is computationally universal in the sense that it can can do anything that **any** other computer can do – given enough time and resources it can compute any computable function

Computation

- Hence, universally **intelligent** machine: given enough time and resources can accomplish any goal as well as any other intelligent entity (UI potential to develop into **Life 3.0**)

- Computation is **substrate neutral/independent**.

Building Artificial Intelligence

	Human Based	Ideal Rationality
Reasoning-Based	Systems that think like humans	Systems that think rationally
Behaviour-Based	Systems that act like humans	Systems that act rationally

- AI researchers can be said to build systems that fall into one of the above quadrants.

Building Artificial Intelligence

	Human Based	Ideal Rationality
Reasoning-Based	Systems that think like humans	Systems that think rationally
Behaviour-Based	Systems that act like humans	Systems that act rationally

- Does any one definition imply (subsume) the other ?

- **Russell and Norvig** (*Artificial Intelligence: A Modern Approach*) fall within the acting rationally camp:

  What is AI?" Question recast as "What is intelligence?" and then identify intelligence with acting rationality.

  ***AI is the field devoted to building intelligent agents**, which are functions taking as input, tuples of percepts from the external environment, and producing behaviour (actions) on the basis of these percepts*.

Ideal Agents

- A perfectly rational agent models the environment **E**, considers the actions **a1**, …,**an** on that environment that result in changes to the states of the world, resulting in new environments **E1**, …,**En** and multiplies the probability each **ai** will result in **Ei** by the value/worth/utility of each **Ei** (i.e., the expected utilities). Then acts to bring about **Ei** with highest expected utility

- Usually not possible to build perfectly rational agents (eg can specify algorithm for invincible chess but not feasible to implement)

- **Calculatively rational** – programs that if executed infinitely fast would result in perfectly rational behavior

- **Bounded Optimality** – given a machine M (with its associated time and space constraints) what is the optimal program that given an environment E results in the agent acting to maximize expected utility

Universal (Artificial) Intelligence: *AIXI*

*AIXI (*Shane Legg and Marcus Hutter*)* = mathematically rigorous framework for defining optimal intelligence and provides a definition of intelligence that generalises the Russell and Norvig definition

Intelligence measures an agent's ability to achieve goals in <u>a wide range of environments</u>

- *AIXI* transforms above definition into meaningful equations and then studies those equations (See https://plato.stanford.edu/entries/artificial-intelligence/aixi.html for a good intuitive explanation of AIXI (also https://jan.leike.name/AIXI.html))

- Unfortunately, optimal (maximally) intelligent agent according to these equations is not Turing computable – challenge is to integrate resource bounds that approximate theoretical ideal

Types of Intelligence

Narrow AI and AGI

- Narrow A!: Purpose built algorithms for specific tasks/goals often outperforming human capabilities e.g. IBM's *Deep Blue*, Google Deep Mind's *Alpha Go* and *DQN AI* system (which can play many simple ATARI games at human level of better)

- Depth (AI) v breadth (human intelligence)

- Holy grail of AI is Artificial *General* Intelligence (AGI):

  - **the ability to accomplish *any* goal at least as well as humans**

  - possessing common sense and an ability to learn, reason and plan, to meet complex information processing challenges across a wide range of natural and abstract domains.

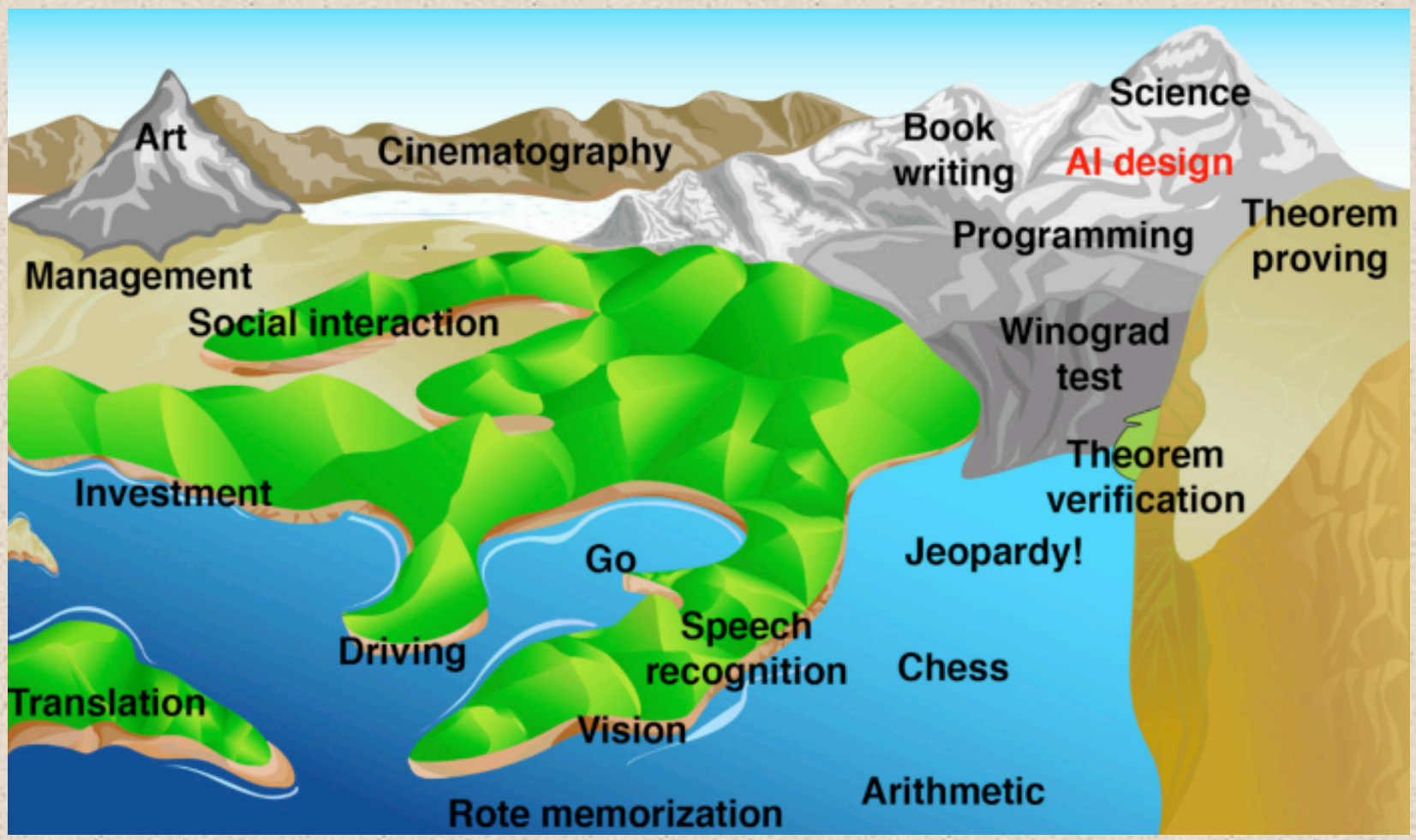
Moravec's Landscape of Human Competence

Computers are universal machines, their potential extends uniformly over a boundless expanse of tasks. Human potentials, on the other hand, are strong in areas long important for survival, but weak in things far removed.

Imagine a "landscape of human competence," having lowlands with labels like "arithmetic" and "rote memorization," foothills like "theorem proving" and "chess playing," and high mountain peaks labeled "locomotion," "hand-eye coordination" and "social interaction."

Advancing computer performance is like water slowly flooding the landscape. A half century ago it began to drown the lowlands, driving out human calculators and record clerks, but leaving most of us dry. Now the flood has reached the foothills, and our outposts there are contemplating retreat. We feel safe on our peaks, but, at the present rate, those too will be submerged within another half century. I propose that we build Arks as that day nears, and adopt a seafaring life!

Moravec's Landscape of Human Competence



Paths to AGI and Superintelligence

- Critical sea level corresponds to when machines can perform AI design, before which humans improve machines, after which machines improve machines (much faster than humans) – the ***singularity***

- But where are we now and what are the prospects ?

In the 40s AGI expected in 20 years ! Nowadays futurists still often settle on 20 years before advent of AGI

20 years is the sweet spot – near enough to be attention grabbing and relevant, yet far enough to imagine breakthroughs (cf shorter time scales – technologies that will have a big impact on the world are already in use while technologies that will reshape world in 15 years probably exist as prototypes)

Paths to AGI and Superintelligence

- Early AI adopted logic/symbolic paradigm – rules applied to data but an AI winter because of problems handling uncertainty, brittleness (small damage to logical knowledge base implies total crash), symbol grounding problem (humans “hard coding” initial beliefs)

- More recent optimism which instead of focussing only on high level symbol manipulation adopted connectionist paradigm (neural networks) that ‘degrade gracefully’ when there is small damage, and learn from experience (potentially solving symbol grounding problem) finding natural ways of generalising from examples and finding hidden statistical patterns in inputs.

- Back propagation algorithms (multilayered networks with hidden layers between input and output layers – abilty to learn a much wider range of functions) + advances in hardware and processing power fuelled explosion in use of neural networks and machine learning

When will AGI be attained ?

- Survey results from expert communities (taken from *Superintelligence*)

10 %	50%	90%
2022	2040	2075

From AGI to Superintelligence

- Once we have AGI *Superintelligence* is arguably a short step away
- Machines recursively improving and designing better versions of themselves, advantaged by massive superiority in speed of processing and data access (cf human AI researchers)
- **Superintelligence**: *any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest*

From AGI to Superintelligence

- Most popular version of the **technological singularity**: a hypothetical future point in time at which technological growth becomes uncontrollable and irreversible, resulting in unimaginable changes to human civilization.

How long from AGI to ASuperintelligence

Within 2 years after AGI	Within 30 years after AGI
10 %	75%

- What would be future impact of superintelligence on humantiy ?
- Range of opinions – approx 20% extremely good, approx 10% extremely bad (existential threat to mankind)
- Precautionary principle dictates that we should seriously address potential risks (more later in this course)

How will we know when we achieve AGI

- Human level performance in common sense reasoning (e.g. *frame problem*) and natural language understanding considered *AI complete* problems:

  Difficulty of solving these tasks essentially equivalent to difficulty of building *AGI*

- linguistic indistinguishability was the approach taken by Alan Turing in his classic 1950 paper *Computing Machinery and Intelligence*

- Turing begins the 1950 paper with the claim, "I propose to consider the question 'Can machines think?

- He rejects traditional approach of starting with defining the terms "machine" and "intelligence" and replaces question with a new one,

- Can machines do what we (as thinking entities) can do?" The advantage of the new question, Turing argues, is that it draws "a fairly sharp line between the physical and intellectual capacities of a man."

The Imitation Game (The Turing Test)

- Suppose an interrogator in a room separated from a person and a machine. The object of the game is for the interrogator to determine which of the other two is the person, and which is the machine. The interrogator is allowed to put questions to the person and the machine. The object of the machine is to try to cause the interrogator to mistakenly conclude that the machine is the other person; the object of the other person is to try to help the interrogator to correctly identify the machine.

- Turing says:

"I believe that in about fifty years' time it will be possible to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well that an average interrogator will not have more than 70 percent chance of making the right identification after five minutes of questioning"

The Imitation Game (The Turing Test)

The Turing Test

- Later in the paper Turing suggests an "equivalent" alternative formulation

  In the Turing Test (TT) a woman and a computer are in sealed rooms, and a human judge who doesn't know which of the two rooms contains which contestant, asks questions by email (actually, by teletype, to use the original term) of the two. If, on the strength of returned answers, the judge can do no better than 50/50 when delivering a verdict as to which room houses which player, we say that the computer in question has passed the TT.

  Passing in this sense operationalizes linguistic indistinguishability.

What is Artificial Intelligence ?

	Human Based	Ideal Rationality
Reasoning-Based	Systems that think like humans	Systems that think rationally
Behaviour-Based	Systems that act like humans	Systems that act rationally

- Which quadrant do you think the Turing Test falls in ?

The Imitation Game (The Turing Test)

- Interestingly, Turing anticipates the importance of (machine) learning as a means of bootstrapping a simpler system to human-level intelligence

"Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain"

- An interesting historical observation is that the TT was prefigured by the famous French philosopher Rene Descartes in his 1669 classic *Discourse on the Method* ("I think therefore I am")

The Imitation Game (The Turing Test)

- "If there were machines which bore a resemblance to our body and imitated our actions as far as it was morally possible to do so, we should always have two very certain tests by which to recognise that, for all that, they were not real men. The first is, that they could never use speech or other signs as we do when placing our thoughts on record for the benefit of others. …..

  …..But it never happens that it arranges its speech in various ways, in order to reply appropriately to everything that may be said in its presence, as even the lowest type of man can do. "

- Descartes thought it inconceivable that a mere machine could produce different arrangements of words so as to give an appropriately meaningful answer to whatever is said in its presence

An Analysis of the Turing Test

- Two kinds of questions about the TT

  Empirical Questions:

  Is it true that we will soon have computers that can play imitation game so well that an average interrogator has no more than a 70 percent chance of making the right identification after 5 minutes of questioning?

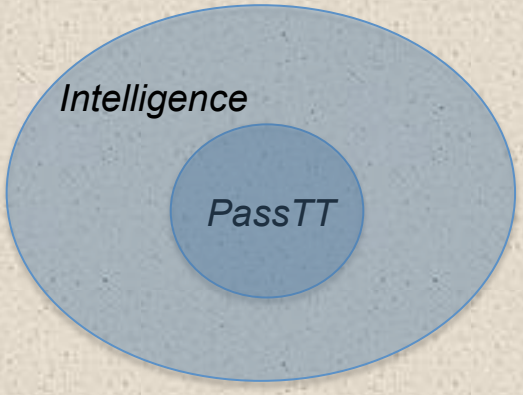
- In 2014, claims emerged that computer program Eugene Goostman fooled 33 percent of judges in the Turing Test 2014 competition (i.e., 70% threshold met)

  But there have been other one-off competitions in which similar results have been achieved.

  Result not reliably reproducible, and size of trial small. No strong grounds for claiming that an average interrogator had no more than a 70 chance of making the right determination about the relevant program after five minutes of questioning.

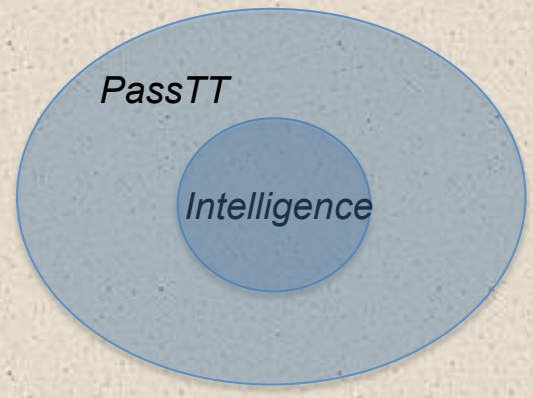
An Analysis of the Turing Test

logically sufficient conditions



⊆

logically necessary conditions



⊆

PassTT $\longrightarrow$ *Intelligence* (logically sufficient conditions - *PassTT* $\subseteq$ *Intelligence*)

Intelligence $\longrightarrow$ *PassTT* (logically necessary conditions *Intelligence* $\subseteq$ *PassTT*)

In general, A $\longrightarrow$ B is false iff A is true and B is false

Hence A $\longrightarrow$ B is true if A is true and B is true, or A is false and B is true, or B is false and A is false

An Analysis of the Turing Test

- Four cases in the philosophical literature are considered

1) **Logically sufficient and necessary conditions** *PassTT* $\leftrightarrow$ *Intelligence*

Claimed by very few people, but some objections to the TT only make sense assuming **logically necessary interpretation** (*Intelligence* $\rightarrow$ *PassTT*) which implies that if something fails TT then it cannot be intelligent

E.g. ***chauvinistic objection***

Intelligent creatures may fail TT because they do not share our way of life (the pragmatic conventions that govern the languages that they speak are so very different from the pragmatic conventions that govern human languages)

An Analysis of the Turing Test

Intelligent creatures may fail TT because they do not share our way of life (the pragmatic conventions that govern the languages that they speak are so very different from the pragmatic conventions that govern human languages)

So this **objection** to TT says that **may not pass TT but still be intelligent**. Hence it is an objection that assumes logically necessary interpretation of TT (which states that not passing TT implies not intelligent).

Indeed one can clearly conceive of achieving other **more** complex goals in a **wider** variety of environments than humans – surely considered intelligent.

Note that objection is an objection only because the question Turing asks

“Can Machines think ?”

is considered (chauvinistically) equivalent to

“Can a machine exhibit intelligent behaviour equivalent to, or indistinguishable from, that of a human ?”

An Analysis of the Turing Test

2) Only logically sufficient conditions *PassTT* $\rightarrow$ *Intelligence*

i.e., logically impossible that if pass TT then not intelligent.

Objections to this and other **behavioral** tests is a being whose behavior (eg passing TT) was produced by brute force methods ought not to count as intelligent (possessing a mind/having thoughts)

E.g. Blockhead is a creature that looks just like a human being, but is controlled by a look-up tree that contains a programmed response for every discriminable input at each stage in the creature's life. So it can use look up tree to create appropriate linguistic response to every possible question

An Analysis of the Turing Test

E.g. Blockhead is a creature that looks just like a human being, but is controlled by a look-up tree that contains a programmed response for every discriminable input at each stage in the creature's life. So it can use look up tree to create appropriate linguistic response to every possible question

If we assume: a) Blockhead is logically possible, and b) that since Blockhead uses brute force method of a look up table Blockhead is not intelligent (does not have a mind, does not think), **then** Blockhead is an objection (counter-example) to *PassTT* $\longrightarrow$ *Intelligence*

But we can deny logical possibility by (controversially) denying

conceivability $\longrightarrow$ *logical possibility*

Or insist that Blockhead is intelligent : after all it is processing information and that, together with its behavior, is grounds for saying Blockhead has some level of intelligence

An Analysis of the Turing Test

3) TT provides (more or less strong) probabilistic support for the attribution of intelligence.

Clear that Turing thought that passing TT would provide **probabilistic** support for the hypothesis of intelligence.

- The prediction Turing makes is itself probabilistic: in about 50 years it will be possible to programme computers to make them play the imitation game so well that an average interrogator will have **no more than a seventy per cent chance of making the right identification after five minutes of questioning**.

- Probabilistic nature of prediction gives good reason to think TT is itself of a probabilistic nature: a given level of success should produce a specifiable level of increase in confidence that the participant in question is intelligent

Since TT does not correlate levels of success with increases in confidence the TT is greatly underspecified. Relevant variables include length of questioning period; skills and expertise of interrogator (depth and difficulty of the questioning that takes place); the skills and expertise of the human player in the game, etc.

An Analysis of the Turing Test

Objections and Alternatives

1) **The TT is too hard.** Because nothing without a human cognitive substrate could pass the test.

2) **The TT is too narrow.** Objection to the notion that TT provides logically sufficient conditions for intelligence (eg Blockhead, can be adapted to show TT is too restrictive).

   Some think success in TT might come for reasons other than the possession of intelligence and is but one example of things that intelligent beings can do

   But success in TT implies cognitive competencies in memory, perception, maths, game playing, understanding politics etc etc

   Ability to pass TT, especially on repeated runs, implies *ability to solve complex goals in a wide range of environments*

An Analysis of the Turing Test

Objections and Alternatives

3) **The TT is too easy.** Some suggest a more demanding test e.g. The Lovelace Test, which judges the AI based on its ability to create, or to form an original idea.

In analysis of Turing Test, the reason Blockhead is said not to possess human level intelligence is that it simply uses a look up table to provide all possible responses to all possible questions but lacks any real *understanding* which is supposedly an important aspect of intelligence

Understanding assumes some relation between syntax (the words in the questions and responses) and semantics (the meaning of the words uttered). Not understanding can be interpreted as lacking any *conscious* apprehension of meaning.

Now, well known philosopher John Searle (in *Minds, Brains and Programs, Cambridge University Press, 1980*) uses a similar thought experiment to Blockhead to **argue against Turing's claim that an appropriately programmed computer could think**.

Arguments against the possibility of AGI

The Chinese Room

Imagine a native English speaker who knows no Chinese locked in a room full of boxes of Chinese symbols (a data base) together with a book of instructions for manipulating the symbols (the program).

Imagine that people outside the room send in other Chinese symbols which, unknown to the person in the room, are questions in Chinese (the input). And imagine that by following the instructions in the program the man in the room is able to pass out Chinese symbols which are correct answers to the questions (the output).

The program enables the person in the room to pass the Turing Test for understanding Chinese but he does not understand a word of Chinese.

Arguments against the possibility of AGI

The Chinese Room

The Chinese Room attempts to show that a computer can only do manipulation of symbols, **but has no true *understanding* which is one of the hallmarks of intelligence**, and so a computer cannot be intelligent

1. The symbol manipulating program *P* is a kind of digital computation that is possible in some remotely conceivable/imaginable world
2. P does not possess understanding and hence is not intelligent
3. Any actual digital computer *C* will not differ from *P* in any sense that makes it more intelligent than *P.* Therefore if P is not intelligent …
4. **Conclusion**: ….any actual digital computer *C* cannot be intelligent

Arguments against the possibility of AGI

1) Given assumption 2 (P does not possess understanding and **hence** is not intelligent) narrow conclusion of argument is that programming a digital computer may make it *appear* to understand language but it does not possess **real** understanding. Hence the TT is an inadequate test for intelligence

Computers merely use syntactic rules to manipulate symbol strings, but have no understanding of meaning or semantics.

2) Broader conclusion is that human minds are not like computers – simply information processing systems - as humans have real understanding but computers cannot have real understanding.

On the other hand, our *general definition* of intelligence - a system achieving (complex) goals in a wide variety of environments – does not necessarily imply requirement for “understanding”

So is Chinese Room argument really an argument against the inadequacy of the TT as a test for intelligence or the possibility of AGI ?

Only if understanding is necessary for intelligence (assumption 2).

Replies to the Chinese Room Argument

Let's looks at a reply from the philosopher Danniel Dennet

The Impossibility Reply challenges the argument's assumption (1) that the symbol manipulating program *P,* simulating Chinese understanding, is possible given the laws of physics. Hence the argument and its claim is not valid.

Dennet argues that *P* is not a **nomic** possibility (possible given the laws of physics)

Nomic possibility is not the same as logical possibility (which only requires that there is no inherent contradiction in what is thought of as being possible)

Such a program, to simulate passing the Chinese Turing test would probably be 100 billion lines of code, and so running the program by hand (or digitally) would probably take many lifetimes (weeks) before any Chinese came out the door.

Replies to the Chinese Room Argument

The System's Reply: Concedes that man in room doesn't understand Chinese. But the man is a part – the CPU – of a larger system that includes a huge database, memory containing intermediate states and the instructions. The system as a whole understands Chinese

The Virtual Mind (VM) Reply: Concedes that man in room doesn't understand Chinese. Unlike systems reply, *VM reply* states that whole system may create new virtual entity, distinct from system and its parts, that understands Chinese (cf Siri, Cortana, characters in video games).

The Robot Reply: Concedes Searle is right that whole system in room cannot understand language and know meaning of words. But if let loose in the world (e.g. as a robot) would come to understand meaning by interacting with world and others.

E.g. to "understand" meaning of Chinese word for hamburger reasonable to assume we know what a hamburger is because we have seen one, and perhaps even made one, or tasted one, or at least heard people talk about hamburgers and understood what they are by relating them to things we do know by seeing, making, and tasting.

Replies to the Chinese Room Argument

Searle-in-the-room, or the room alone, cannot understand Chinese. But if you let the outside world have some impact on the room, meaning or “semantics” and hence understanding might begin to get a foothold.

But of course, this concedes that thinking cannot be simply symbol manipulation.?