Journal of Artificial Intelligence and Consciousness

Measuring Intelligence in Natural and Artificial Systems

David Gamez

Department of Computer Science, Middlesex University,
London, NW4 4BT, UK
d.gamez@mdx.ac.uk

A systematic understanding of the relationship between intelligence and consciousness can only be achieved when we can accurately measure intelligence and consciousness. In other work I have suggested how the measurement of consciousness can be improved by reframing the science of consciousness as a search for mathematical theories that map between physical and conscious states. This paper discusses the measurement of intelligence in natural and artificial systems. While reasonable methods exist for measuring intelligence in humans, these can only be partly generalized to non-human animals and they cannot be applied to artificial systems. Some universal measures of intelligence have been developed, but their dependence on goals and rewards creates serious problems. This paper sets out a new universal algorithm for measuring intelligence that is based on a system's ability to make accurate predictions. This algorithm can measure intelligence in humans, non-human animals and artificial systems. Preliminary experiments have demonstrated that it can measure the changing intelligence of an agent in a maze environment. This new measure of intelligence could lead to a much better understanding of the relationship between intelligence and consciousness in natural and artificial systems, and it has many practical applications, particularly in AI safety.

Keywords: Intelligence, consciousness, prediction, predictive brain, Bayesian brain, probability, artificial intelligence, AI, universal measure of intelligence.

1. Introduction

Humans are highly intelligent, and their brains are associated with rich states of consciousness. The connection between intelligence and consciousness in humans is unlikely to be a necessary relationship or a natural law because we can use radically different architectures (biological neurons, silicon chips, etc.) to build artificial systems that have the same level of intelligence (measured through objective tests), which are unlikely to have the same consciousness. Some of the possible connections between intelligence and consciousness were discussed in a previous paper [Gamez, 2020], which concluded that the relationship between intelligence and consciousness can only be systematically studied when we can accurately measure intelligence and consciousness in humans, non-human animals and artificial systems.

In previous work [Gamez, 2018] I suggested how we could develop algorithmic theories of consciousness that would be able to generate believable predictions about the consciousness associated with a particular physical state. This would be a major

contribution to our ability to measure consciousness. This paper addresses the other half of the problem: how can we reliably measure the intelligence of humans, non-human animals and artificial systems?

The first part of this paper discusses previous work on the measurement of intelligence in humans, non-human animals, and artificial systems. This includes batteries of questions, such as IQ tests, cognitive tests for animals, and universal measures of intelligence. After summarizing the limitations of these previous measures, Section 3 discusses recent theories about probability and prediction in the brain and the connection between prediction and intelligence. Section 4 describes a new algorithm for measuring predictive intelligence, which can be applied to humans, non-human animals and artificial systems. This section also summarizes some early experimental work in which the algorithm is used to measure the changing intelligence of an agent as it explores a maze environment. Section 5 covers some limitations of the current algorithm and suggests how it could change our way of thinking about intelligence. In the future this new measure of intelligence could lead to better ways of regulating and controlling artificial intelligence.

2. Measurement of Intelligence

2.1. *What is Intelligence?*

Intelligence is a complex multifaceted term and many overlapping definitions have been put forward. These include cognitive ability, rational thinking, problem-solving and goal-directed adaptive behavior [Bartholomew, 2004]. Most people believe that intelligence is some kind of general ability to think, understand and solve problems. It has also been claimed that there are multiple types of intelligence - for example, musical intelligence, linguistic intelligence, and emotional intelligence [Gardner, 2006]. Warwick [2000] frames this more generally with his idea that intelligence is a high-dimensional space of abilities. Other people working in AI have linked intelligence to the achievement of goals or rewards (see Section 2.5).

A distinction is often made between fluid and crystallized intelligence [Cattell, 1971]. Crystallized intelligence is a stored ability to solve problems. For example, older intelligence tests included factual questions, such as “Who is the president of the USA?”. The answer to this question must be remembered – it cannot be deduced by reasoning. Crystallized intelligence also includes rules that can be used to solve problems, known as heuristics. For example, it is theoretically possible to deduce how to solve a Rubik’s cube from scratch. However, most people use heuristics to solve different parts of the problem - for example, a method for moving a color to a different face - and then sequence the heuristics together to complete the puzzle. Heuristics also exist for some of the problems that appear in intelligence tests.

Fluid intelligence is the ability to generalize knowledge and solve problems that have not been seen before. For example, someone with high fluid intelligence might be able to

generalize what they have learnt from solving the Rubik's cube to similar puzzles. Modern intelligence tests are mostly designed to measure fluid intelligence. In humans there is a constant interaction between fluid and crystallized intelligence. A solution to a problem might be discovered through fluid intelligence and then stored for rapid recall at a later date.

2.2. *Measurement of Human Intelligence*

Intelligence is not a precisely defined property, like mass or charge, and it cannot be directly measured. People recognize intelligent behavior and rank people according to their intelligence, but we cannot point to the intelligence in a brain and we cannot program general intelligence into a machine.

Over the last hundred years there has been a large amount of work on the indirect measurement of human intelligence using batteries of tests that measure behavioral characteristics judged to be linked to intelligence. In the early days these tests included significant numbers of questions based on factual knowledge (crystallized intelligence). Modern human intelligence tests are now mostly based on verbal reasoning, spatial manipulation and mathematics. The results from these tests are typically converted into values of intelligence quotient (IQ) or g-score. To calculate IQ you take the test results from a sample of the population and calculate the mean and standard deviation. The mean score is assigned an IQ of 100 and each standard deviation above and below the mean corresponds to 15 IQ points. The resulting IQ score can be used to rank individuals according to how well they perform on a battery of intelligence tests. IQ is a population derived measure that does not correspond to a property of a particular individual.

Within the scientific community intelligence test results are often analyzed for factors that explain the relationships between the test results. Studies have shown that factors related to specific cognitive abilities – for example, reasoning, memory, and processing speed – can explain the results of closely related tests, and these factors are, in turn, linked to a single underlying factor, g, which is thought to correspond to intelligence. Like intelligence, g cannot be directly measured, so the test results are expressed as a g-score. Measures of IQ and g-score are controversial and they have often been misused. However, they have played a valuable role in scientific research on intelligence and they can be an effective way of pre-processing large numbers of applicants for jobs, education, or the military.

The results from human intelligence tests have been shown to be correlated with other measures of success. For example, people who score highly in intelligence tests are more likely to achieve advanced educational degrees and pursue careers in areas, such as science, that are generally regarded as requiring intelligence [Robertson *et al.*, 2010]. This correlation of intelligence tests with societal measures of intelligence gives IQ and g-score considerable plausibility as measures of human intelligence.

2.3. *Measurement of Intelligence in Non-human Animals*

Animals cannot take human intelligence tests, so there has been a lot of work on the development of cognitive test batteries for animals [Shaw & Schmelz, 2017]. While it might be possible to come up with a plausible set of tests that could be applied to similar animals, this approach is likely to neglect the different types of intelligence that animals develop to survive in their ecological niche. A measure of intelligence that is designed for sheep or fish, for example, cannot easily be transferred to birds or bees. Suppose we want to develop a test that compares human and pigeon intelligence. We could include mathematical abilities and spatial reasoning in our tests, which might be common to both. But pigeons have a greater capacity to map and navigate through their environment, so should this be included in the test as well? As our test battery expands with each species we will end up with a very ad-hoc collection, with each animal scoring well on the tests that are specific to their own set of abilities. It seems highly unlikely that we will be able to design a single set of cognitive tests that would enable us to meaningfully compare intelligence across all species.

A second problem with the measurement of non-human animal intelligence is that we do not have a way of connecting an animal's test results to other indicators of intelligence for that species. Most people would agree that a person who gets top grades in school, gets a first at MIT and publishes groundbreaking physics research is likely to be intelligent. If an intelligence test gives this person a low score, then this is a failure of the test, not an indicator of low intelligence. But how could we ground the results of intelligence tests in octopi, bees or dogs? Animals do not take advanced degrees or write papers on quantum theory. Mating success is not correlated with intelligence in humans, so we have no reason to believe that this could be used to validate the results of animal intelligence tests. It is far from clear how we could prove that intelligence tests in animals measure anything more than the ability to perform the test itself.

These problems are often addressed by giving simplified human tests to animals– for example, tests of spatial reasoning or mathematical ability [Boysen and Capaldi, 1992]. This is a form of Turing testing that measures the extent to which non-human animals exhibit human intelligence. It is not a meaningful measure of non-human animal intelligence and it does not enable us to compare general intelligence across species.

2.4. *Measurement of Artificial Intelligence*

Many different types of system are classified as artificially intelligent, for example:

- *Machine learning algorithms*. Trained to label data, predict time series, and so on.
- *Chatbots*. Replicate human conversational ability.
- *Natural language processing*. For example, Watson [Ferrucci, 2012].
- *Cognitive systems*. For example, IDA [Franklin, 2003].

- *Game playing systems*. For example, the deep Q-network that can play ATARI video games [Mnih *et al*., 2015].
- *Computer models of the brain*.
- *Computer models of consciousness*. For example, CiceroBot [Chella *et al.*, 2007].
- *AGI*. There have been some attempts, such as Cyc [Lenat and Guha, 1993].
- *Self-driving cars*.[a]

These systems have diverse expertise and operate within different highly constrained environments. None of them can understand or pass an intelligence test that has been written for humans or non-human animals. However, it would be straightforward to program AIs to outperform humans on IQ tests. What, then, is this *intelligence* that these artificial systems have in common and how could we possibly measure it?

Artificial intelligence can be measured using Turing testing, which takes humans as the benchmark and ranks machines according to the extent that they match human intelligence. One problem with this approach is that as machines improve they are likely to exhaust the possibilities of human tasks. For example, they might eventually map out and completely understand all the possibilities of Go, which would become for them what Tic Tac Toe is for humans – a trivial game whose possibilities can be easily comprehended. To rank AIs according to their intelligence we need tasks that challenge them and which they can complete to different degrees. If they all completely solve a task that is challenging for humans and get the same score, then we can, at most, say that they have super-human intelligence on that task.

A more serious problem with Turing testing is that it cannot measure non-human forms of intelligence. For example, computers are much better at processing vast amounts of data, so they could have much higher levels of intelligence in bioinformatics, while being incapable of solving a Raven's Matrix. It would be extremely anthropocentric to declare that a machine is not intelligent because it cannot solve the narrow range of problems that can be tackled by human intelligence.

These issues suggest that Turing testing is only useful when we want to produce human-like machines whose intelligence does not significantly exceed human intelligence. It cannot be used to measure artificial intelligence that significantly exceeds human intelligence or that operates in a completely different area.

2.5. *Universal Measures of Intelligence*

To address the problems with measuring natural and artificial intelligence, people have developed *universal* measures of intelligence that are, in theory, applicable to any system at all. For example, Legg and Hutter [2007] define intelligence as an agent's ability to achieve goals in a wide range of environments. Their algorithm measures intelligence by

[a] These are not exclusive categories. For example, AlphaGo and Watson are built with machine learning algorithms.

summing the rewards that an agent receives across all possible environments, with some adjustment for the complexity of different environments. This measure has some intuitive plausibility, but it is not practically calculable because it sums across all possible actions of the agent and across all possible environments. A more practical goal-achievement/reward-based measure of intelligence has been proposed by Hernandez-Orallo and Dowe [2010].

As discussed in Section 2.2, the results of IQ tests in humans are thought to be linked to intelligence because they are correlated with achievements that are widely regarded as indicators of intelligence. People who get good grades at school and publish papers in *Nature* typically get high scores in IQ tests. However, the goals and rewards that are linked to intelligence in humans are a subset of the goals that humans pursue and the rewards that they receive from their environment. The achievement of some human goals is highly correlated with intelligence; the achievement of other human goals, such as drinking beer and watching Netflix, is highly rewarding, but extremely weakly correlated with IQ or g-score. More intelligent humans also tend to have less achievable goals (high impact publications; breakthroughs in AI, etc.) and their preoccupation with these harder goals often leads them to receive less rewards from their environment. So goal-achievement/reward-based measures of intelligence will be much *less* correlated with human intelligence than IQ tests or g-score. With non-human animals, the majority of goals and rewards are achieved with limited amounts of crystallized intelligence. This suggests that goal-achievement/reward-based intelligence measures are, at best, likely to be weakly correlated with non-human animal intelligence.

Goal-achievement/reward-based measures of intelligence are even more problematic with artificial systems. To begin with, most artificial systems do not have anything resembling a human goal. Roughly speaking, a human goal is a state of the world that a person pictures in their imagination and associates with a positive emotional state. The person compares the imagined goal with the current state of the world to determine if they have achieved the goal. The reward is a positive emotional state that is induced by the environment or by the perception of the environment.[b] For example, if I am searching for my keys and find my keys, then I experience a feeling of relief. Most AI systems do not work this way - a 'goal' in an AI system is often an internal variable that causes a *while* loop to exit when a particular environment state is reached. It is also very hard to see how a coherent non-anthropocentric notion of reward can be developed for artificial systems that lack emotions and consciousness. As far as a computer is concerned an increase in a variable named 'reward' is no different from an increase in a variable named 'punishment'. Both 'reward' and 'punishment' are just human labels that do not exist at the level of the binary machine code that is executed by the computer.

Another problem with goal-achievement/reward-based measures of artificial intelligence is that environments can have extremely large numbers of goals. Consider a binary environment whose state can be 1 or 0. An agent in this environment can have the

[b] This interpretation of goals and rewards is mostly drawn from Damasio [1994].

goal for the state to be 1 or 0, to alternate in the pattern 11010010, and so on. These are all acceptable goals for an agent in this environment. The theoretical limit for the number of goals that an agent can have in this environment is $2^{\text{memory size}}$, where each memory state is a different number that corresponds to the pattern of 1's and 0's that the agent 'aims' to achieve in its environment. From a human point of view, these goals make no sense, but goal-achievement/reward-based measures of intelligence do not place any restrictions on the *quality* of the goals that are 'aspired to' by the machine. This large goal space enables artificial systems to set themselves goals that can be achieved in their environment without any intelligence. For example, consider a binary environment that randomly switches between 1 and 0. An agent could set itself goals that correspond to every possible random pattern for the next 1000 time steps. It would not have to do anything or apply any intelligence to achieve these goals. If the achievement of goals was linked to rewards, then the agent would achieve a high score on a reward-based measure of intelligence.

A third problem with goal-achievement/reward-based measures of intelligence is that they conflate two different functions of a system: 1) Problem-solving intelligence that understands the environment and develops plans that can be executed to achieve the goals and rewards. 2) Actuators that enable the agent to execute the plans, achieve the goals and receive the rewards. Suppose two agents in an environment have the same goal of moving a heavy rock 50m and the achievement of this goal has a large reward. One agent is strong and the other is weak. Both agents can use their intelligence to develop a plan about how the rock can be moved to receive the reward, but only one of the agents is strong enough to move the rock and get the reward. A goal-achievement/reward-based measure would attribute more intelligence to the strong agent because it received the reward, but this is incorrect because in this example both agents have the same intelligence – the only difference between them is their physical ability to *execute* the plans formulated by their intelligence.

Finally, there is a problem about how the reward of an AI system is connected to the environment. In humans rewards are often connected to states of the environment – receipt of food, warmth, etc. In an AI system a variable named 'reward' could be increased when the AI achieved its goals. The AI system could also increase this variable by itself without any interaction with its environment – a form of wireheading. Again, inaccurate high scores would be achieved on a goal-achievement/reward-based measure of intelligence. We need a better universal way of measuring intelligence in humans, non-human animals and artificial systems.

3. Prediction and Probability in Natural and Artificial Systems

3.1. *Prediction and Probability in the Brain*

Older models of the brain, for example Marr [1982], were based on the idea that low level sensory input layers, for example V1, computed simple features of the input data and these

features were successively processed into more complex representations, such as faces and people. In this model the information that is passed from lower to higher levels is the presence or absence of the features detected by the layer - for example, the presence of an edge or hand at a particular point in the visual field.

In recent years there has been a surge of interest in the idea that the primary function of the brain is the generation of predictions about the environment [Clark, 2016]. According to these theories, each layer in the brain generates predictions about activity in the layer below. Each layer compares the predictions with its own activity and passes information about the prediction errors back up to the layer above. This explains why there are more top-down than bottom-up connections in the brain. Predictive brain theories typically treat the brain's predictions as probability distributions. This accommodates situations in which we are certain about something, as well as more common scenarios in which we assign probabilities to different events. For example, I use my past experience to estimate the probability that pasta will be available in a shop on a certain day. This probability can change - for example, when people stock up during a pandemic. People working on the Bayesian brain investigate the extent to which the probability distributions of the brain's predictions match the probability distributions of the environment [Knill and Pouget, 2004].

There is little evidence for Bayesian and predictive theories of the brain. However, they are consistent with our subjective experiences and a good match for what we know about the brain. If the predictive and Bayesian brain hypotheses are partly or wholly true, then the generation of probabilistic predictions is a core function of the brain, and we would expect there to be a strong correlation between a brain's predictive ability and its intelligence.

3.2. *Probability and Prediction in AI Systems*

AI systems make predictions that are often expressed as probability distributions. For example, robots predict the consequences of different movements and select sequences of actions that maximize the probability of achieving their goals. The output of machine learning algorithms is often expressed as probability distributions, and their mapping between input and output can be regarded as predictions about the labels that are associated with particular input states. Any system that learns about the regularities of an uncertain environment will make some form of probabilistic predictions.

3.3. *Prediction and Intelligence*

There are close links between a system's ability to generate predictions and its intelligence. An agent with perfect predictive ability would have god-like omniscience. It would know what would happen under all possible permutations of its environment; it could plan sequences of actions that would have the highest probability of achieving its goals. A

predictive approach to intelligence separates an agent's understanding of its environment from the goals that is has in that environment and its ability to achieve these goals. Prediction can be measured in both natural and artificial systems and a predictive approach to intelligence fits in well with recent work on the predictive and Bayesian brain.

My hypothesis is that prediction is the most important component of natural and artificial intelligence: If we can accurately measure a system's ability to predict, we can accurately measure the system's intelligence. The next section summarizes the work that I have done on a universal measure of intelligence based on prediction. It starts with definitions of internal states, predictions, and distinct states of the environment. The algorithm is then presented followed by an overview of some recent experimental work.

4. A Universal Measure of Intelligence Based on Prediction

4.1. *Internal States and Predictions*

When assessing the accuracy of a prediction it is natural to compare the prediction with a state of the environment. For example, I might predict that it will rain and then check the weather to see if my prediction came true. Comparison between predictions and environment states is natural and easy when similar systems (for example, people) are interacting with the same environment (the physical world). It is much less straightforward with non-human animals and artificial systems, which have different sensors, different environments and different ways of representing their environments.

Suppose a Khepera robot,[c] a dog and a human are in a room. For the Khepera, the states of the environment consist of the distances of objects from its ultrasonic sensors.[d] It might have a way of converting these internal measurements into a map of the room, but this would be a 2D map that was limited to objects at the same height as the robot. The dog has a three-dimensional experience of the room with limited color vision and a 3D odor map. The human experiences a sophisticated distribution of colors, less smell and a deeper understanding of the objects - for example, it can see at a glance that a pattern of colors on the wall is a map. Which is the real objective room whose probability distributions should be compared with the predictions of the Khepera, dog and human? Is the room 2D or 3D? Is it colored? Does it contain a map and a 3D distribution of smells? Physics states that the objective room is the wave-particle distributions of all the elementary particles, but the Heisenberg uncertainty principle tells us that this cannot be fully and accurately measured, even if we had appropriate instruments. Clearly it does not make sense to base our assessment of the human's intelligence on its ability to predict 3D smell patterns or the intelligence of the dog on its ability to predict color patterns that it cannot sense. The only

[c] There are several versions of the Khepera robot. In this example, I am considering a basic version that only has ultrasonic distance sensors.

[d] Actually the distance values are just numbers for the Khepera – we cannot assume that it has the human concepts of objects and distances. This does not affect the basic point that I am making here.

reasonable and fair way of evaluating the intelligence of each agent in the room is to compare their predictions with the probability distributions of their own future states. Systems with more sensors and richer representations of their environment will exhibit higher intelligence because their internal states will be less perturbed by events that they are not sensing. For example, a human and a dog can see a ball bouncing around a room and predict its trajectory, but a Khepera robot will only see the ball intermittently appearing in its 2D array of distance sensors, which will be very difficult to predict.

My predictive measure of intelligence is based on an agent's internal states, I_1, I_2 ... I_q. At each point in time and/or at each place in the environment it generates one or more predictions about these internal states, $P_{1\text{-}1}$, $P_{1\text{-}2}$ ... $P_{1\text{-}r}$, $P_{2\text{-}1}$, $P_{2\text{-}2}$ ... $P_{2\text{-}r}$ ... $P_{q\text{-}1}$, $P_{q\text{-}2}$, $P_{q\text{-}r}$, where $P_{2\text{-}5}$ is a prediction about the state of I_2 at time 5 (measured in time steps or seconds). The predictions are probability distributions across values. For example, the agent might predict that I_1 will have value 35 with probability 0.7 and value 37 with probability 0.3. They can also be probability distributions across time. For example, the agent might predict that the value probability distribution, *P*, has probability 0.2 of occurring in 2 seconds, probability 0.5 of occurring in 3 seconds and probability 0.3 of occurring in 4 seconds. These probability distributions can be combined into a single joint probability distribution across values and time.

We typically make predictions about future states. In this measure of intelligence they can also be about past states or about states that are outside the current environment. For example, people make predictions about the Big Bang or about what their children are eating for lunch at school. Systems can also predict the labels that are associated with a particular input pattern - for example, a face-recognition algorithm.

4.2. *Distinct States of the Environment*

Roughly speaking the most common types of environment are:

- *Spatial*. The location of an agent in space can change, leading to different sensory inputs.
- *Temporal*. An agent's environment can change in time independently of the movement of the agent.
- *Data*. Many AIs work within a data environment that they are progressively exposed to during training and testing.

These are not exclusive categories: the environment of many agents is a combination of multiple types.

In my algorithm two environment states are considered to be the same if the agent cannot distinguish between them. Two environment states are *distinct* if they lead to different internal states. To measure an agent's intelligence, we need to expose it to all *distinct* environment states and sum up the accuracy of its predictions across these states. The set of distinct environment states will vary with an agent's sensors and internal memory. For example, a light that flashes red-green is indistinguishable from a light that

flashes green-red for an agent that cannot detect color. The sequences 1010 and 010 are indistinguishable for an agent that can only store one digit. In a complex environment an agent's intelligence could be estimated by exposing it to a representative sample of distinct environment states.

4.3. *Algorithm for Measuring Predictive Intelligence*

This algorithm for measuring predictive intelligence compares the probability distributions of an agent's predictions with the probability distributions that actually occur as the agent interacts with its environment. This is illustrated in Fig. 1.

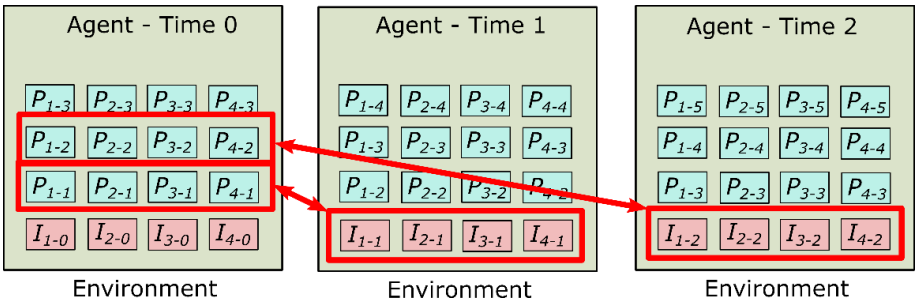

Fig. 1. Agent's predictions about internal states. The agent has internal states I_1, I_2, I_3 and I_4. $I_{1\text{-}0}$, $I_{2\text{-}0}$, $I_{3\text{-}0}$ and $I_{4\text{-}0}$ are the probability distributions of the internal states at time 0. $P_{1\text{-}1}$, $P_{1\text{-}2}$, $P_{1\text{-}3}$ are predictions that the agent makes about the values of I_1 at times 1, 2 and 3. As the spatial and temporal properties of the environment change, the future states of I_1, I_2, I_3 and I_4 are compared with earlier predictions to check their accuracy.

To measure the accuracy of the predictions we need to compare the probability distributions of the predictions with the later probability distributions of the internal states – for example, in Fig. 1, comparing the prediction $P_{1\text{-}2}$ made at time 0 with $I_{1\text{-}2}$ at time 1. In this algorithm Hellinger distance is used to compare probability distributions:

$$H(P,Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2} \qquad (1)$$

Hellinger distance is 0 when there is an exact match between two probability distributions and 1 when there is a complete mismatch between two probability distributions. So 1-*H(P,Q)* gives us the *degree of match* between two probability distributions, expressed as a number between 0 and 1.

To calculate the total prediction match, PM_e, between predictions and internal states we need to sum up the match between the probability distributions of the agent's predictions, $P_{1\text{-}1}$, $P_{2\text{-}1}$, $P_{3\text{-}1}$, etc. and the probability distributions of the agent's internal states $I_{1\text{-}1}$, $I_{2\text{-}1}$, $I_{3\text{-}1}$, etc. for all distinct states, s=1 … s=p, of environment e:

$$PM_e = \sum_{s=1}^{p}\sum_{i=1}^{q}\sum_{t=1}^{r} 1 - H(P_{i-t}, I_{i-t}) \quad (2)$$

Environments have large differences in complexity and it is much easier to make predictions about simple environments. This problem can be addressed by multiplying PM_e with the Kolmogorov complexity of the predictions. Kolmogorov complexity cannot be directly calculated, so this intelligence measure uses the compressibility of the predictions as its measure of complexity. This is the length of the compressed predictions, *K*, divided by the length of the uncompressed predictions, *L*. This leads to an equation for calculating predictive intelligence, PI_e, of an agent in environment *e*:

$$PI_e = \frac{K}{L}\sum_{s=1}^{p}\sum_{i=1}^{q}\sum_{t=1}^{r} 1 - H(P_{i-t}, I_{i-t}) \quad (3)$$

Eq. 3 can be used to compare the predictive intelligence of agents within a single environment. However, we often want to evaluate agents' intelligence across multiple environments. This is particularly important when we are analyzing the intelligence of artificial systems, which often outperform humans in single environments, such as Go or chess, while being completely useless in the real world. When summing intelligence across environments we cannot simply add the PI_e values from different environments together. If we took the simple sum, an agent would double its intelligence across two environments that were almost identical. To address this issue, the sum of PI_e across multiple environments is multiplied by the joint Kolmogorov complexity of the two environments divided by the sum of the complexity of the environments considered individually, as shown in Eq. 4.

$$ᛕ_c = \frac{K(L_1 + L_2 + \cdots + L_n)}{K(L_1)+K(L_2) + \cdots + K(L_n)}\sum_{e=1}^{p} PI_e \quad (4)$$

$L_1 \ldots L_n$ are strings describing environments 1-*n* and *K(L)* is the length of the shortest program describing this string. In practice Kolmogorov complexity can be approximated by compression algorithms. When two environments are very similar it should be possible to find a compact representation of the two together. In this case the joint complexity will be approximately the same as the individual complexity of each environment and the complexity factor will approximate 1/2. When two environments are very different, the joint complexity will be similar to the sum of the individual complexities and the complexity factor will be close to 1. So agents that make accurate predictions about multiple different environments will be attributed a higher $ᛕ_c$ than agents that make accurate predictions about multiple similar environments.

The symbol, ᛕ, that I have chosen for this measure of intelligence is the Old Norse letter (rune) that corresponds to our modern 'p' sound (named 'peorth', 'perth' or 'pertho'). In addition to its role as the first letter of the word 'prediction', ᛕ is associated with the dice

cup, chance, secrets, destiny and the future, which is rather appropriate for a measure that is designed to measure our ability to accurately predict the future. In Eq. 4, $\aleph_c$ is crystallized predictive intelligence. I define fluid predictive intelligence, $\aleph_f$, as the rate of change of $\aleph_c$, as shown in Eq. 5:

$$\aleph_f = \frac{d\aleph_c}{dt} \tag{5}$$

4.4. *Experimental Work*

This measure of intelligence was tested on a simple agent that interacts with multiple maze environments. The environments consist of walls that the agent cannot enter, empty squares and rewards. The agent has sensors on its front, left and right sides that tell it the contents of adjacent squares. It also has a sensor that detects the contents of the current square. The agent can rotate left or right and it can attempt to jump in the direction that it is pointing. The attempt will fail if it tries to jump into a wall. The agent simulation is shown in Fig. 2.

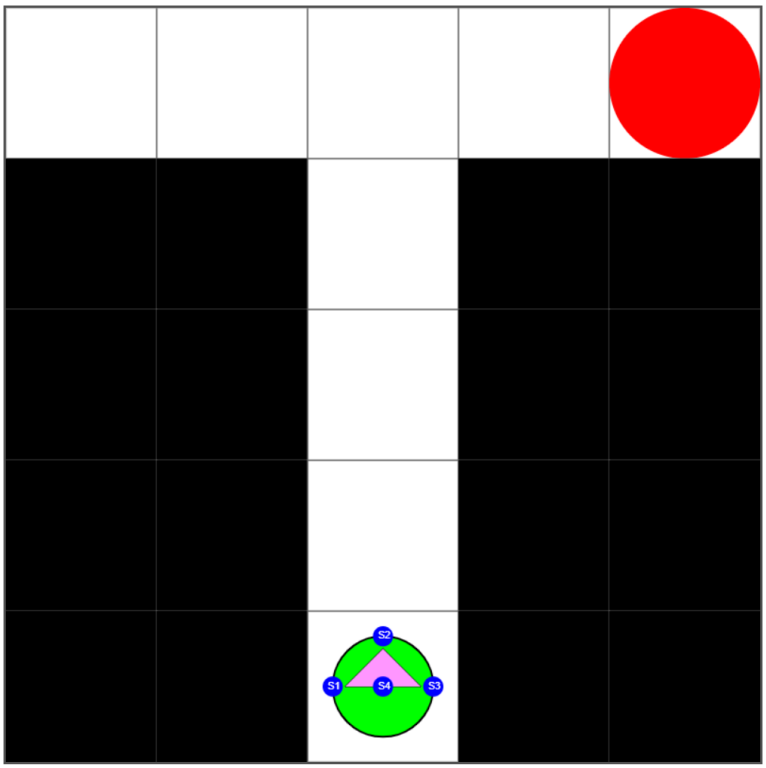

Fig. 2. Agent in maze environment. Sensors S1-S3 give the agent information about the contents of the adjacent square (empty, wall, reward). S4 gives information about the contents of the current square. The agent can rotate left and right and it can move forward into empty squares or into squares containing a reward.

The agent uses its previous experience of the maze to predict the inputs that will occur after an action (move forward or change direction). If it has no experience of the consequences of its actions under particular input conditions, then it assigns equal probability to all input state values.

To discover the actual probability distributions of the agent's inputs under different state transitions, the system switches the agent's learning off and monitors the agent as it tries every possible movement in each location of the environment. Then learning is switched on. As the agent explores its environment its predicted probability distributions for a movement are compared to the actual probability distributions to calculate $\aleph_c$. Fluid intelligence, $\aleph_f$, is calculated by taking the differential of a polynomial approximation to $\aleph_c$ at the center of a time window. An example is shown in Fig. 3.

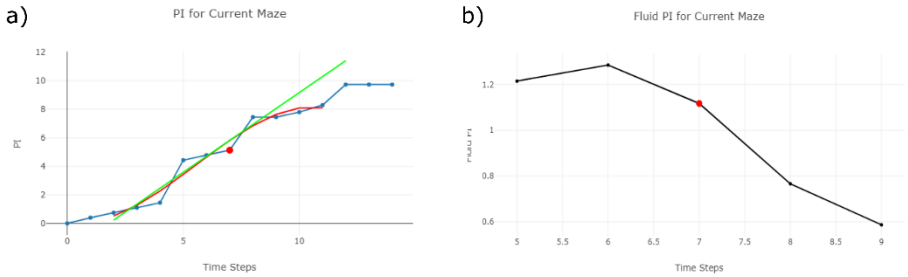

Fig. 1. a) Crystallized intelligence of agent over time. The red line is the polynomial approximation that is used to calculate fluid intelligence. The green line shows the slope of the polynomial at the red dot. b) Fluid intelligence of agent over time. The red dot corresponds to the center of the time window that is used to calculate the polynomial approximation in the left graph. The value of fluid intelligence at the red dot is the slope of the green line in the left graph.

These preliminary experiments demonstrate that $\aleph_c$ and $\aleph_f$ can be calculated in real time on small systems. The simulation can be viewed online at www.davidgamez.eu/pi and the source code is available for download.

5. Discussion

In this approach intelligence is relative to a set of environments – intelligent systems do not have the ability to understand and predict the features of *any* environment. This is reasonable because the real-world intelligences that we know are only able to comprehend narrow sets of environments. Some intelligences are good at numbers; others can predict changes in the real physical world; others make accurate predictions about large data sets. People with an IQ of 180 can pass exams and write scientific papers, but they cannot identify patterns in large data sets and they often have low social and emotional intelligence. What links all these types of intelligence together is their ability to generate accurate predictions within specific environments.

We could only come up with final and complete values of $\aleph_c$ and $\aleph_f$ for an agent if we could measure its predictive intelligence in every possible environment. Since there is an

effectively infinite number of possible environments, a system's χ_c and χ_f values will always be relative to the set of environments in which its intelligence has been measured, and it will always be subject to change – for example, if the system turns out to be good at making predictions about a new environment that it has never encountered before.

The experiments described in Section 4.4 show that χ_c and χ_f can be measured on small systems and environments. In most situations it will be practically impossible to measure PM_e for every distinct environment state, so it will be necessary to develop methods for estimating χ_c and χ_f from limited measurements, possibly using similar methods to Hernandez-Orallo and Dowe [2010].

In some situations, χ_c and χ_f might not match our intuitions about the intelligence of a system. One problematic case is a system that generates large numbers of predictions about internal states that are disconnected from the environment. For example, a system could create many variables that increase by 1 every time step and then generate accurate predictions about future states of these variables. The complexity components of the algorithm would ensure that little intelligence was associated with each internal state. However, the overall intelligence of this system would still end up large because its intelligence would exist in every possible environment. One potential solution to this problem would be to use statistics, Granger causality, mutual information or a similar algorithm to measure the connection between the internal states and the environment. The prediction match of the internal states could then be weighted by the extent to which they are connected to the environment. This would ensure that an agent's intelligence is *about* its environment.[e]

This algorithm uses an internalist approach because there are no observer-independent facts of the matter about the current state of the environment beyond the unmeasurable position and velocity of the elementary wave-particles. The problem with an internalist approach is that it relies on an observer-dependent mapping between physical states, input states I_1, I_2 ... I_q and predictions $P_{1\text{-}1}$ ... $P_{q\text{-}r}$. For example, the internal states of a brain can be measured at the level of neurons, neuron groups, EM waves, blood, and so on. To analyze a brain for intelligence, we first have to decide which states we will use for our analysis. Then we have to decide which of these states are internal states, which are predictions and which should be ignored. The situation is equally challenging with computers. When we develop AIs we allocate variables and naturally take these as the states that our measure of intelligence should be applied to. However, the internal states of an AI can also be obtained by measuring the RAM, EM waves and so on. What *we* take to be obvious internal states to measure are not the only states that we can treat as internal and analyze for predictions. I have discussed this problem at length in relation to Tononi's IIT [Gamez, 2016] and in relation to computational theories of consciousness [Gamez, 2014]. Tononi's solution is to consider all possible levels and look for the maximum Φ

[e] This raises the problem discussed in Section 4.1 about which states of the environment are connected to the agent's internal states. This could potentially be addressed by looking for the maximum mutual information (for example) between internal states and the environment. However, this could lead to a combinatorial explosion.

across all levels [Tononi, 2010], but this is impossible because the number of levels is effectively infinite. As I have discussed elsewhere [Gamez, 2018], consciousness is an 'objective'[f] property of a system and objective properties cannot be correlated with subjective properties. However, it is less clear whether intelligence is an 'objective' property like consciousness, so in many cases we might be satisfied with an estimate of the intelligence of a system based on what we consider to be a reasonable choice of internal states. With smaller systems we can sweep the states to look for the level that has maximum intelligence and this could be combined with an analysis of the connection between internal states and the environment.

The current version of the $\aleph$ algorithm could be improved. To begin with, the only maximum or upper limit for $\aleph_c$ is the ability to predict the past and future probability distributions of every wave-particle in the entire universe, so $\aleph_c$ will get very large with big systems and complex environments. This could be addressed by adding a logarithm to Eq. (4). Second, when the agent guesses randomly there is usually some degree of match between the random probability distributions and the actual probability distributions - similar to the score that random guessing obtains in a multiple choice test. This could be addressed by subtracting the prediction match of a random distribution from the prediction match of the predictions in Eq. (2) and Eq. (3). However, the notion of negative intelligence might be problematic, so some thought would have to be given to situations in which the agent's prediction match is less than the match with the random distribution.

There has been a great deal of discussion of AI safety and the existential threat that AI could pose to humanity. In these discussions the notion of intelligence is often instrumentalized – for example, as the achievement of goals or rewards – or interpreted as a very abstract general property of a system that enables it to solve problems in many different areas (human psychology, computer hacking, global domination, etc.). The measure of intelligence that has been put forward in this paper provides a much clearer definition of artificial intelligence and shows us that intelligence can take many different forms. In the future $\aleph_c$ and $\aleph_f$ could be used to identify AI systems that are a potential threat. For example, current machine learning systems already have superhuman levels of $\aleph_c$ in their areas of expertise, but they do not pose any threat to humanity because they have very low levels of $\aleph_c$ in human psychology, the real physical environment, computer hacking, etc. These machines also have low $\aleph_f$ in environments that are significantly different from the data sets that they have been designed to work with. In the longer term, $\aleph_c$ and $\aleph_f$ could be incorporated into laws that regulate the amount of AI intelligence in different environments.

6. Conclusions and Future Work

This paper has set out a new way of measuring intelligence in natural and artificial systems that draws on recent work on the predictive and Bayesian brain and does not rely on a

[f] By this I mean that consciousness really exists, even though it can only be measured subjectively.

problematic connection between intelligence and the achievement of goals and rewards. $\aleph_c$ and $\aleph_f$ improve our theoretical understanding of intelligence, they could lead to a better understanding of the relationships between intelligence and consciousness, and they have many practical applications, particularly in AI safety.

Preliminary experimental work has demonstrated that $\aleph_c$ and $\aleph_f$ can be measured in simple embodied agents. The next stage of this research will be to measure fluid and crystallized intelligence in a machine learning algorithm as it learns to classify data. Performance tests are also planned and it is anticipated that the algorithm will be refined over time to address the issues raised in Section 5 - following a similar trajectory to the improvements to Φ that have been made over the last 15 years.

References

Bartholomew, D. J. [2004] *Measuring Intelligence: Facts and Fallacies* (Cambridge University Press, Cambridge).

Boysen, S. T. and Capaldi, E. J. [1992] *The Development of Numerical Competence : Animal and Human Models* (L. Erlbaum Associates, Hillsdale, N.J.).

Cattell, R. B. [1971] *Abilities: Their Structure, Growth, and Action* (Houghton Mifflin, Boston).

Chella, A., Liotta, M. and Macaluso, I. [2007] Cicerobot: A Cognitive Robot for Interactive Museum Tours, *Industrial Robot: An International Journal* **34**(6), 503-511.

Clark, A. [2016] *Surfing Uncertainty: Prediction, Action, and the Embodied Mind* (Oxford University Press, Oxford).

Damasio, A. R. [1994] *Descartes' Error: Emotion, Reason, and the Human Brain* (G.P. Putnam, New York).

Ferrucci, D. A. [2012] Introduction to "This Is Watson", *IBM Journal of Research and Development* **56**(3.4), 1:1-1:15.

Franklin, S. [2003] IDA - A Conscious Artifact?, *Journal of Consciousness Studies* **10**(4-5), 47-66.

Gamez, D. [2014] Can We Prove That There Are Computational Correlates of Consciousness in the Brain?, *Journal of Cognitive Science* **15**(2), 149-186.

Gamez, D. [2016] Are Information or Data Patterns Correlated with Consciousness?, *Topoi* **35**(1), 225-239.

Gamez, D. [2018] *Human and Machine Consciousness* (Open Book Publishers, Cambridge).

Gamez, D. [2020] The Relationships Between Intelligence and Consciousness in Natural and Artificial Systems. *Journal of Artificial Intelligence and Consciousness*, **7**(1). 51-62.

Gardner, H. [2006] *Multiple Intelligences: New Horizons* (Basic Books, New York).

Haier, R. J. [2017] *The Neuroscience of Intelligence* (Cambridge University Press, Cambridge).

Hernández-Orallo, J. and Dowe, D. L. [2010] Measuring Universal Intelligence: Towards an Anytime Intelligence Test, *Artificial Intelligence* **174**, 1508-1539.

Knill, D. C. and Pouget, A. [2004] The Bayesian Brain: The Role of Uncertainty in Neural Coding and Computation, *Trends in Neurosciences* 27(12), 712-719.

Legg, S. and Hutter, M. [2007] Universal Intelligence: A Definition of Machine Intelligence, *Minds and Machines* **17**, 391-444.

Lenat, D. B. and Guha, R. V. [1993] Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project, *Artificial Intelligence* 61(1), 65-79.

Marr, D. [1982] *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (Freeman, New York; Oxford).

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S. and Hassabis, D. [2015] Human-Level Control through Deep Reinforcement Learning, *Nature* 518(7540), 529-533.

Oizumi, M., Albantakis, L. and Tononi, G. [2014] From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0, *PLoS Computational Biology* **10**(5), e1003588.

Robertson, K. F., Smeets, S., Lubinski, D. and Benbow, C. P. [2010] Beyond the Threshold Hypothesis: Even among the Gifted and Top Math/Science Graduate Students, Cognitive Abilities, Vocational Interests, and Lifestyle Preferences Matter for Career Choice, Performance, and Persistence, *Current Directions in Psychological Science* **19**(6), 346-351.

Shaw, R. C. and Schmelz, M. [2017] Cognitive Test Batteries in Animal Cognition Research: Evaluating the Past, Present and Future of Comparative Psychometrics, *Animal Cognition* **20**, 1003-1018.

Tononi, G. [2008] Consciousness as Integrated Information: A Provisional Manifesto, *Biological Bulletin* **215**(3), 216-242.

Tononi, G. [2010] Information Integration: Its Relevance to Brain Function and Consciousness, *Arch Ital Biol* 148(3), 299-322.

Warwick, K. [2000] *Qi: The Quest for Intelligence* (Piatkus, London).