A Formal Definition of Intelligence for Artificial Systems

Shane Legg and Marcus Hutter
IDSIA, Galleria 2, Manno-Lugano 6928, Switzerland
{shane,marcus}@idsia.ch

A fundamental difficulty in artificial intelligence is that nobody really knows what intelligence is, especially for systems with senses, environments, motivations and cognitive capacities which are very different to our own. In our work we take a mainstream informal perspective on intelligence and formalise and generalise this using the reinforcement learning framework and algorithmic complexity theory. The resulting formal definition of intelligence has many interesting properties and has received attention in both the academic [4, 5] and popular press [2, 1].

Although there is no strict consensus among experts over the definition of intelligence for humans, most definitions share many key features. In all cases, intelligence is a property of an entity, which we will call the *agent*, that interacts with an external problem or situation, which we will call the *environment*. An agent's intelligence is typically related to its ability to succeed with respect to one or more objectives, which we will call the *goal*. The emphasis on learning, adaptation and flexibility common to many definitions implies that the environment is not fully known to the agent. Thus true intelligence requires the ability to deal with a wide range of possibilities, not just a few specific situations. Putting these things together gives us our informal definition: *Intelligence measures an agent's general ability to achieve goals in a wide range of environments.* We are confident that this definition captures the essence of many common perspectives on intelligence. It also describes what we would like to achieve in machines: A very general capacity to adapt and perform well in a wide range of situations.

To formalise this we combine the extremely flexible reinforcement learning framework with algorithmic complexity theory. In reinforcement learning the agent sends its *actions* to the environment and receives *observations* and *rewards* back. The agent tries to maximise the amount of reward it receives by learning about the structure of the environment and the goals it needs to accomplish in order to receive rewards. To denote symbols being sent we will use the lower case variable names o, r and a for observations, rewards and actions respectively. 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 simply a function, denoted by π, which is a probability measure over actions conditioned on the current history, for example, $\pi(a_3|o_1 r_1 a_1 o_2 r_2)$. How the agent generates this distribution over actions is left completely open, for example, agents are not required to be Turing computable.

The environment, denoted μ, is similarly defined: $\forall k \in \mathbb{N}$ the probability of $o_k r_k$, given the current history is $\mu(o_k r_k|o_1 r_1 a_1 o_2 r_2 a_2 \ldots o_{k-1} r_{k-1} a_{k-1})$. As we desire an extremely general definition of intelligence for arbitrary systems, 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 require the measures to be computable. This allows for an infinite space of possible environments with no bound on their complexity. It also permits environments which are non-deterministic as it is only their probability distributions which need to be computable. Additionally we bound the total reward to be 1 to ensure that the future value $V_\mu^\pi := \mathbf{E} \sum_{i=1}^{\infty} r_i$ is finite. This space, denoted E, appears to be the largest useful space of environments.

We want to compute the general performance of an agent in unknown environments. As there are an infinite number of environments, we cannot simply take an expected value with respect to a uniform distribution — we must weight some environments more heavily than others. If we consider the agent's perspective on the problem, it is the same as asking: Given several different hypotheses which are consistent with the observations, which hypothesis should be considered the most likely? This is a fundamental problem in inductive inference for which the standard solution is to invoke Occam's razor: *Given multiple hypotheses which are consistent with the data, the*

simplest should be preferred. As this is generally considered the most intelligent thing to do, we should test agents in such a way that they are, at least on average, rewarded for correctly applying Occam's razor. This means that our a priori distribution over environments should be weighted towards simpler environments.

As each environment is described by a computable measure, we can measure the complexity of these in the standard way by considering their Kolmogorov complexity. Specifically, 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, formally $K(\mu) := \min_p\{l(p) : \mathcal{U}(p) = \mu\}$. We can now define the *universal intelligence* of an agent π to simply be its expected performance,

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

It is clear by construction that universal intelligence measures the general ability of an agent to perform well in a very wide range of environments, as required by our informal definition of intelligence given earlier. 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 which includes a reward signal. Universal intelligence also reflects Occam's razor in a natural way; like standard intelligence tests for humans which define the correct answer to a question to be the simplest consistent with the given information.

By considering V_μ^π for a number of basic environments, such as small MDPs, and agents with simple but very general optimisation strategies, it is clear that Υ correctly orders the relative intelligence of these agents in a natural way. If we consider a highly specialised agent, for example IBM's DeepBlue chess super computer, then we can see that this agent will be ineffective outside of one very specific environment, and thus would have a low universal intelligence value. This is consistent with our view of intelligence as being a highly adaptable and general ability.

A very high value of Υ would imply that an agent is able to perform well in many environments. Such a machine would obviously be of large practical significance. The maximal agent with respect to Υ is the theoretical AIXI agent which has been shown to have many strong optimality properties, including being self-optimising in all environments in which this is at all possible for a general agent [3]. Such results confirm the fact that agents with high universal intelligence are very powerful and adaptable.

Universal intelligence spans simple adaptive agents right up to super intelligent agents like AIXI, unlike the pass-fail Turing test which is useful only for agents with near human intelligence. Furthermore, the Turing test cannot be fully formalised as it is based on subjective judgements. Perhaps an even bigger problem is that the Turing test is highly anthropocentric, indeed many have suggested that it is really a test of humanness rather than intelligence. Universal intelligence does not have these problems as it is formally specified in terms of the more fundamental concept of complexity.

References

[1] C. Fiévet. Mesurer l'intelligence d'une machine. In *Le Monde de l'intelligence*, volume 1, pages 42–45, Paris, November 2005. Mondeo publishing.

[2] D. Graham-Rowe. Spotting the bots with brains. In *New Scientist magazine*, volume 2512, page 27, 13 August 2005.

[3] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2004. 300 pages, http://www.idsia.ch/~marcus/ai/uaibook.htm.

[4] S. Legg and M. Hutter. A universal measure of intelligence for artificial agents. In *Proc. 21st International Joint Conf. on Artificial Intelligence (IJCAI-2005)*, Edinburgh, 2005.

[5] S. Legg and M. Hutter. A formal measure of machine intelligence. In *Proc. Annual machine learning conference of Belgium and The Netherlands (Benelearn-2006)*, Ghent, 2006.