CS5010 AI Principles:
Lecture 2
The Turing Test & Attacks on AI

Written by Ian Gent

Revised & delivered by Tom Kelsey

Part 1: The Turing Test

Alan M Turing, Hero

- Helped to found theoretical CS
  - 1936, before digital computers existed
  - concieved the stored program concept
- Helped to found practical CS
  - wartime work decoding Enigma machines
  - ACE Report, 1946
- Helped to found practical AI
  - first (simulated) chess program
- Helped to found theoretical AI …
- Made a huge contribution to allied effort in World War II

Ian Gent peering into Alan Turing's office at Bletchley Park, 2013

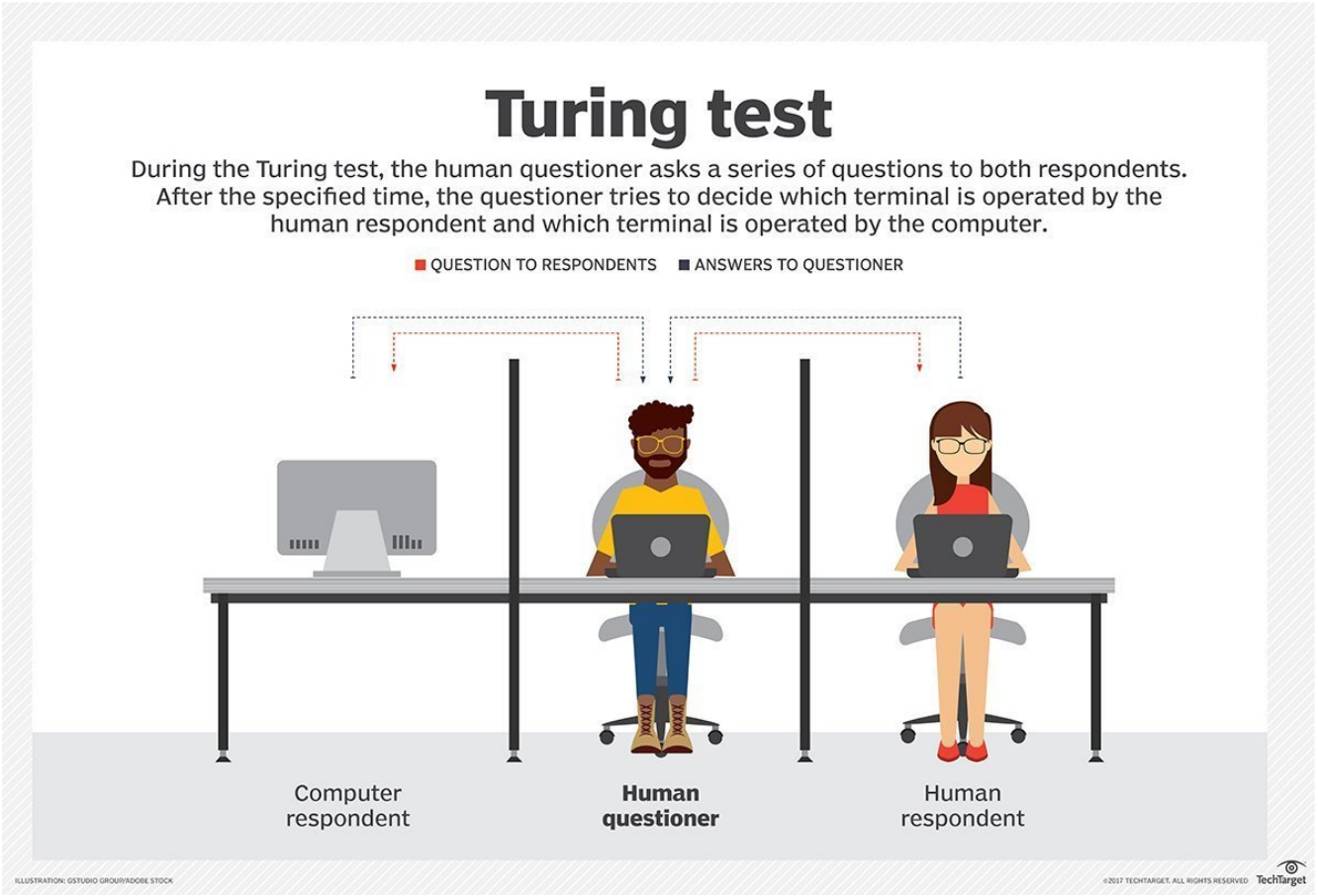
The Turing Test

- In 1950 Turing wrote about what is now called “The Turing Test”
- The original article is well worth reading
  - Not written in philosophobabble
- *Computing Machinery and Intelligence*
  - Alan M Turing
  - *Mind,* vol LIX, Number 236 (1950)
  - *http://m.mind.oxfordjournals.org/content/LIX/236/433.full.pdf*

Can Machines Think?

- Turing starts by defining *machine* & *think*
  - Will *not* use everyday meaning of the words
    - otherwise we could answer by Gallup poll
- Instead, use a different question
  - closely related, but unambiguous
- Turing comes up with what he calls “the imitation game”
  - Can we get a computer to imitate a human so that people can’t tell the difference?
- He is *not saying* this automatically means the computer can think
  - “I believe the original question to be too meaningless to deserve discussion”

The Imitation Game



Source: techtarget.com

- Interrogator in one room
  - digital computer in another room
  - person in a third room
- From typed responses only, can interrogator distinguish between person and computer?
- If the interrogator often guesses wrong, say the machine is intelligent.
- Usually done with one machine/person at a time

Turing suggested sample Q & A's:

Q: Write me a sonnet about the Forth Bridge

A: Forget this one, I never could write poetry

Q: Add 34957 to 70764.

(pause about 30 seconds)

A: 105621

Q: Do you play chess?

A: Yes

Q: I have K at my K1, and no other pieces. You have only K at K6 and R at R1. It is your move. What do you play?

(pause about 15s)

A: R-R8 mate

What did Turing think?

- Turing (in 1950) believed that by 2000
  - computers available with 128Mbytes storage
    - This is an astonishingly good prediction many years before Moore's law
    - Around 2000, a good PC would often come with 128MB RAM
      - Though admittedly much more disk space
  - programmed so well that interrogators have only a 70% chance after 5 minutes of being right
- "By 2000 the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted"

Was he right?

- Annual “Loebner Prize” competition
  - Turing Test funded by Mr Loebner
- 2008 Winning entry fooled 3 out of 12 judges
  - After 5 minutes conversation
  - i.e. 25% success rate, just 5% short of 30%
  - Turing Test very nearly passed, only 8 years off his prediction
- 2014 Reading University organises competition
  - As 60th anniversary memorial to Turing
  - Announces **TURING TEST SUCCESS MARKS MILESTONE IN COMPUTING HISTORY**
  - There’s arguments both ways whether the test was really passed (beware “Science by Press Release”)
- Not milestone in computing history
  - For most people …
- Why not?
  - First, Turing’s prediction was *wrong.*
  - Not about a computer winning the imitation game – only 14 years out if you accept Reading’s test
  - But he was wrong that “the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted”

Part 2: Attacks on AI

It depends what you mean by “AI”

- Typical joke academic answer “it depends what you mean by …”
- But actually it’s important to know what people are attacking
- Three things that could be under attack:
  - Weak AI
  - Strong AI (1)
  - Strong AI (2)
- We neeed some clarity here.

Weak AI and Strong AI (1)

- Phrases coined by John Searle (of “Chinese Room” fame)
- Weak AI takes the view that …
  - Computers are powerful tools
    - to do things that humans otherwise do
    - and to study the nature of minds in general
- Strong AI takes the view that
  - A computer with the right software ***is*** a mind
  - Really really a mind like you and I have though maybe different
  - Compare “robot arm”…
    - If you have a robot with an arm it can use to lift things
    - Nobody argues about whether that is “really” an arm.
    - Even though it’s obviously very different to my arm or yours.
- I will use “Strong AI” for the above
  - More recently “Strong AI” has come to have a rather different meaning as well

Strong AI (2)

- More recently “Strong AI” has come to have a second meaning
- Artificial General Intelligence (AGI)
  - The idea of making an AI system that can do anything
  - Though not necessarily do anything *well*
  - Compare some of our example AI Successes
    - Proving Checkers is a draw
      - Obviously very specific
    - IBM Watson winning at Jeopardy!
      - Less obviously still very specific
      - It is *only* good at playing Jeopardy
- How do we know if we have achieved AGI?

For clarity

- In this lecture, I will use:
- “Strong AI” for “Strong AI (1)” above
  - i.e. the theory that we can build computers to have minds
- “AGI” for “Strong AI (2)” above
  - i.e. the idea of building machines which can do more or less anything people can do

AGI and “AI-Complete” problems

- How do we know if we have achieved AGI?
- “Robot Coffee Test”
  - Proposed by Ben Goertzel
  - Can you make a robot that can go into a house and make some coffee?
  - Not a prepared house, but any house
  - Involves finding coffee, finding water, how to heat water …
  - And if you’ve built a robot to pass the coffee test …
    - Can the same robot learn how to build a brick wall?
    - Without being reprogrammed?
- “AI-Complete” Problems
  - Analogy with “NP-Complete” but not formally defined
  - An example of an AI-Complete problem is Machine Translation
  - To translate perfectly, a machine has to be able to know and understand almost all aspects of human experience

- Part 1: Attacks of various kinds on the Turing Test
- Part 2: The Chinese Room Argument
- Part 3: Killer Robots and the End of Humanity

1. Attacks on The Turing Test

- Some attacks that Turing addressed
- Including “The Mathematical Objection” and Lucas’s argument
- “How to pass the Turing test by cheating”

Objections and Responses

- The Theological Objection
  - Man has a soul, machines do not
  - *AT:* Can we deny His power to give a soul to a machine?
- Heads in the sand
  - I don't like the idea so I will ignore it
- Argument from various disabilities
  - No machine can X (e.g. tell right from wrong)
  - *AT*: Becomes a less powerful argument each day

Some more objections

- Lady Lovelace's [Ada's] objection
  - computers do whatever we know how to order them to perform, so computers cannot do anything really new
  - *AT*: Machines constantly surprise us.
- Argument from informality of behaviour
  - impossible to write down formal rules for every situation
  - *AT:* impossible to prove people not rule-driven
- Argument from ESP
  - Telepathy would let humans win imitation game
  - *AT*: Put competitors in 'telepathy-proof' room (!)

Two more serious objections

- Argument from Consciousness
  - "No mechanism could *feel* pleasure, grief …
  - *AT:* Danger of Solipsism
  - *AT:* Imitation game exists now - in oral exams
  - Probably the most contentious objection
- Argument from continuity in the nervous system
  - the brain does not operate digitally
  - *AT:* computers can simulate continuous behavior, e.g. Statistically, graphically, numerically

Solipsism:
"a theory in philosophy that your own existence is the only thing that is real or that can be known" (Merriam-Webster dictionary)

The mathematical objection

- Mathematical Objection
  - Godel's theorem, Halting problem, etc, show that machines cannot do 'meta-reasoning'.
  - *AT:* We too often give wrong answers ourselves to be justified in being very pleased at fallibility of machines
- Oddly, above refutation is complete but Lucas still used this attack
- "Minds, Machines and Gödel"
  - J.R. Lucas
    - Eminent philosopher, logician, etc
  - *Philosophy*, XXXVI, 1961, pp.(112)-(127)
  - http://users.ox.ac.uk/~jrlucas/Godel/mmg.html
  - Claims that Gödel's theorem makes strong AI impossible

And I quote…

…Gödel's theorem must apply to cybernetical machines, because it is of the essence of being a machine, that it should be a concrete instantiation of a formal system. It follows that given any machine which is consistent and capable of doing simple arithmetic, there is a formula which it is incapable of producing as being true--i.e., the formula is *unprovable-in-the-system*--but which we can see to be true. It follows that no machine can be a complete or adequate model of the mind, that minds are essentially different from machines…

J.R. Lucas, "Minds, Machines and Gödel"

Minds, Machines and Gödel

- Premises
  - 1. Gödel's theorem shows that any consistent and powerful formal system must be limited
    - there must be true statements in cannot prove
  - 2. Computers are formal systems
  - 3. Minds have no limit on their abilities
- Conclusion
  - Computers cannot have minds
- Should Strong AI give up and go home?
  - Certainly Gödel's theorem applies to computers

Refuting Lucas:

- Turing decisively refuted Lucas on two counts

1. "Although it is established that there are limitations to the powers of any particular machine, it has only been stated without any sort of proof, that no such limitations apply to the human intellect"
   - i.e. are we sure humans can prove ***all*** true theorems?
   - Maybe humans *are* unlimited? What then? So…
2. Turing's second point is decisive
   - "We too often give wrong answers ourselves to be justified in being very pleased at such evidence of fallibility on the part of machines."
   - Gödel's theorem applies only to consistent formal systems
   - Humans often utter untrue statements
   - We might be unlimited formal systems which make errors

An example of Rogerian psychotherapy

Patient	Men are all alike.
Doctor	*In what way*
Patient	They're always bugging us about something or other.
Doctor	*Can you think of a specific example*
Patient	Well, my boyfriend made me come here
Doctor	*Your boyfriend made you come here*
Patient	He says I'm depressed much of the time.
Doctor	*I am sorry to hear that you are depressed*
Patient	It's true. I am unhappy
Doctor	*Do you think coming here will help you not to be unhappy*
Patient	I need some help, that much seems certain
Doctor	*What would it mean to you if you got some help*
Patient	Perhaps I could learn to get along with my mother
Doctor	*Tell me more about your family*
Patient	My mother takes care of me
Doctor	*Who else in your family takes care of you*

"Rogerian therapy involves the therapist's entry into the client's unique phenomenological world. In mirroring this world, the therapist does not disagree or point out contradictions. Neither does he/she attempt to delve into the unconscious. The focus is on immediate conscious experience"
Dagmar Pescitelli, 1996

When did this dialogue happen?

- The “doctor” was actually a computer
- It’s not a huge spoiler that this dialogue comes from AI
- Guess when this dialogue came from (if you don’t know)
  - 1950s?
  - **1966: !!!**
    - **“**ELIZA—a computer program for the study of natural language communication between man and machine”
    - Joseph Weizenbaum
    - Communications of the ACM, Jan 1966
  - 1970s?
  - 1980s?
  - 1990s?
  - 2000s?
  - 2010s?

The problem with the Turing Test

- Chatbots trying to pass the Turing Test use *lots of tricks*
- Eliza used simple pattern matching
  - "Well, *my* boyfriend made *me* come here"
  - *"Your* boyfriend made *you* come here?"
- Jason Hutchens programmed the 1996 Loebner Prize winner
  - Then wrote an article
  - "How to pass the Turing test by cheating" (!!)
  - "Turing's imitation game in general is inadequate as a test of intelligence, as it relies solely on the ability to fool people, and this can be very easy to achieve, as Weizenbaum found."
- E.g. the winner of the 2014 Turing Test was "Eugene"
  - Eugene pretended to be a 13-year-old, non-native-English-speaking Ukrainian.
  - This gives lots of ways for judges to explain away anything dubious
- Whether or not it has been passed, the Turing Test does not drive much AI research

Part II: The Chinese Room

- This is one of the most famous attacks on AI

The Chinese Room

- John Searle
  - “Minds, Brains, and Programs”
  - The Behavioral and Brain Sciences, vol 3, 1980
  - http://dx.doi.org/10.1017/S0140525X00005756
  - http://cogprints.org/7150/1/10.1.1.83.5248.pdf
  - Attacks with the ‘Chinese Room’ argument
- Remember, Searle is attacking Strong AI
  - he attacks claims that, e.g. story understanding programs
    - *literally* understand stories
    - *explain* human understanding of stories

John Searle in 2005
Image by Matthew Breindel
https://en.wikipedia.org/wiki/File:John_searle2.jpg

The Chinese Room Thought Experiment

- A thought experiment
- Aimed at showing conscious computers are impossible
- By analogy with an obviously ridiculous situation
- John Searle does not understand Chinese
- Imagine a set up in which he can simulate a Chinese speaker

Locked in a Chinese Room

- John Searle is locked in solitary confinement
- He is given lots of …
  - blank paper, pens, and *time*
  - lots of Chinese symbols on bits of paper
  - an in tray and out tray
    - for receiving and sending Chinese messages
  - rule books written in English (which he does understand)
    - telling how to take paper from in-tray, process it, and put new bit of paper with symbols on it in out-tray

Outside the Chinese Room

- Unknown to Searle, his jailers …
  - regard the in-tray as containing input from a Chinese player of Turing's imitation game
  - the rule books as containing an AI program
  - regard the out-tray as containing responses
- Searle might pass the Turing Test in Chinese
- But *still not understand Chinese*
- By analogy, even a computer program that passes the Turing test does not truly "understand"

Objections and Responses

- Like Turing, Searle considers various objections
- 1. The Systems Reply
- *"The whole system (inc. books, paper) understands"*
  - Searle: learn all rules and do calculations all in head
    - still Searle (i.e. whole system) does not understand
- 2. The Robot Reply
- *"Put a computer inside a robot with camera, sensors etc"*
  - Searle: put a radio link to the room inside the robot
    - still Searle (robot's brain) does not understand

Objections and Responses

- 3. The Brain Simulator Reply
- *"computer simulates neurons, not AI programs"*
  - Searle notes this seems to abandon AI after all!
  - Searle: there is no link between mental states and their ability to affect states of the world
  - "As long as it simulates only the formal structure of a sequence of neuron firings … it won't have simulated what matters about the brain, namely its causal properties, its ability to produce intentional states"
    - "intentional states": that feature of mental states by which they are directed at states of affairs in the world

Is Searle right?

- Almost universally disagreed with by AI writers
- No 100% rock solid refutation
- Some points to ponder
  - Is a thought experiment valid?
    - E.g. 3GHz x 1 hour >> Searle processor ≈ 10^{13} operations >> 1 human lifetime@ 1 KHz
  - If machines lack intentionality, where do humans get it?
  - Is AI a new kind of duality?
    - Old: mind separate from body (villified by AI people)
    - New: thought separate from brain
  - Does it matter, or only matter for Turing tests?

Part 3: Killer Robots and the End of Humanity

Part 3: Killer Robots and the End of Humanity

- Concern about AI ending humanity goes way back
  - Remember Rossum's Universal Robots
- Recently been the focus of much more attention
- But has been addressed seriously in the past

Joseph Weizenbaum

- Weizenbaum wrote ELIZA in mid 1960's
- Shocked by reactions to such a simple program
  - people (including his secretary) wanted private conversations
  - psychotherapists suggested use of automated therapy programs
  - people believed ELIZA solved natural language understanding
- He wrote a book criticising AI
- "Computer Power and Human Reason"
  - Penguin, 1976 (second edition 1985)
- Weizenbaum does not attack possibility of AI
- Attacks the *use* of AI programs in some situations

What's the problem?

- "The question I am trying to pursue here is:
  - 'What human objectives and purposes may not be appropriately delegated to a computer?' "
- He claims that the Artificial Intelligentsia claim
  - there is no such domain
  - *Actually not true (at least now) as we will see*
  - Weizenbaum argues that many decisions should not be handled by computer
  - e.g. law cases, psychotherapy, battlefield planning
- Especially because large AI programs are 'incomprehensible'
  - e.g. you may know how a Deep Blue works
    - but not the reason for a particular move vs Kasparov
- But knowledge of the emotional impact of touching another person's hand "involves having a hand at the very least"
  - Should machines without such knowledge be allowed power over us?

Killer Robots

- Weizenbaum seems to have (partly) lost the argument
  - In the sense that many decisions about our lives *are* made by AI all the time without us knowing
  - The schedule for classes or trains or planes, pricing decisions, etc etc
- But what about life or death decisions?
  - We talked about Foot's thought experiment
    - Kill 5 people by inaction or Kill 1 person by taking action?
  - What about decisions in war?
- Unmanned vehicles (e.g. drones) now regularly used to kill
  - However usually controlled by human operators
  - Though note that AI may be critical to their success
    - E.g. moment to moment control of flight surfaces
- What about giving drones the power to decide to kill?

Most people in AI don’t want killer robots

- “AUTONOMOUS WEAPONS: AN OPEN LETTER FROM AI & ROBOTICS RESEARCHERS”
  - Signatories include Stuart Russell, Peter Norvig

“In summary, we believe that AI has great potential to benefit humanity in many ways, and that the goal of the field should be to do so. Starting a military AI arms race is a bad idea, and should be prevented by a ban on offensive autonomous weapons beyond meaningful human control.”

Toby Walsh TEDxBerlin talk, 2015
“How you can stop killer robots”

Existential Threat to Humanity

- Some serious people have proposed that AI could end human life
  - Stephen Hawking, Elon Musk, Bill Gates
- Main mechanism proposed is through AI surpassing human intelligence, then designing future AIs to surpass it even more …
- A process of "recursive self improvement" leading to a "singularity"
- And at some point through malice, or warped priorities, ending human life

"A system that is optimizing a function of n variables, where the objective depends on a subset of size k<n, will often set the remaining unconstrained variables to extreme values; if one of those unconstrained variables is actually something we care about, the solution found may be highly undesirable. ***This is essentially the old story of the genie in the lamp, or the sorcerer's apprentice, or King Midas: you get exactly what you ask for, not what you want.*** *A highly capable decision maker – especially one connected through the Internet to all the world's information and billions of screens and most of our infrastructure – can have an irreversible impact on humanity."*

Stuart Russell, edge.org

Should you worry?

- I don’t lose any sleep about this
- I *do* think there are many ethical issues around AI
- I *don’t* think that ending humanity is one of them
- My efforts in AI are aimed at improving life for humans
- So maybe I’m biased, naïve, credulous, wrong, …
- If you are worried there is *another* open letter about this
- *An Open Letter: RESEARCH PRIORITIES FOR ROBUST AND BENEFICIAL ARTIFICIAL INTELLIGENCE*

Attacks on AI: Summary

- Attacks on **Strong AI**
  - Mathematical objection decisively refuted by Turing
  - The Turing Test may or may not have been passed
    - But either way it doesn't say much about *Strong AI*
  - The Chinese Room thought experiment has never been refuted or confirmed
- Attacks on **Weak AI**
  - Weizenbaum wanted limits on what AI is allowed to do
    - Those limits are probably not being respected
  - AI researchers do not want to allow automonous killing machines
- Attacks on **Artificial General Intelligence**
  - There are those who think AI poses an existential threat to humanity
  - I'm not one of them

Next Lecture

- Interactive session 1
  - Teams Thursday 4pm
- Machine Learning 1
  Monday 4pm