Artificial Intelligence Foundation – JC3001

Lecture 46: Ethics in AI - I

Prof. Aladdin Ayesh (aladdin.ayesh@abdn.ac.uk)

Dr. Binod Bhattarai (binod.bhattarai@abdn.ac.uk)

Dr. Gideon Ogunniye, (g.ogunniye@abdn.ac.uk)

October 2025

1495 UNIVERSITY OF ABERDEEN

Material adapted from:
Russell and Norvig (AIMA Book): Chapter 28

UNIVERSITY OF ABERDEEN

Course Progression

- Part 1: Introduction
  1. Introduction to AI ✓
  2. Agents ✓
- Part 2: Problem-solving
  1. Search 1: Uninformed Search ✓
  2. Search 2: Heuristic Search ✓
  3. Search 3: Local Search ✓
  4. Search 4: Adversarial Search ✓
- Part 3: Reasoning and Uncertainty
  1. Reasoning 1: Constraint Satisfaction ✓
  2. Reasoning 2: Logic and Inference ✓
  3. Probabilistic Reasoning 1: BNs ✓
  4. Probabilistic Reasoning 2: HMMs ✓
- Part 4: Planning
  1. Planning 1: Intro and Formalism ✓
  2. Planning 2: Algorithms & Heuristics ✓
  3. Planning 3: Hierarchical Planning ✓
  4. Planning 4: Stochastic Planning ✓
- Part 5: Learning
  1. Learning 1: Intro to ML ✓
  2. Learning 2: Regression ✓
  3. Learning 3: Neural Networks ✓
  4. Learning 4: Reinforcement Learning ✓
- Part 6: Conclusion
  1. **Ethical Issues in AI**
  2. Conclusions and Discussion

Objectives

- Limits of Current and Future AI
- Ethical Issues

Outline

Philosopher John Searle (1980):

- **weak AI:** the idea that machines could act as if they were intelligent
- **strong AI:** the assertion that machines that do so are actually consciously thinking (not just simulating thinking)

The argument from informality

Turing's "argument from informality of behaviour" says that human "behaviour" is far too complex to be captured by any formal set of rules

Good Old-Fashioned AI (GOFAI)

- The simplest logical agent design has limitations
  - qualification problem: difficult to capture every contingency of appropriate behaviour in a set of necessary and sufficient logical rules
- Dreyfus's strongest arguments is for situated agents rather than disembodied logical inference engines
- Embodied cognition approach claims that it makes no sense to consider the brain separately
  - cognition takes place within a body, which is embedded in an environment

The argument from disability

- The “argument from disability” makes the claim that “a machine can never do X.”
- Turing’s lists of X:
  Be kind, resourceful, beautiful, friendly, have initiative, have a sense of humour, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, make someone fall in love with it, learn from experience, use words properly, be the subject of its own thought, have as much diversity of behaviour as man, do something really new.
- Some of these are rather easy to be replicated by AI.
  However, some are not possible
- Overall, programs exceed human performance in some tasks and lag behind on others.
- The one thing that it is clear they cannot do is to be exactly human.

The mathematical objection

Turing (1936) and Gödel (1931) proved that certain mathematical questions are in principle unanswerable by particular formal systems.

Gödel sentence $G(F)$ with the following properties:

- $G(F)$ is a sentence of F, but cannot be proved within F.
- If F is consistent, then $G(F)$ is true.

Philosophers such as J. R. Lucas (1961) have claimed that this theorem shows that machines are **mentally inferior to humans**,

- machines are **formal systems** that are limited by the incompleteness theorem
- cannot establish the truth of their own Gödel sentence
- Problems with Lucas' claim:
  - Example sentence which cannot consistently assert by human else contradiction: Lucas cannot consistently assert that this sentence is true.
  - **No entity**–human or machine—-can prove things that are impossible to prove
  - incompleteness theorem technically applies only to formal systems that are powerful enough to do arithmetic.

Measuring AI

- Turing test: whether machines can pass a behavioural test
- The test requires a program to have a conversation (via typed messages) with an interrogator for five minutes
- ELIZA program and Internet chatbots such as MGONZ and NATACHATA
- Eugene Goostman fooled 33% of the untrained amateur judges in a Turing test
- Recently, LamDA even fooled a Google “engineer”
- AI researchers who crave competition are more likely to concentrate on playing chess or Go or StarCraft II, or taking an 8th grade science exam, or identifying objects in images.

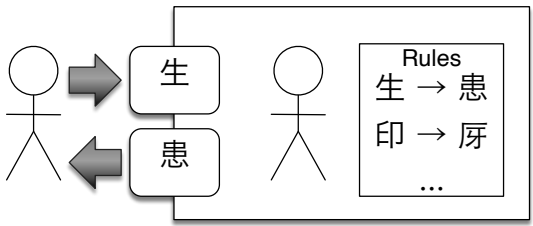
Outline

Can Machines Really Think?

Some philosophers claim that a machine that acts intelligently would not be actually thinking, but would be only a simulation of thinking.

Turing argues the **polite convention** that everyone and machines think.

John Searle rejects the polite convention: The Chinese Room



To continue in the next session.