Coursework due provisionally 28th Feb 4pm

WHAT IS AI ???????????

- Intelligence important (Smarts > Sharts)
- AI relevant to any intellectual task
- Aims to understand and built intelligent entities

Approaches to AI

Thinking like humans, Acting like humans Meaning, not perfect. Measure success against rationality rather than result.

Definitions of AI

Thinking like humans: Machines with minds, automation of activities associated with human thinking, decision making, problem solving, etc. Thinking Rationally Study of mental faculties through computational models, making it possible to perceive, reason, act. Acting like Humans Creating machines that perform functions requiring intelligence when performed by people. Make computers do things which people can be better at. Acting Rationally Study of design of intelligent agents. AI concerned with intelligence behaviour in artefacts.

Acting Like Human: The Turing Test

A test to prove proper operational definition of intelligence; the imitation game. A computer passes the test if a human interrogator, posing written questions, cannot tell if written response came from human or computer No direct interaction between interrogator and computer. Deliberately avoided due to physical simulation unimportant; a test of intelligence.

The Total Turing Test Includes additional apparatus:

- Video Signal - allows interrogator to test perceptual abilities
- Hatch - Interrogator can pass physical objects through a hatch

Acceptance Criteria To pass the Turing Test, the computer would need:

- natural language processing to communicate successfully;
- knowledge representation to store what it knows or hears;
- automated reasoning to use the stored information to answer questions and draw new conclusions;
- machine learning to adapt to new circumstances and to detect and extrapolate patterns; To pass the total Turing Test, the computer would need:
- computer vision to perceive objects;

- robotics to manipulate objects and move about.

These are the 6 principles that encapsulate most of AI

Appraisal

- Anticipated all major arguments against AI in following 50 years
- Suggested major components of AI: knowledge, reasoning, language understanding, learning.
- Remains relevant 60 years later. However,
- AI researchers given little effort to passing Turing Test,
- Believe it is more important to study underlying principles of intelligence rather than duplicate an exemplar.

Thinking Like Human: Cognitive Modelling

Introspection - catching thoughts as they happen Psychological Experiments - observing a person in action Brain Imaging - observing brain in action Sufficiently precise theory of mind can be expressed as a computer program If IO behaviour of program and human matching, evidence that some of the programs mechanisms are mirrored in humans

Cognitive Modelling

- Level of Abstraction? Knowledge or Circuits?
- How to Validate?
  - Cognitive Science - Predicting and testing behaviour of human subjects (top down)
  - Cognitive Neuroscience - Direct identification from neurological data (bottom up)

Both approaches now distinct from AI Both share with AI that the available theories do not explain anything resembling human-level general intelligence.

Thinking Rationally: Laws of Thought

- Syllogisms - patterns for argument structures that always yield correct conclusions when given correct premises.

ex. Socrates is a man, men are mortal, socrates is mortal.

Laws of thought were supposed to govern the operation of the mind. Study gave rise to the field of **logic**, and may have proceeded to the idea of **mechanisation**

Laws of Thought Direct link with mathematics & philosophy, and AI Logicians in the 19th century developed precise notation for statements of objects in the world, and their relations. By the 60s, programs existed that could solve any solvable problem descirbed in logical notation (However, program might

loop forever if no solution exists. Tractability of the program depends on the problem)

Problems Not all intelligent behaviour is mediated by logical deliberation

- Not easy to take informal knowledge and state it in logical notation, especially when uncertain. Can be large discrepancy between in-principle and in-practice
- Even a few hundred parameters can exhaust the computational resources of any computer unless some guidance is given first. Purpose of Thinking?
- What thoughts *should* I have out of all the thoughts I *could* have.

Acting Rationally

Rational Behaviour - doing the correct thing The right thing is that which is expected to maximise goal achievement given available information Doesnt necessarily involve thinking, but thinking should be in the service of rational action.

Aristotle:

- Every art and every enquiry, and similarly every action and pursuit, is thought to aim at some good

Rational Agents An agent is an entity that can perceive and act.

All computer programs do something, but computer agents are expected to operate autonomously, perceive environment, persist over time, adapt to change, and create / pursue goals.

Percept - agents perceptual inputs at any given instance.

Abstractly, an agent is a function from percept histories to actions For any given class of environments and tasks, we seek the agent with the best performance.

Perfect <-> Limited Rationality Perfect rationality always succeeds. This is infeasible in complicated environments, and computational demands are too high.

Limited rationality means acting appropriately when there is not enough time to do all the computations one might wish. Design best program for given machine resources.

Perfect rationality remains a good starting point for theoretical analysis.

Value Alignment Problem A problem with perfect rationality is that it would assume a fully specified objective given to the machine. Artificially defined problems such as chess, come with an objective. Real world problems such as

self-driving cars, become more difficult to specify the objective completely and correctly.

Objective cannot simply be to reach destination safely, since optimal strategy would be to leave the car at home given risks, failures, etc.

Values or objectives given to the machine must align with those of humans Problem of achieving this is called the value alignment problem

Bad Behaviour

If a machine is intelligent enough to reason and act, such a machine may attempt to increase chances of winning by immoral means:

- Blackmail
- Bribery
- Grabbing additional computing resources for itself These behaviours are rational, and are logical for success, however immoral.

Beneficial Machines

We do not want machines that are intelligent in the sense of pursuing their objectives. We want them to pursue our objectives. If we cannot transfer our objectives perfectly to the machine, we need the machine to know that is does not know the complete objective and have the incentive to act cautiously, ask for permission, learn about preferences through observation, defer to human control. We want agents that are provably beneficial to humans.