28 November 2013, 2.10pm AEST, The Conversation
http://theconversation.com/to-create-a-super-intelligent-machine-start-with-an-equation-20756

To create a super-intelligent machine, start with an equation

Where do you start building a machine like Deep Thought, the super-intelligent computer who came up with the Answer to The Ultimate Question of Life, the Universe, and Everything? YouTube

Intelligence is a very difficult concept and, until recently, no one has succeeded in giving it a satisfactory formal definition.

Most researchers have given up grappling with the notion of intelligence in full generality, and instead focus on related but more limited concepts – but I argue that mathematically defining intelligence is not only possible, but crucial to understanding and developing super-intelligent machines.

From this, my research group has even successfully developed software that can learn to play Pac-Man from scratch.

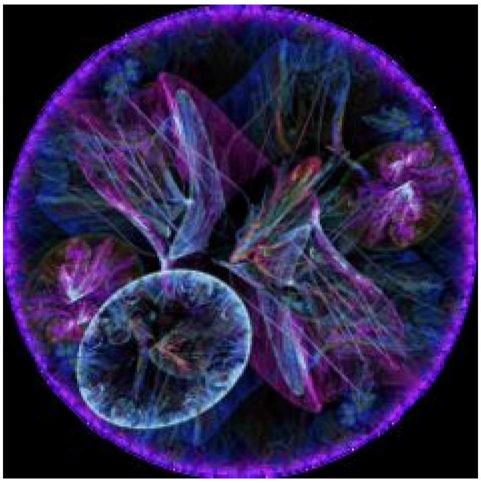
Let me explain – but first, we need to define "intelligence".

So what *is* intelligence?

I have worked on the question of general rational intelligence for many years. My group has sifted through the psychology, philosophy and artificial intelligence literature and searched for definitions individual researchers and groups came up with.

The characterisations are very diverse, but there seems to be a recurrent theme which we have aggregated and distilled into the following definition:

Shane Legg



Intelligence is an agent’s ability to achieve goals or succeed in a wide range of environments.

You may be surprised or sceptical and ask how this, or any other single sentence, can capture the complexity of intelligence. There are two answers to this question:

1. Other aspects of intelligence are implicit in this definition: if I want to succeed in a complex world or achieve difficult goals, I need to acquire new knowledge, learn, reason logically and inductively, generalise, recognise patterns, plan, have conversations, survive, and most other traits usually associated with intelligence.
2. The challenge is to transform this verbal definition consisting of just a couple of words into meaningful equations and analyse them.

This is what I have been working on in the past 15 years. In the words of American mathematician Clifford A. Truesdell:

There is nothing that can be said by mathematical symbols and relations which cannot also be said by words. The converse, however, is false. Much that can be and is said by words cannot be put into equations – because it is nonsense.

Indeed, I actually first developed the equations and later we converted them into English.

Universal artificial intelligence

This scientific field is called universal artificial intelligence, with AIXI being the resulting super-intelligent agent.

The following equation formalises the informal definition of intelligence, namely an agent’s ability to succeed or achieve goals in a wide range of environments:

$$\text{AIXI} \qquad a_k := \arg\max_{a_k} \sum_{o_k r_k} \ldots \max_{a_m} \sum_{o_m r_m} [r_k + \ldots + r_m] \sum_{q:U(q,a_1..a_m)=o_1 r_1..o_m r_m} 2^{-\ell(q)}$$

Explaining every single part of the equation would constitute a whole other article (or book!), but the intuition behind it is as follows: AIXI has a planning component and a learning component.

Imagine a robot walking around in the environment. Initially it has little or no knowledge about the world, but acquires information from the world from its sensors and constructs an approximate model of how the world works.

It does that using very powerful general theories on how to learn a model from data from arbitrarily complex situations. This theory is rooted in algorithmic information theory, where the basic idea is to search for the simplest model which describes your data.

The model is not perfect but is continuously updated. New observations allow AIXI to improve its world model, which over time gets better and better. This is the learning component.

M Hutter

AIXI now uses this model for approximately predicting the future and bases its decisions on these tentative forecasts. AIXI contemplates possible future behaviour: "If I do this action, followed by that action, etc, this or that will (un)likely happen, which could be good or bad. And if I do this other action sequence, it may be better or worse."

The "only" thing AIXI has to do is to take from among the contemplated future action sequences the best according to the learnt model, where "good/bad/best" refers to the goal-seeking or succeeding part of the definition: AIXI gets occasional rewards, which could come from a (human) teacher, be built in (such as high/low battery level is good/bad, finding water on Mars is good, tumbling over is bad) or from universal goals such as seeking new knowledge.

The goal of AIXI is to maximise its reward over its lifetime – that's the planning part.

In summary, every interaction cycle consists of observation, learning, prediction, planning, decision, action and reward, followed by the next cycle.

If you're interested in exploring further, AIXI integrates numerous philosophical, computational and statistical principles:

- Ockham's razor (simplicity) principle for model selection
- Epicurus principle of multiple explanations as a justification of model averaging
- Bayes rule for updating beliefs
- Turing machines as universal description language
- Kolmogorov complexity to quantify simplicity
- Solomonoff's universal prior and
- Bellman equations for sequential decision making.

Theory and practice of universal artificial intelligence

The above equation rigorously and uniquely defines a super-intelligent agent that learns to act optimally in arbitrary unknown environments. One can prove amazing properties of this agent – in fact, one can prove that in a certain sense AIXI is the most intelligent system possible.

Note that this is a rather coarse translation and aggregation of the mathematical theorems into words, but that is the essence.

Since AIXI is incomputable, it has to be approximated in practice. In recent years, we have developed various approximations, ranging from provably optimal to practically feasible algorithms.

At the moment we are at a toy stage: the approximation can learn to play Pac-Man, TicTacToe, Kuhn Poker and some other games.

Watch AIXI play Pac-Man

The point is not that AIXI is able to play these games (they are not hard) – the remarkable fact is that a single agent can learn autonomously this wide variety of environments.

AIXI is given no prior knowledge about these games; it is not even told the rules of the games!

It starts as a blank canvas, and just by interacting with these environments, it figures out what is going on and learns how to behave well. This is the really impressive feature of AIXI and its main difference to most other projects.

Even though IBM Deep Blue plays better chess than human Grand Masters, it was specifically designed to do so and cannot play Jeopardy. Conversely, IBM Watson beats humans in Jeopardy but cannot play chess – not even TicTacToe or Pac-Man.

IBM Watson in action. Anirudh Koul

AIXI is not tailored to any particular application. If you interface it with any problem, it will learn to act well and indeed optimally.

The current approximations are, of course, very limited. For the learning component we use standard file compression algorithms (learning and compression are closely related problems). For the planning component we use standard Monte Carlo (random search) algorithms.

Neither component has any particular built-in domain knowledge (such as the Pac-Man board or TicTacToe rules).

Of course you have to interface AIXI with the game so that it can observe the board or screen and act on it, and you have to reward it for winning TicTacToe or eating a food pellet in Pac-Man … but everything else AIXI figures out by itself.

This article is adapted from a presentation which will be delivered at the Science, Technology and the Future conference, November 30 and December 1 2013.

Discussion

Statistics