

Machine Super Intelligence

Shane Legg

Gatsby Computational Neuroscience Unit
University College London

Halloween 2009

Overview

- “intelligence” is not a 4 letter word
- the ultimate predictor of Solomonoff
- upgrade to Hutter’s AIXI
- making it real: Monte Carlo AIXI
- a formal definition of machine intelligence
- the brain is not (quite) a black box
- early 2020’s: the Halloween scenario

Where to find the details for the first half of the talk

Machine Super Intelligence

by Shane Legg

See my website for free pdf, or hardcover for \$17 (printing cost):

`www.vetta.org`

If you want all the mathematical details, for \$70 from Amazon:

Universal Artificial Intelligence

by Marcus Hutter

Three main attitudes to the I in AI

“I’m not really working on intelligence, what ever that is.”

– many people working in the area called “AI”

“My reaction to intelligence is the same as my reaction to pornography, I can’t define it, but I like it when I see it.”

– Hugh Loebner

“...we need a definition of intelligence that is applicable to machines as well as humans or even dogs... Then it will be possible to assess whether progress is being made...”

– W. L. Johnson

What is intelligence?

After reviewing over 70 definitions of intelligence I identified the following commonly occurring features:

- intelligence is a property of an active **agent**
- the agent interacts with an **environment**
- intelligence is a **matter of degree**
- intelligence is related to the agent's **success** in achieving goals
- the environment is not fully known to the agent and so the agent must be **adaptable** to many different environments

An Informal Definition of Intelligence

Putting these features together in a very general way gives me my informal definition of intelligence:

Intelligence measures an agent's ability to achieve goals in a wide range of environments.

Now let's look at a few definitions given by others to see how my definition compares...

Definitions of Intelligence from Psychologists

“We shall use the term ‘intelligence’ to mean the ability of an organism to solve new problems ... “ – **Bingham**

“Intelligence is part of the internal environment that shows through at the interface between person and external environment as a function of cognitive task demands.” – **Snow**

“A person possesses intelligence insofar as he has learned, or can learn, to adjust himself to his environment.” – **Colvin**

“A global concept that involves an individual’s ability to act purposefully, think rationally, and deal effectively with the environment. – **Wechsler**

“... a cluster of cognitive abilities that lead to successful adaptation to a wide range of environments.” – **Simonton**

Definitions of Intelligence from AI Researchers

“Intelligent systems are expected to work, and work well, in many different environments. Their property of intelligence allows them to maximise the probability of success...” – **Gudwin**

“Doing well in a broad range of tasks is an empirical definition of ‘intelligence’ ” – **Masum et al.**

“... the ability of a system to act appropriately in an uncertain environment, where appropriate action is that which increases the probability of success, and success is the achievement of behavioural subgoals that support the system’s ultimate goal.” – **Albus**

“Any system . . . that generates adaptive behaviour to meet goals in a range of environments can be said to be intelligent.” – **Fogel**

An Informal Definition of Intelligence

Once again, my informal definition of intelligence:

Intelligence measures an agent's ability to achieve goals in a wide range of environments.

At present no *practical* machine has enough generality to have much intelligence according to the above definition.

Suggestion: Make the problem easier by ignoring computational cost, and then try to define a *theoretical* machine that is intelligent.

It turns out that this is possible...

The inductive inference problem

You observe the sequence of numbers:

1, 3, 5, 7, ?

Why 9?

You have used *Occam's razor* and assumed that the simplest explanation for the sequence was the most likely: $2n - 1$

Actually, the next number is 57 because the sequences is being generated by $2n - 1 + 2(n - 1)(n - 2)(n - 3)(n - 4)$.

Inductive Inference: Epicurus, Occam and Bayes

Epicurus' principle:

Keep all hypothesis that are consistent with the data.

Occam's razor:

*Among all hypotheses consistent with the data,
the simplest is the most likely.*

Bayes' Rule:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Epicurus says: $P(h) > 0$ for all h , and h belongs to a very large set

Occam says: $P(h)$ depends on the complexity of h

Kolmogorov complexity

Represent an hypothesis h by the probability distribution ν and define the **Kolmogorov complexity** of ν :

$$K(\nu) = \text{"length of the shortest program that computes } \nu \text{"}$$

Intuition: simple hypotheses have short programs
while complex hypotheses only have long programs.

$2n - 1 \Rightarrow$ has short program \Rightarrow low complexity

$2n - 1 + 2(n - 1)(n - 2)(n - 3)(n - 4)$
 \Rightarrow needs longer program \Rightarrow higher complexity

Solomonoff's Universal prior distribution

The **universal prior probability** of hypothesis ν is,

$$P(\nu) := 2^{-K(\nu)}$$

This prior respects Epicurus' rule as it assigns a positive probability to all hypotheses consistent with the observed data, and is defined for all probability distributions.

It also formalises Occam's razor as simpler hypotheses are given higher probability than complex ones.

In terms of some data observed x the **universal prior probability** is:

$$\xi(x) := \sum_{\nu} P(\nu) \nu(x)$$

Solomonoff induction: prediction using ξ

Let sequence $\omega = 010110\dots$ be from an unknown distribution μ .

Having observed $\omega_{1:n}$ try to predict ω_{n+1} :

$$\xi(\omega_{n+1} = 0 | \omega_{1:n}) = \frac{\xi(\omega_{1:n}0)}{\xi(\omega_{1:n})}$$

The key result:

For *any* unknown computable μ , the expected **total** prediction error over the infinite length of ω is upper bounded by $\frac{\ln 2}{2} K(\mu)$.

The catch:

Only works in theory as ξ is not computable.

Approximating Solomonoff Induction

By making various simplifying assumptions that make Solomonoff induction computable we can derive standard statistical methods:

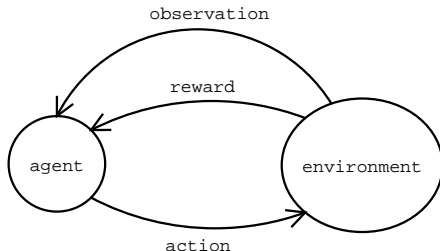
- maximum a posteriori estimation
- minimum message length estimation
- minimum description length estimation
- maximum entropy estimation

One view of Solomonoff Induction is that it's a model of ideal inductive inference which practical statistics approximates.

For the details see [*An Introduction to Kolmogorov Complexity and Its Applications*](#) by Li and Vitányi.

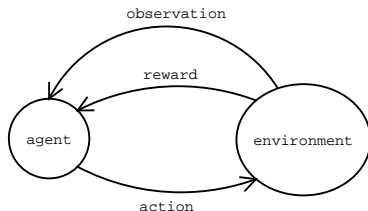
The Reinforcement Learning Framework

A very general framework in artificial intelligence for an agent interacting with an environment is **reinforcement learning**



The agent tries to choose its actions so as to receive as much reward as possible from the environment.

Formalising the Agent and the Environment



The process of interaction produces an increasing history of observations, rewards and actions,

$$a_1 o_1 r_1 a_2 o_2 r_2 a_3 o_3 r_3 \dots a_t o_t r_t = aor_{1:t}$$

Formalising the Agent and the Environment

The **agent** is a probability measure over actions conditioned on the history,

$$\pi(a_k \mid a_1 o_1 r_1 a_2 o_2 r_2 \dots o_k r_k) = \pi(a_k \mid aor_{1:k})$$

The **environment** is a probability measure over observations and rewards conditioned on the history,

$$\mu(o_k r_k \mid a_1 o_1 r_1 a_2 o_2 \dots a_{k-1}) = \mu(or_k \mid aor_{1:k} a_{k-1})$$

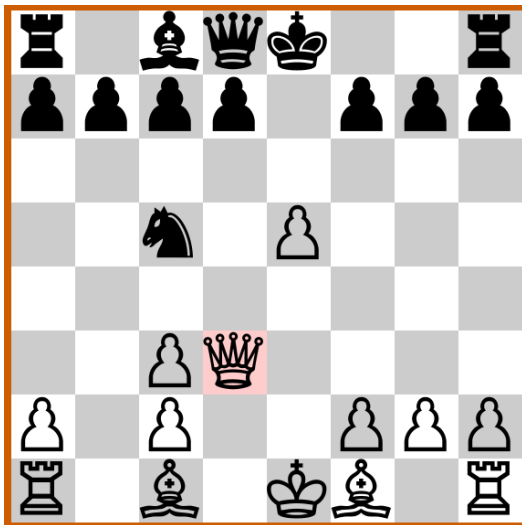
The agent's goal is to maximise the **expected total reward**:

$$V_{\mu}^{\pi} = \mathbf{E}(r_1 + r_2 + r_3 + \dots)$$

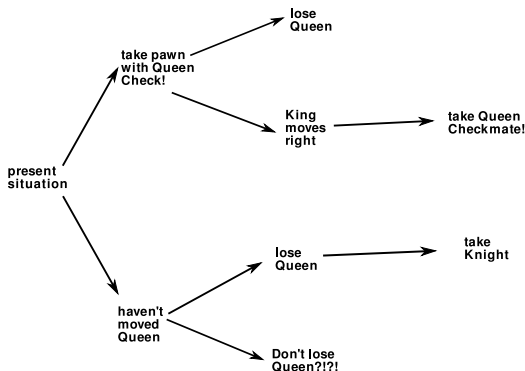
Reinforcement Learning is Extremely General

- Playing chess
- Navigating a robot through a maze
- Answering questions on an IQ test
- Passing a Turing test
- Writing an award winning book
- etc...

Maximising total reward: Chess



Maximising total reward: Chess



Equation for the optimal behaviour:

$$\operatorname{argmax}_{a_t} \lim_{m \rightarrow \infty} \sum_{or_t} \max_{a_{t+1}} \sum_{or_{t+1}} \cdots \max_{a_m} \sum_{or_m} [r_t + \cdots + r_m] \mu(or_{t:m} | a_{or_{<t}} a_{t:m})$$

Hutter's universal agent: AIXI

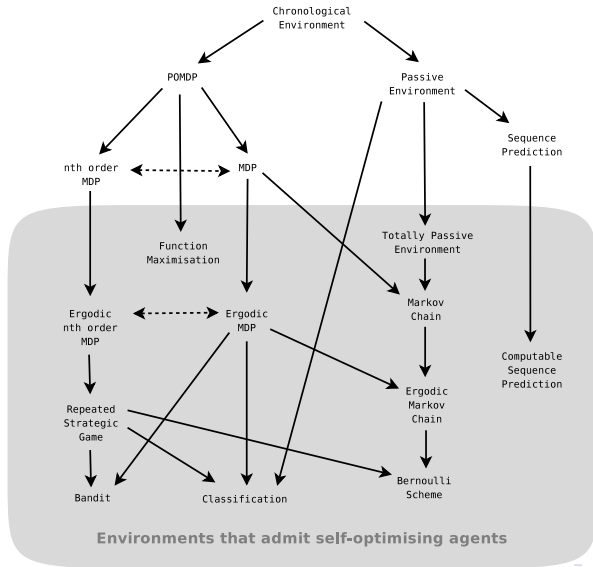
Hutter's idea: replace the unknown μ with Solomonoff's ξ .

With a history of $or_{t:m}$, the next action selected by AIXI is:

$$\operatorname{argmax}_{a_t} \lim_{m \rightarrow \infty} \sum_{or_t} \max_{a_{t+1}} \sum_{or_{t+1}} \cdots \max_{a_m} \sum_{or_m} [r_t + \cdots + r_m] \xi(or_{t:m} | a_{<t} or_{<t} a_{t:m})$$

Key result: Hutter proved that AIXI converges to optimal in *any* environment where this is possible for a general agent.

Environments in which AIXI converges to optimal



Approximating AIXI: Monte Carlo AIXI

Equation for AIXI:

$$\operatorname{argmax}_{a_t} \lim_{m \rightarrow \infty} \sum_{or_t} \max_{a_{t+1}} \sum_{or_{t+1}} \cdots \max_{a_m} \sum_{or_m} [r_t + \cdots + r_m] \xi(or_{t:m} | a_{or < t} a_{t:m})$$

In Monte Carlo AIXI:

Expecti-max tree search \rightarrow Monte Carlo tree search

Solomonoff predictor $\xi \rightarrow$ Context tree weighting predictor
+ Complexity weighting

(for details see Veness, Ng, Hutter and Silver, 2009)

Is MC-AIXI intelligent?

Currently it can learn to solve/play:

- simple prediction problems
- Tic-Tac-Toe
- Paper-Scissors-Rock
- mazes where it can only see locally
- various types of Tiger games
- simple computer games, e.g. Pac-Man

The future of MC-AIXI

“Only a small community has concentrated on general intelligence. No one has tried to make a thinking machine and then teach it chess ... The bottom line is that we really haven’t progressed too far toward a truly intelligent machine.” – Marvin Minsky

Currently, they are trying to teach MC-AIXI to play checkers.

MC-AIXI has a number of interesting properties:

- embarrassingly parallel algorithm
- anytime algorithm
- not too difficult to try new things with the compressor to learn deeper patterns in the environment

Universal Intelligence: A formal definition of intelligence

Return again to our informal definition of intelligence:

Intelligence measures an agent's ability to achieve goals in a wide range of environments.

In constructing AIXI we have already formalised agents, environments, how to measure an agent's success and so on.

Applying this formalism to our informal definition of intelligence yields the **universal intelligence measure**:

$$\Upsilon(\pi) := \sum_{\mu \in E} 2^{-K(\mu)} V_{\mu}^{\pi}$$

Is Universal Intelligence any good?

Is it “intelligence”? By construction it measures the general ability of an agent to succeed in a wide range of environments, capturing the essence of many informal definitions.

Very general The definition places no restrictions on the internal workings of the agent; it only requires that the agent is capable of generating output and receiving input containing a reward signal.

Agents with high universal intelligence are very powerful

The maximal agent with respect Υ is the universal agent AIXI which has been proven to have powerful optimality properties.

Is Universal Intelligence any good?

Practically meaningful A high value of universal intelligence would imply that an agent was able to learn to perform well in a wide range of environments. Such a machine would obviously be of large practical significance.

Non-anthropocentric Universal intelligence is based on fundamentals of information and computation theory. In contrast, other tests such as the Turing test are largely a measure of a machine's “humanness”, rather than its intelligence.

Formal definition Universal intelligence can be discussed and studied in a precise way — unlike some informal notions of intelligence that take pages of text to describe and depend on equally hard to define concepts such as “creativity”, “understanding”, “wisdom”, “consciousness” and so on.

Can complexity theory tests be made practical?

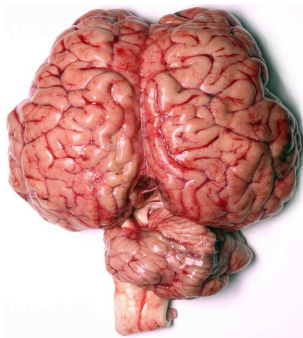
Matt Mahoney has been sampling from an approximation to ξ and using this to benchmark compressors.

J. Hernández-Orallo has developed a complexity based sequence prediction test by using Levin complexity to approximate Kolmogorov complexity.

Complexity	Sequence	Answer
9	a, d, g, j, - , ...	m
12	a, a, z, c, y, e, x, - , ...	g
14	c, a, b, d, b, c, c, e, c, d, - , ...	d

Ben Goertzel is about to publish a paper which describes a number of practical measures of machine intelligence which use our formal definition of intelligence as a foundation.

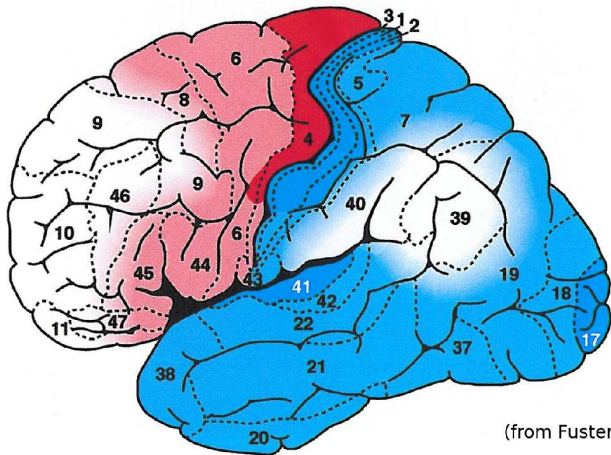
Does the brain tell us anything useful about AGI design?



\neq

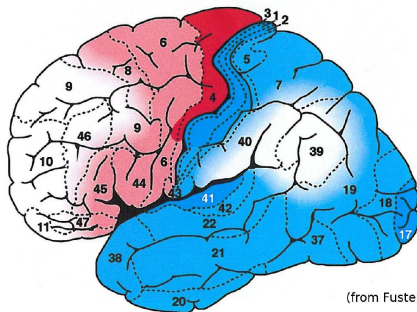
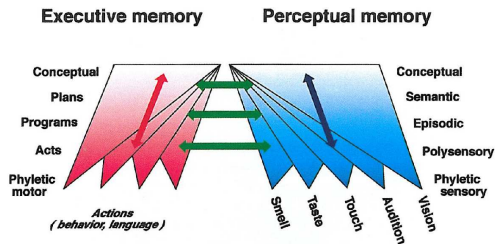


A high level view of the cerebral cortex



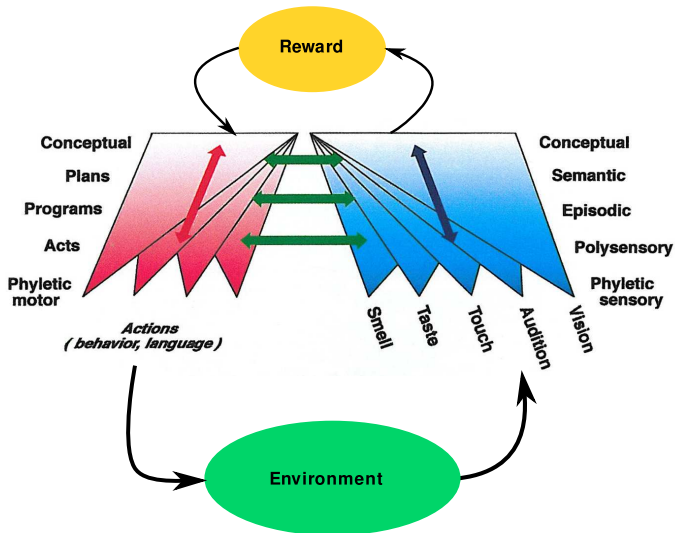
(from Fuster 2003)

A high level view of the cerebral cortex

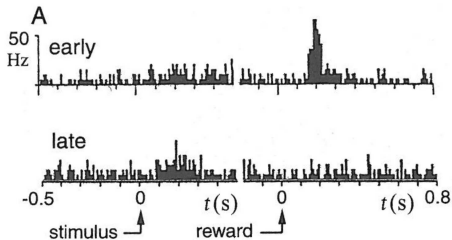


(from Fuster 2003)

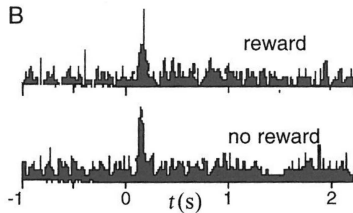
A high level view of the brain



Temporal difference learning



(from Mirenowicz and Schultz, 1994)



(from Schultz, 1998)

We have since learnt a lot more about RL in the brain

Negative values in a spiking neural system are problematic.

The habenula computes the negative prediction error.
(recent Bromberg-Martin)

Does the brain use model-based or model-free RL?

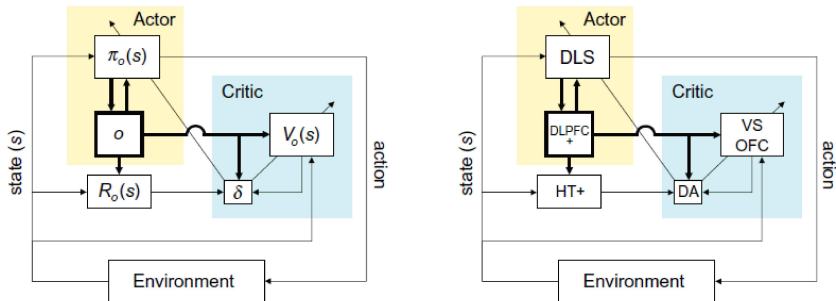
It has both and intelligently switches. (Daw, Niv and Dayan, 2005)

Pseudo rewards are useful to direct behaviour that isn't strictly rewarding, e.g. informative cues. What does the brain do?

The brain is using this trick too. (recent Bromberg-Martin)

How does the brain deal with complex temporal sequences?

Hierarchical reinforcement learning in the brain



(from Botvinick, Niv and Barto, 2008)

Where is brain RL going?

This area of research is currently progressing *very quickly*.

New genetically modified mice allow researchers to precisely turn on and off different parts of the brain's RL system in order to identify the functional roles of the parts.

I've asked a number of researchers in this area:

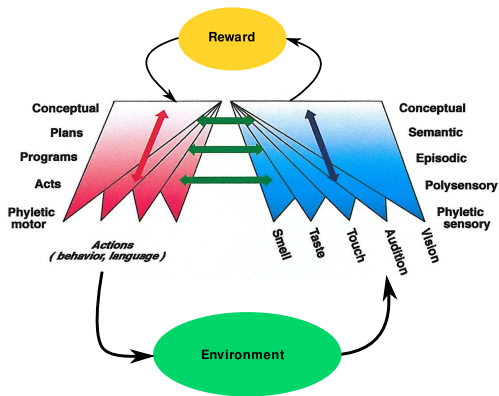
"Will we have a good understanding of the RL system in the brain before 2020?"

Typical answer:

"Oh, we should understand it well before then. Indeed, we have a decent outline of the system already."

If we can do the RL, why can't we build an AI?

The main problem with applying RL to real problems is scaling the system to deal with very large action and perception spaces.



In the brain the cortical hierarchy seems to solve this problem.

Deep belief networks and the cortical hierarchy

A lot of progress in last 10 years, in particular with restricted Boltzmann machines (RBMs).

- Can train deep hierarchical networks
- Increasing abstract in each level
- Local learning rule
- Generative model: can recognise and generate behaviour
- Able to do multiple constraint satisfaction and filling in
- Can be made temporal
- Can even be implemented with spiking neural networks

Clearly not the same as cortex, but they appear to be in the same general class of learning algorithms.

Restricted Boltzmann Machines in action

Digits video is from (Hinton, Osindero and Teh, 2006)

Walking videos are from (Taylor and Hinton, 2009)

Is building an intelligent machine a good idea?

- if we can build human level, we can almost certainly scale up to well above human level
- a machine well above human level will understand its own design and be able to design even more powerful machines...
- we have almost no idea how to deal with this

A group called the Singularity Institute for Artificial Intelligence is studying this problem.

One of their core goals is to develop a “Friendly AI” that has a very high level of safety for humanity no matter how super intelligent it may become.

A vision of the early 2020's: the Halloween Scenario

- 1 Petaflops desktops
- 2 Powerful algorithms for deep belief networks?
- 3 Brain reinforcement learning fairly well understood

⇒ many groups working on brain-like AGI architectures

⇒ success leads to access to exaflops supercomputers

No practical theory of Friendly AI