AI Philosophy

Please do not circulate

1 AI Philosophy

The first week of the course was spent on the philosophical aspect of AI. We mainly took a *cognitive* approach to intelligence, in that, we looked at[1] several instance of intelligent behavior in animals (see Table 1).

Experiment	Aspect of Intelligence	$https://youtube.com/$
Crow (using stick to get food)	Problem solving; Counting; Internal States	$watch?v = cbSu2PXOTOc$
Chimpanzee digit identification	Ordering of numbers; Memory	$watch?v = qyJomdyjyvM$
Elephant mirror recognition	Self-Awareness	$watch?v = -EjukzL - bJc$
Chimp, Crow	Reward/Utility optimization	See above
Chimp, Crow, Elephant	Vision	Trivial
Koko - Gorilla	Language	$watch?v = G4QQ8Mfjb_g$

Table 1: Aspects of AI identified in Lecture 1

We also mentioned that Octupus are also intelligent animals, and pointed out the fact that their nervous system is decentralized as opposed to the centralized nervous system found in mammals. We took the discussion forward and identified more aspects of AI in Lecture 2 (see Table 2).

Experiment	Aspect of Intelligence	$https://youtube.com/$
Capuchin monkey	Fairness and morality	$watch?v = -KSryJXDpZo$
Rafael Nadal's analysis of a point	Adversarial decision making use of randomness	$watch?v = 2Cq7sw9O0LA$
Andre Agassi	Prediction and representation	$watch?v = suXcaC61gRk$ $watch?v = 57BMzCM6hQI$
Cat Perception	Representation and algorithmic bias	$watch?v = QzkMo45pcUo$
Crow (water displacement with stones)	Generalization across tasks	$watch?v = ZerUbHmuY04$

Table 2: Aspects of AI identified in Lecture 2

An interesting case of non-example for animal intelligence was Clever Hans ($watch?v = r7850Yl1rbg$): A horse was believed to have the ability to solve numerical puzzles. However, it turned out later that the horse was picking cues/signals form its trainer (even when he was not explicitly giving them). It is still relevant in current day context, i.e., algorithms can pick up (and are supposed to) hidden signals in the training data. Thus, it is a very interesting question to ask 'what is that the algorithms learn when they do manage to learn from data' (interested students can look at articles on interpretable machine learning).

The main objective of looking at the videos is to translate our intuition into keywords. We now expand them as follows:

- Problem solving is part of "Good Old-Fashioned AI" (GOFAI), and is part of our course, where we cover search techniques leading to the A^* algorithm.

[1] Mostly youtube videos, the snippets of which have been uploaded in Moodle.

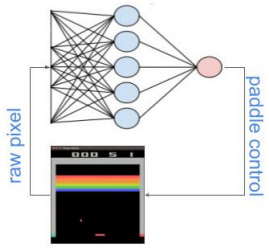

Figure 1: Shows the architecture of AI playing Atari-Breakout game. Here, the input is the vision, i.e., image in the form of raw-pixels. The deep neural network learns to convert this image input into actions, i.e., control of the paddle.

- Memory relates to *space-complexity* of any algorithm.
- Reward/Utility maximization is related to *optimization* theory. Most problems in machine learning and reinforcement learning are formulated as either maximization of a reward objective or minimization of a cost objective.
- Vision is at the heart of current state of art deep reinforcement learning algorithms that have been successful in tasks such as AlphaGo, Atari gaming etc. The fascinating fact about such vision driven systems is that they can convert input in the form of raw pixels to real time decision making (see Figure 1).
- Fairness is an important sub-topic of AI that has been gather the attention of researchers in the recent past, and will be a huge factor in the future as AI systems begin to make their way into society. While AI systems are good at learning the hidden dependencies, it also important for us to design them in a manner that the decisions do not favour or adversely affect a particular group or person.
- Prediction is a very key sub-problem in machine learning and will continue to do so in the future.
- Representation is a key in successful prediction. The recent success of deep learning has been attributed to the fact that the hidden layers of the deep neural network are able to learn the right representation. Automatic learning of representations removes the need for designing hand-crafted features, an approach heavily dependent on domain knowledge and prone to trial and error. Further, an advantage of such automatically learned representation is that all that we need is the input data and the AI can take care of the rest.
- AI systems have been successful in solving several 'hard'[2] problems. This is possible because i) in most cases AI systems only find approximate but near-optimal solutions, ii) use the right architecture and initialization that enables to the algorithm to search in useful regions of the solution space.
- Generalization of learning algorithms has been a very important theme in theory as well as practice. Loosely speaking, the notion of generalization distinguishes learning from memorization or 'hard-wired' behaviour. Statistical learning theory address this question as follows: Given *training* data from an unknown distribution, how to learn a model so that it performs well on *test* (or unseen) data.

1.1 What is intelligence?

When can we say that machines are intelligent? Historically, several philosophical positions have been taken, a brief description of them is given below:

Think Humanly: The idea is to replicate human like *thinking* in AI systems. This idea dates back to Simons and Newell's "General Problem Solver", wherein, in addition to problem solving, tracing the steps of reasoning and comparing it to human like thinking was also given importance.

Act Humanly: Alan Turing proposed a test of intelligent behavior; it should not be possible for a human to chat with a computer and tell it apart from chatting to a different human being. This is

[2]The complexity theoretic notion of hardness such as NP-hardness.

famously known as the Turing's test. Starting with Eliza, several 'chatbots' have been designed to converse with humans.

Strong AI vs Weak AI: The philosophical position that AI is purely computational is known as strong AI. This position was challenged by John Searle's Chinese Room experiment: say a computer can read input in Chinese and produce output in Chinese. However, the intermediary steps are logical, which when specified could be executed as well by a human being (here John Searle himself) who has know knowledge of Chinese, and hence has no intelligence in Chinese. Thus acting humanly does not imply intelligence

Further, either human like action or thinking also bring us to the question of whether or not AI has to replicate the inefficiencies in the human behaviour, and whether or not such replication is even possible given the completely different hardware (humans hardware is bio-chemical, whilst computer hardware is electronic).

Think Rationally: is an idea related to the logicist tradition which holds the position that given set of facts about the objects of the world one can make right inferences. For example, given the facts i) humans have two legs ii) we are human, we can conclude that we have two legs. An important issue with this approach is scalability, i.e., it becomes increasingly tedious to arrive at the right conclusions even if we have only a few hundred facts.

Act Rationally: A very 'practical' approach to AI, where the position is that AI systems are nothing but 'smart' programs. Most of the course will deal with what we call *rational agents* (see handout2.pdf).

The following fields of study can be said to be foundational to AI:

- Optimization; Probability; Theoretical Computer Science; Neuroscience; Logic; Game Theory; Economics; Control Theory; Linguistics; Robotics; Computer Vision

We also concluded with some of the successes and also some words of caution. We list them below:

Success: Imagenet, AlphaGo, Automatic Driving, Google Duplex.

Words of Caution: Jeopardy challenge of IBM-Watson ($watch?v = WFR3lOm_xhE$), Adversarial Examples of deep learning ($watch?v = oeQW5qdeyy8$).