AI1001 - Assignment 3

Shivram S
ai24btech11031@iith.ac.in

The definition of "intelligence" varies along two dimensions: human vs. rational and thought vs. behaviour, each with different methods. The six main disciplines are natural language processing, knowledge representation, automated reasoning, machine learning, computer vision, and robotics.

- The **Turing Test** (Alan Turing, 1950) was proposed to test the intelligence of a computer. A computer passes the test if a human interrogator cannot tell the computer's responses from those of a human.
- **Cognitive science** brings together AI and psychology. We can learn about human thought through introspection, brain imaging, and psychological experiments. A sufficiently precise theory of the mind allows us to express the theory as a computer program.
- The **logicist tradition** of artificial intelligence builds on programs that can solve any problem written in logical notation (solvers). The uncertainty of the real world is filled using **probability**, allowing rigorous reasoning even with uncertain information.
- Computer agents are programs that percieve their environment and operate autonomously for extended periods. A **rational agent** acts to achieve the best expected outcome.

The rational-agent approach is popular due to its generality and flexibility. The **standard model** involves agents that "do the right thing.". However, the values given to the machine may not match our true values (value alignment problem), and may have negative consequences. For example a chess-playing machine with the sole goal of winning might try to kill the opponent. We want the agent to be cautious when the objective is unknown, and ultimately, be **provably beneficial**.

The foundations of artificial intelligence are drawn from a large number of disciplines:

- **Philosophy**: **Aristotle's syllogisms** for logical reasoning were an attempt to formulate the working of the rational mind. **Dualism** held that these was a part of the mind exempt from physical laws, and **materialism** held otherwise. The **principle of induction** is used to reduce problems, The actions of an agent are governed either by **utilitarianism** (best possible outcome) or by **deontological ethics** (agreement with value system).
- **Mathematics**: **Probability** generalizes logic to uncertain situations. **Statistics** and **decision theory** provide a framework for making decisions. **Game theory** lets us account for the actions of other agents. The notion of **satisficing** - making decisions that are "good enough", gives a better description or actual human behaviour.
- **Neuroscience**: The exact working of the human brain is still a great mystery. Digital computers operate at a clock rate that is a million times faster than a brain but the brain makes up for that with storage and interconnection. The development of **brain-machine interfaces** is promising and may shed light on many aspects of neural systems.
- **Psychology**: **Cognitive physchology** views the brain as an information- processing device. **Behaviourism** looks at the response of systems to stimulus.
- **Computer Engineering**: Faster computers allow for speedup in AI algorithms. **Quantum computing** might offer a large speedup. Specialized hardware such as GPUs, NPUs and wafer scale engine (WSE) have been developed for AI applications.
- Artificial intelligence also borrows **regulatory mechanisms** from **control theory**, and developments in **knowledge representation** from **linguistics**.

Preprint. Under review.