A Formal Measure of Machine Intelligence

Shane Legg and Marcus Hutter

Dalle Molle Institute for Artificial Intelligence
Manno-Lugano
Switzerland

Benelearn 2006, Gent, Belgium

What is intelligence?!?!

While defining human intelligence is difficult, for machines with senses, environments, motivations and cognitive capacities which are very different to our own — it seems to be impossible.

> *How can we hope to create "artificial intelligence"*
> *if we can't even say what intelligence is?!?!*

A good place to start is by looking at well known definitions of intelligence that have been given by psychologists... what we find is that they have many similarities.

Some well known Definitions of Intelligence

"The capacity to learn or to profit by experience." – *Dearborn*

"Ability to adapt oneself adequately to relatively new situations in life." – *Pinter*

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

"We shall use the term 'intelligence' to mean the ability of an organism to solve new problems. . . ." – *Bingham*

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

Common Features in Definitions of Intelligence

From these definitions (and many similar ones), the following key elements are apparent:

- intelligence is a property of an *individual*
- intelligence is a *matter of degree*
- the individual interacts with an *environment*
- intelligence is related to the individual's *success*
- the environment is not fully known to the individual and so the agent must be *adaptable* and *learn from experience*

An Informal Definition of Intelligence

Combining these gives us the following definition:

> *Intelligence measures an individual's general ability to succeed in a range of environments.*

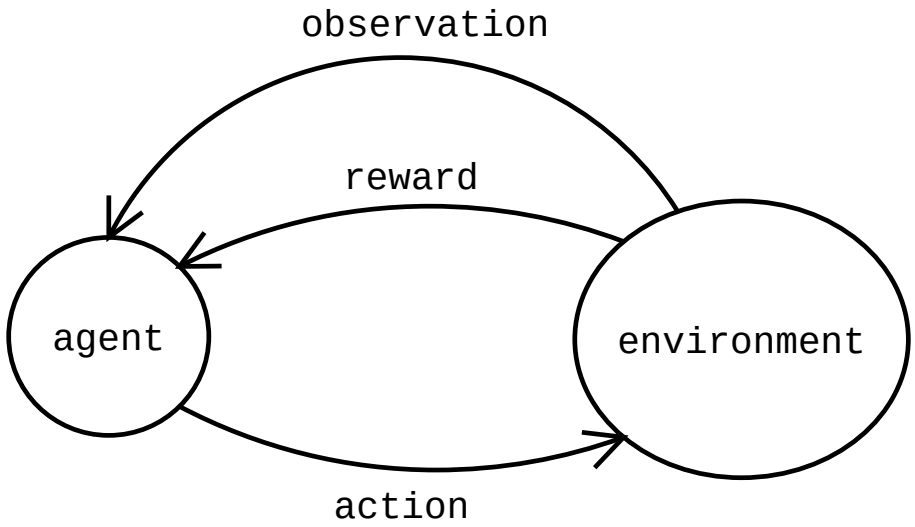
This captures the essence of many definitions of intelligence.

However the definition is still *informal*.

... what we would like is a *formal* definition for *arbitrary systems*.

The Reinforcement Learning Framework

The most 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.

Reinforcement Learning is Extremely General

- Playing chess
- Navigating a robot through a maze
- Making money on the stock market
- Answering questions on an IQ test
- Passing a Turing test
- etc...

Formalising the Agent and the Environment

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

$$o_1\ r_1\ a_1\ o_2\ r_2\ a_2\ o_3\ r_3\ a_3\ o_4 \ldots$$

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

$$\pi(a_k \mid o_1\ r_1\ a_1\ o_2\ r_2 \ldots o_k\ r_k)$$

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

$$\mu(o_k\ r_k \mid o_1\ r_1\ a_1\ o_2\ r_2 \ldots a_{k-1})$$

Formalising the Success of an Agent in an Environment

The agent tries to maximise the total reward. What this means depends on how we value reward at different points in time.

A simple solution is to require that the total reward received from an environment is bounded.

Thus the *future reward*,

$$V_\mu^\pi := \mathbf{E}\left(\sum_{i=1}^{\infty} r_i\right) \leq 1,$$

where the expected value is taken over all possible interaction histories of π and μ.

The Space of Possible Environments

As we desire a very general definition of intelligence our space of environments should be as large as possible.

An obvious choice is the space of *all probability measures*. However this causes serious problems as we cannot even describe some of these measures in a finite way.

The solution is to use *all computable probability measures*. This allows for an infinite space of environments with no upper bound on their complexity. Clearly environments can be stochastic.

This space, denoted E, appears to be the largest useful space of environments.

Defining a General Measure of Performance

We want to compute the *general performance* of an agent.
As there are an infinite number of environments in E,
we cannot simply take a uniform distribution over them.

If we consider the agent's perspective, this is the same as asking:
Given several different hypotheses which are consistent with the data, which hypothesis should be considered the most likely?

This is a standard problem in inductive inference for which the usual solution is to invoke *Occam's razor*

> *Given multiple hypotheses which are consistent with the data, the simplest should be preferred.*

Thus we should test agents in such a way that they are on average rewarded for considering simpler environments to be more likely.

How to Measure the Complexity of an Environment?

As each environment is described by a computable measure, we can use standard Kolmogorov complexity.

If $\mathcal{U}$ is a prefix universal Turing machine then the *Kolmogorov complexity* of an environment μ is the length of the shortest program on $\mathcal{U}$ that computes μ,

$$K(\mu) := \min_p \{l(p) : \mathcal{U}(p) = \mu\}.$$

It can be shown that K depends on $\mathcal{U}$ only up to a small constant that is independent of p.

As each program p is a binary string from a prefix-free set, a natural way to express the probability of μ is $2^{-K(\mu)}$.

Formal Definition of Intelligence

The *universal intelligence* of an agent π is thus,

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

Compare to our informal definition,

> *Intelligence measures an individual's general ability to succeed in a range of environments.*

Is Universal Intelligence any good?

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

Incorporates Occam's razor In this respect it is similar to intelligence tests for humans which usually define the "correct" answer to a question to be the simplest consistent with the given information.

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.

Is Universal Intelligence any good?

Correctly orders simple adaptive agents By considering V_μ^π for a number of basic environments, such as small MDPs, and agents with simple optimisation strategies, it can be shown that Υ orders the agents' intelligence in a natural way.

Agents with high universal intelligence are extremely powerful The maximal agent with respect Υ is the universal agent AIXI. AIXI has been proven to have powerful optimality properties, including Pareto optimality and the ability to be self-optimising in all environments in which this is possible for a general agent.

Thus Υ spans very low intelligence up to super intelligence.

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.

Simple and intuitive 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.

What next?

How bad is the dependency on $\mathcal{U}$? Can this be helped?

Can Υ be used to create practical tests of machine intelligence? There is a related complexity based test for sequence prediction was used to create a usable IQ test with some encouraging results. We should be able to do something similar.

Get more feed back. A poster on this was presented at IJCAI, a short article appeared in New Scientist magazine, a 3 page article in a neuroscience magazine, and the paper gets down loaded several times a day... plenty of curiosity, but little real feed back.

If we want to create intelligent machines,
then clarifying what this means would be a good start!