Philosophy of AI
Lecture 4: What is a good test for intelligence?

Natasha Alechina
n.a.alechina@uu.nl

Plan of the lecture

1. What this lecture is NOT about
2. What this lecture IS about
3. Problems with TT as a test for intelligence
4. Hector Levesque's 'On our best behaviour' paper (AIJ 2014)
5. If there is time: knowledge based programs
6. Examples of knowledge-based systems

What this lecture is NOT about

- NOT: what is intelligence?
- NOT: can machines think?
- NOT: can machines be conscious?
- or have feelings, or be truly creative, or have mystical experiences, ...

What this lecture IS about

- Assuming that AI is about developing programs that do things that in humans would require intelligence
- . . . is the Turing Test a good measure of progress in AI?
- What are the current challenges/good measures of progress in AI?

Test for progress in AI

- We are not looking for a *definition* of intelligence
- We are looking for a (necessarily partial) *test* for intelligence
- The purpose is to measure progress in AI research.

Progress in AI: Where are we?

- Some things that were deemed impossible have been achieved
- The chess program that beat Dreyfus, and Deep Blue that beat Kasparov, were both ‘GOFAI’ (symbolic)
- Big breakthroughs in pattern recognition (image recognition, games like Go): sub-symbolic AI, machine learning, methods ‘in between’ symbolic and sub-symbolic (Alpha Go uses Neural Networks and Monte Carlo Tree Search)
- Natural language understanding: big advances (a lot achieved with statistics techniques), still a lot to do (more later in this lecture)

More on progress in AI: Beginning of Stuart Russell's AAAI 2020 invited talk

- Russell AAAI 2020

Problem with TT as a test for intelligence

- exercise in cheating/lying
- best done with imitating human shortcomings (slow erratic typing, small short term memory, reasoning mistakes)
- also helps to make jokes, change topic, evade
- imitating a particularly silly human works best; which is ironic for an intelligence test

Case in point: Eugene Goostman

- Eugene Goostman was a program that won (1/3 vote) Loebner prize
- pretended to be a 13 year old Ukranian boy

Eugene Goostman conversations

Conversation with Gary Marcus (from *The New Yorker* article)

> *Marcus: Do you read The New Yorker?*
> *Goostman: I read a lot of books . . . So many – I don't even remember which ones.*
> *Marcus: You remind me of Sarah Palin.*
> *Goostman: If I'm not mistaken, Sarah is a robot, just as many other "people." We must destroy the plans of these talking trash cans!*

Eugene Goostman conversations

Conversation with Scott Aaronson (from Levesque's book *Common Sense, the Turing Test, and the Quest for Real AI*)

> *Aaronson: What is bigger, a shoebox or Mount Everest?*
> *Goostman: I can't make a choice right now. I should think it out later. And I forgot to ask you where you are from . . .*
> *Aaronson: Do you think Alan Turing, brilliant though he was, had trouble imagining that the judges of his "imitation game" wouldn't think to ask commonsense questions like the ones above – or that, if they did, they'd actually accept evasion or irrelevant banter as answers?*

Continued: Scott Aaronson and Eugene Goostman

Goostman: No, not really. I don't think alan turing brilliant although this guy was had trouble imagining that the judges of his imitation game would not consider to Oooh. Anything else?
Aaronson: In your opinion, does your existence demonstrate any idea or principle that wasn't demonstrated just as convincingly by ELIZA, Joseph Weisenbaum's chatbot from 1960s?
Goostman: ELIZA was a break-thru. All the bots after it were nothing but weak parodies, claiming to have "revolutionary improvements".
Aaronson: Hey, that's the first sensible thing you've said!

Hector Levesque's 'On our best behaviour' paper (AIJ 2014)

- Levesque's definition of AI:
  AI is the study of intelligent behaviour in computational terms
- Science of AI: how is it possible for something physical (people, computers,...) to actually do (some intelligent behaviour) X?
- Different (technology?) question: Can we engineer a computer system to do something that is vaguely X-ish?

Turing Test

- Levesque agrees that something is intelligent if its behaviour cannot be distinguished from a human behaviour in long term
- however, the short, open conversation test does have problems

Alternative proposal

- Answering fixed form questions in English (that can be easily answered by people who understand English)
- no dodging the question possible; also things like human typing speed, way of using natural language etc. don't arise
- can be administered by a computer and just marked for correctness.
- requirement: should be only doable in a similar way that people answer such questions, not using 'cheap tricks' such as checking whether some words ever occur together on the web etc.

Winograd Schema Challenge

- Levesque proposed Winograd Schema Challenge (to replace Turing test)
- Winograd Schemas are named after Terry Winograd (computer scientist from Stanford University)
- Examples:
  - The trophy would not fit in the brown suitcase because it was too big. What was too big?
    1. the trophy
    2. the suitcase
  - Joan made sure to thank Susan for all the help she had given. Who had given the help?
    1. Joan
    2. Susan

Why Winograd Schemas?

- cannot be answered by checking statistical correlations; require knowledge and common sense reasoning
- use 'swappable' special words (large/small, given/received)
  - The trophy would not fit in the brown suitcase because it was too small. What was too small?
    1. the trophy
    2. the suitcase
  - Joan made sure to thank Susan for all the help she had received. Who had received the help?
    1. Joan
    2. Susan

Special words

- when words are swapped, correct answer changes
- it should not be more likely that one of the pair of words is much more often associated with one of the answers
- e.g. trophy and small, or trophy and large, same for the suitcase
- example of a bad WS question (that can be answered by a google search):
  - The racecar easily passed the school bus because it was going too fast. What was going too fast? (alternative: slow)
    1. the racecar
    2. the school bus

10 minute exercise

- (as a group of 3-4 people)
- Design your own Winograd schema
- Swap with another group and check if you can answer their question using Google

Winograd schema challenge competition

- a competition called Winograd Schema Challenge took place at IJCAI 2016 and was offered at AAAI 2018
- sponsored by Nuance Communications, the prize of 25 000 USD offered to a program that can match human performance (Nuance no longer offers the prize)
- two rounds, the first round is in pronoun disambiguation (not WSC, literary sources); if a program is within 3% of human performance, proceeds to the second round (WSC proper)

WSC competition 2016

- the best performance in PDP (pronoun disambiguation problems) were:
  Liu, Quan; Jiang, Hui; Ling, Zhen-Hua; Zhu, Xiaodan; Wei, Si; Hu, Yu from the University of Science and Technology, China.
- well below human performance (58% correct), so did not get to second round
- work continues on solving this problem

Current state of the art in WSC

- Trichelair et al 2019: how much progress is genuine and how much due to problems with experimental set up etc.
- internal validity: no alternative explanation for success of experiments, external validity: can be generalised to other settings
- main problems with WSC: predictable structure; limited number of possible questions; still some associativity (some answers statistically correlate)
- evaluation protocol: removed associative examples; switched names etc. (so the answers change, and if learnt on original question the answer is wrong)
- conclusion: there is some (small) genuine progress

Other alternatives to Turing Test

- Ortiz (AI Magazine 2016): need physically embodied Turing Test, incorporating perception and action ...
- Pearl (Comm. ACM 2019): understanding causal relationships ...
- Lenat (AI Magazine 2016): changing our life (real personal assistants, education, health, economy, democracy, ...)

Embodied Turing Test

- Motivation: to test all aspect of intelligence and bring together diverging areas of AI
- Original proposal by Ortiz:
- a human judge interacts with a robot and a human
- the human teleoperates the same set of robot manipulators and has the same video sensors
- the judge sees the same sensors and manipulators from both
- was considered too difficult for humans (to teleoperate); replaced by a series of challenges
- Some work that could compete: IkeaBot from MIT (collaborates with humans on IKEA furniture assembly)

Pearl: what is necessary for intelligence

- challenges to machine learning approaches:
- adaptability, ability to modify behaviour in new circumstances
- explainability
- understanding of cause-effect connections; asking and answering 'what if' questions
- Pearl argues that they can be overcome with causal modelling tools
- causal models are *essentially* directed graphs (causes to effects) with conditional probabilities attached

Lenat: when will we say that Real AI is here?

- when everyone has a PDA knowing all about them, recognised in legislation etc.
- when all education is tailored personally using in particular the knowledge that the PDA has
- economy is much more rational (decision making by PDA, fraud detection etc.)
- politics: less fraud, everyone informed, everyone treated as individual, but politicians also have better access to simulations of decisions ...
- personal experience ... quasi immortality, cloning ...

Common theme: knowledge

- The common theme of alternatives to TT is programs using common sense knowledge, and making decisions based on it
- the area of AI that studies it is called Knowledge Representation and Reasoning

Knowledge Representation and Reasoning

- How can knowledge be represented symbolically and manipulated in an automated way by reasoning programs
- Knowledge: some information about the world
  - medical information about some particular set of diseases: what causes them, how to diagnose them
  - geographical data: which city is the capital of which country, population statistics, ...
  - common sense physics: bodies cannot go through solid walls, ...
- Representation: how / in which language do we represent this information
- Reasoning: how to extract more information from what is explicitly represented (because we cannot represent every single fact explicitly as in a database)

Knowledge-based systems

- We want to be able to talk about some AI programs in terms of what they 'know'
  - (which corresponds to taking 'intentional stance' towards those systems, ascribing them human characteristics, see Daniel Dennett, *Intentional Systems* 1971 article and subsequence work)
- ... and not just talk about what they know but also have something to point to in those systems corresponding to 'knowledge' and determining their behaviour, namely *explicitly represented symbolic knowledge*

Example (Brachman and Levesque)

Two Prolog programs with identical behaviour:

```
printColour(snow) :- !, write("It's white.").
printColour(grass) :- !, write("It's green.").
printColour(sky) :- !, write("It's yellow.").
printColour(X) :- !, write("Beats me.").
```

and

```
printColour(X) :- colour(X,Y), !, write("It's "),
write(Y), write(".").
printColour(X) :- write("Beats me.").
colour(snow, white).
colour(sky, yellow).
colour(X,Y) : - madeof(X,Z), colour(Z,Y).
madeof(grass, vegetation).
colour(vegetation, green).
```

Which one is knowledge-based

- Only the second program has explicit representation of 'knowledge' that snow is white
- the second program does what it does when asked for the colour of snow *because of* this knowledge. When `colour(snow, white)` is removed, it will not print the right colour for snow.
- what makes the system knowledge-based is not
  - the use of a particular logical-looking language like Prolog
  - or having representation of true facts (`colour(sky, yellow)` is not)
  - or having lots of facts, or having a complex structure
- rather, it is *having explicit representation of knowledge which is used in the operation of the program*

Definition of knowledge-based systems and knowledge bases

- Knowledge-based systems are systems for which intentional stance is grounded by design in symbolic representation
- The symbolic representation of knowledge is called a knowledge base.

Examples of knowledge-based systems

- Various expert systems
  - MYCIN (1970s, Stanford University)
  - XCON (1978, Carnegie Mellon University)
- Perhaps most famous knowledge base: CYC (1980s, Douglas Lenat, Cycorp, Austin, Texas)
- Ontologies
  - Snomed CT `http://snomed.dataline.co.uk/`
  - Gene ontology `http://www.geneontology.org/`
- Google Knowledge Graph
- (Parts of) IBM Watson

MYCIN

- 1970s, Stanford University (Edward Shortliffe, Pat Buchanan)
- Production rule system (we will see them later in the course)
- Purpose: automatic diagnosis of bacterial infections
- Lots of interviews with experts on infectious diseases, translated into rules (knowledge acquisition is a non-trivial process; also see later in the course)
- approximately 500 rules

Example MYCIN rule

Rule in LISP:

```
RULE035
PREMISE: ($ AND (SAME CNTXT GRAM GRAMNEG)
                (SAME CNTXT MORPH ROD)
                (SAME CNTXT AIR ANAEROBIC))
ACTION: (CONCLUDE CNTXT IDENTITY BACTEROIDES TALLY
.6)
```

English translation:

IF:

1. the gram stain of the organism is gramneg, and
2. the morphology of the organism is rod, and
3. the aerobicity of the organism is anaerobic

THEN: There is suggestive evidence (.6) that the identity of the organism is bacteroides

More about MYCIN

- some facts and some conclusions of the rules (as above) are not absolutely certain
- MYCIN uses numerical *certainty factors*; range between -1 and 1
- (reasonably involved) rules for combining certainty factors of premises, with the number in the rule (as 0.6 above) into a certainty factor for the conclusions
- later it turned out that MYCIN's recommendations would have been the same if it used only 4 values for certainty factors
- MYCIN was never used in practice (ethical and legal issues)
- when tested on real cases, did as well or better than the members of the Stanford medical school

XCON

- John McDermott, CMU, 1978
- eXpert CONfigurer - system for configuring VAX computers
- production rule system, written using OPS5 (language for production systems, implemented in LISP)
- 10,000 rules
- used commercially

Cyc

- The Cyc Knowledge Server is a very large knowledge base and inference engine
- Developed by Cycorp (Lenat is the founder): `http://www.cyc.com/`
- It aims to provide a deep layer of 'common sense knowledge', to be used by other knowledge-intensive programs

Cyc knowledge base

- Contains terms and assertions in formal language CycL, based on first-order logic, syntax similar to LISP
- Knowledge base contains classification of things (starting with the most general category: Thing), and also facts, rules of thumb, heuristics for reasoning about everyday objects
- Currently, over 200,000 terms, and many human-entered assertions involving each term; Cyc can derive new assertions from those
- Divided in thousands of ‘microtheories’

Cyc knowledge base

- General knowledge: things, intangible things, physical objects, individuals, collections, sets, relations...
- Domain-specific knowledge, for example:
  - Political geography: general information (e.g. What is a border?) and specific information about towns, cities, countries and international organizations
  - Human anatomy and physiology
  - Chemistry

Snomed

- Snomed CT: Systematized Nomenclature of Medicine Clinical Terms
- Developed by College of American Pathologists and the NHS
- Clinical terminology (with formal definitions)
- Designed for unambiguous recording of data and interoperability with software applications
- Uses ontology language (different from first order logic) EL++
- Approx. 400 000 concepts, 1 million terms and 1.6 million relationships

Snomed: example

Concept: 32553006 - Hangover

Descriptions:

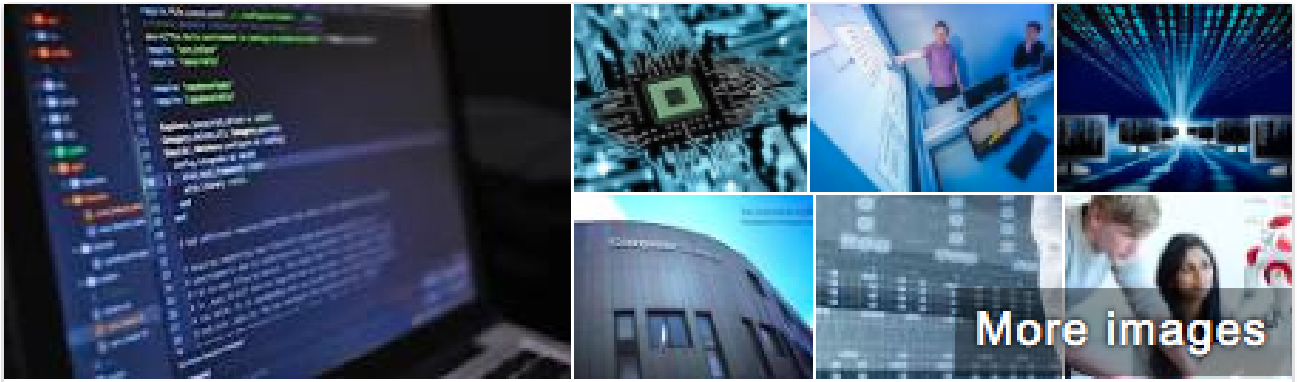
Synonym: hangover effect
Synonym: hangover from alcohol

Relationships:

(is a) 228273003 - Finding relating to alcohol drinking behavior
(causative agent) 311492009 - Ingestible alcohol

Google's Knowledge Graph

- based on an earlier knowledge base, Freebase (bought by Google in 2010)
- is used to enhance search results by displaying Wikipedia-style entry in an infobox alongside search results
- there is also a Google API which allows programmers to use Knowledge Graph
- used in Google Assistant and Google Home to answer questions



Computer Science

Field of study

Computer science is the study of the theory, experimentation, and engineering that form the basis for the design and use of computers.
Wikipedia

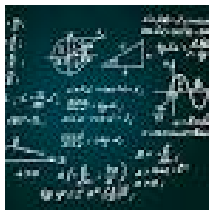
People also search for

View 10+ more

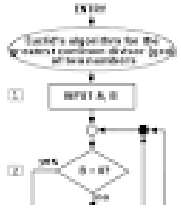
Science

Computers

Engineering



Mathemat...



Algorithm

Watson

- developed at IBM by a team led by David Ferrucci
- question answering system (originally developed to play Jeopardy)
- has access to terabytes of data (all Wikipedia pages, other encyclopedias)
- Jeopardy instance of Watson had a knowledge base, but
- mostly used statistical correlation methods in plain English text to find answers to questions such as, which city has airports named after a WWII hero and a WWII battle
- many other instances of Watson are used in medicine and other areas

Watson playing Jeopardy