Using Universal Measures of Intelligence to Govern Artificial Intelligence

David Gamez
Department of Computer Science, Middlesex University, UK
d.gamez@mdx.ac.uk / www.davidgamez.eu

Abstract

There are ongoing debates about AI safety and the possibility that an uncontrolled expansion of artificial intelligence could pose an existential threat to humanity. These issues could potentially be addressed by legislation that regulates AI. However, any attempt to govern AI faces two key challenges. First, there is no commonly agreed definition of artificial intelligence, so we do not have a clear way of specifying which systems the regulations should apply to. Second, it is impossible to effectively govern AI without an accurate way of measuring the amount of intelligence in artificial systems. This paper argues that there is a close connection between the intelligence of a system and its ability to generate accurate predictions. This enables us to develop algorithms that can measure intelligence in natural and artificial systems, and the paper briefly describes some experimental work in this area. In the future this definition and measure of intelligence could play an important role in the governance of AI. We will be able to distinguish between systems that are genuinely artificially intelligent, such as game-playing AIs, from systems that replicate human behaviour without intelligence, such as classifiers and chatbots. We could use this measure of intelligence to establish whether a particular AI system exceeds human intelligence in a particular area, and potentially develop legislation that limits the amount of artificial intelligence in sensitive areas, such as cybersecurity.

1. Introduction

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 behaviour (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. Intelligence has also been linked to skill acquisition and the achievement of goals and rewards (see Section 3.4).

Over the last hundred years we have developed reasonably effective ways of measuring intelligence in humans. While IQ and g-score have limitations, there is a broad consensus that they are correlated with intelligence in humans and with other measures of human intelligence, such as academic grades, publication of scientific papers and success in professional careers (Robertson et al. 2010). Research on the measurement of intelligence in animals has been less successful, partly because it often conflates the measurement of human-like intelligence in animals with the measurement of animal intelligence considered more broadly in all its many forms. There has been very little work on the measurement of intelligence in artificial systems, apart from some interesting proposals about universal measures of intelligence.

There are ongoing discussions about AI safety and the possibility that an uncontrolled expansion of artificial intelligence could pose an existential threat to humanity (Bostrom 2016). A key limitation of these debates is that there is no clear commonly agreed definition of artificial *intelligence*. At the moment a very broad class of systems, including self-driving cars, face-recognition systems, chatbots and game-playing systems, are loosely classified as artificially intelligent. These systems all reproduce, to a greater or lesser extent, particular human behaviours, but only a few of them can plausibly be described as intelligent. Future attempts to regulate artificial intelligence should be based on a much clearer definition of the systems to which the regulations apply.

The effective governance of AI also depends on accurate ways of measuring and comparing the *amount* of intelligence in natural and artificial systems. Without accurate measurement, AI regulations could, at best, specify the *types* of AI that are allowed (for example, banning reinforcement learning) or limit the *actions* that AIs are allowed to perform (for example, banning 'killer robots'). This would do little to mitigate AI threats and it could seriously limit the benefits that AI could bring to society.

The first part of this paper argues for a close link between intelligence and prediction. This leads to three hypotheses about intelligence that enable us to distinguish between artificial systems that are and are not intelligent. Section 3 reviews some of the problems with the current measures of intelligence and then Section 4 describes a new method for the measurement of predictive intelligence in natural and artificial systems, along with experiments that have been carried out to test this measure. The last part of the paper discusses the role that this definition and measure of intelligence could play in the governance of AI.

2. Prediction and Intelligence

2.1 Prediction and the Brain

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. The layers compare the predictions from higher layers with their own activity and pass information about the prediction errors back up to the layers 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. 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, these theories 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.

2.2 Prediction and Action

As I interact with the world, I am constantly predicting the results of different possible actions and selecting the ones that lead to my goals. For example, when I am hungry, I consider the location of different supermarkets and plan how I can get to the best one, taking into account traffic, petrol, crime, and so on. A system that cannot predict cannot plan – it can only react to changes in its environment as they occur. A system with perfect predictive ability would have god-like omniscience. It would know what would happen under all possible permutations of its environment and could plan sequences of actions that would have the highest probability of achieving its goals.

As animals increase in intelligence and behavioural sophistication there is a shift from hard-wired reactions to planned behaviour based on prediction. Snails follow chemical trails and retreat when danger threatens. The world does something to the snail and it responds in an evolutionarily determined way that, on average, leads to the future survival of the species. More sophisticated animals, such as sheep, can classify features of their environment (food, enemies, mates, etc.) and they have a limited ability to predict how their environment will respond to their actions (Marino and Merskin 2019; Gamez 2019). Corvidae (crow family) combine reactive behaviours with actions based on richer predictions about their environment, which enables them to solve more complex problems and build tools. Humans combine their reactive behaviours with planning based on complex predictions on multiple time scales.

2.3 Prediction and Artificial Intelligence

Systems that are classed as artificially intelligent replicate behaviours that typically require intelligence in humans. Self-driving cars or chess-playing programs are regarded as artificially intelligent because human intelligence is required to drive cars and play chess. Dialysis machines, that replicate the functions of the human kidneys, are not regarded as artificially intelligent because blood filtration does not require intelligence in humans.

The problem with this definition of artificial intelligence is that computers can imitate human behaviours in simple ways without any intelligence. For example, natural language conversation requires intelligence in humans, but it can be reproduced in chatbots using a simple pattern matching algorithm. Classification systems, such as face recognition algorithms, are also not likely to be intelligent. While classification plays an important role in natural and artificial intelligence, it is best viewed as a pre-condition for intelligence, rather than as an important component of intelligence itself. For example, identifying an *Amanita virosa* and knowing that it is poisonous is certainly helpful information, but it only plays a role in intelligence when we can make predictions about the physical consequences of eating poisonous mushrooms.

AI systems that are intelligent include game-playing systems, such as AlphaGo, which predict the consequences of different actions in the space of the game. Robots and self-driving cars contain intelligence that enables them to predict the consequences of different actions in the world. Other systems that generate predictions about the future, such as climate models, also exhibit intelligence.

2.4 Retrodiction / Postdiction

Humans use their intelligence to discover facts about the past as well as the future. Historians debate the economic and social consequences of the plague; physicists develop theories about the origins of the universe. This work clearly requires intelligence, and it is typically called retrodiction or postdiction.

Retrodiction is a form of prediction, but to count as intelligence the retrodictions have to be generated from limited information in the same way as predictions about the future. Consider two systems: one is an omniscient recorder that has stored everything that has occurred over the last 100 years; the other reconstructs the course of history from fragmentary archaeological evidence. The history recorder can make extremely accurate retrodictions, but it is not intelligent because it is just using a simple database search to discover the historical facts. The historian that pieces together the facts is intelligent because she is generating retrodictions from limited evidence.

2.5 Predictive Intelligence and Environments

Some people think of intelligence as an abstract property that is completely independent of the environment. A person has a certain amount of intelligence regardless of whether they are working in the natural world or studying a genomics database. However, a system's ability to generate predictions is relative to its environment. One system predicts weather patterns; another system predicts protein folding. No one has developed a completely general prediction system that can outperform systems that are built for specific environments. Even the human brain has been honed by evolution to work well in a hunter gatherer environment: our limited working memory and sensory input make us very poor at handling large data sets, which is one of the reasons why we build AI systems to handle these kinds of tasks. On the other hand, the AI systems that we build to process large data sets cannot make accurate predictions in a hunter-gatherer environment.

When we accept that intelligence is relative to a set of environments, it becomes clear that there are many different forms of natural and artificial intelligence, which are specific to their different environments. We don't have to broaden our concept of intelligence to handle this (embracing Gardener's multiple intelligences or Warwick's high dimensional space of abilities). Instead, we can say, for example, that system A has a high level of intelligence in a musical environment and system B has a high level of intelligence in a chess environment. These environments can be natural, simulated, data, and so on.

The relativization of intelligence to a set of environments helps us to understand and appreciate the many different forms of human intelligence. IQ tests measure intelligence within an academic environment of mathematical symbols, abstract shapes, and so on. This is why IQ tests are correlated with measures of academic success (school grades, advanced degrees, publication of papers, professional careers, etc.). But the academic environment is just one area where human intelligence operates. A successful plumber has a high level of intelligence within the environment of pipes, fittings, water flow, etc. and can make many accurate predictions about this environment. The same is true of other trades and professions. A predictive approach leads to a much broader conception of intelligence than the academic intelligence measured by IQ tests.

2.6 Definition of Intelligence

The discussion in the previous sections can be summarized as three hypotheses about natural and artificial intelligence:

H1. Prediction is the most important component of intelligence.

H2. Prediction and intelligence are relative to a set of environments.

H3. The amount of a system's intelligence varies with the number of accurate predictions that it makes in a set of environments.

These hypotheses lead to a new way measuring the amount of intelligence in natural and artificial systems, which is described in Section 4. The next section summarises some of the previous work on the measurement of intelligence.

3. Previous Work on the Measurement of Intelligence

3.1 Measurement of Human Intelligence

Over the last hundred years there has been a large amount of work on the indirect measurement of human intelligence. People have developed sets of tests that measure behavioural characteristics judged to be linked to intelligence. In the early days these tests included significant numbers of questions based on factual knowledge. 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 analysed 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.

3.2 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 and 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 ground-breaking 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). These measure the extent to which non-human animals exhibit human intelligence. They are not a meaningful measure of non-human animal intelligence and they do not enable us to compare general intelligence across species.

3.3 Measurement of Artificial Intelligence

Turing testing is often used to measure intelligence in artificial systems. The Turing test was originally proposed by Turing (1950) as a way of answering the question whether a machine could think. In the original version of the test, a human and a machine were connected to an electronic typing system and placed in a separate room. The human tester asked the two systems questions and had to decide which was the human and which was the machine. This test is extremely challenging for a machine to pass because the interrogator can ask questions about any topic. While claims have been made about AI systems passing constrained versions of the Turing test, our current AI systems are extremely far from passing the full version. Many variants of the Turing test have been proposed. These including embodied Turing tests (Harnad 1994), behaviour in game environments (Hingston 2009) and the Animal-AI Olympics (Crosby et al. 2019), which provides an environment in which artificial systems can attempt tasks that are believed to require intelligence in animals.

One problem with Turing testing 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.

Turing testing also relies on a clear definition of the human behaviours that require intelligence. In the past it was thought that chess playing was a paradigmatic example of intelligent behaviour and that any system that could play chess well would be highly intelligent. Now we know that low and medium ability chess systems can be built without much intelligence. We are also coming to realize how much of our intelligence is linked to our ability to understand and interact with the natural environment.

Turing testing also 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.

3.4 Universal Measures of Intelligence

To address the limitations of Turing testing, people have developed *universal* measures of intelligence that, in theory, can be applied 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 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 in all possible environments. A more practical goal-achievement/reward-based measure of intelligence has been proposed by Hernandez-Orallo and Dowe (2010).

One problem with goal/reward approaches to intelligence is that it is possible for a system to have intelligence without having any goals. The receipt of rewards from the environment is also conditional on having a particular set of actuators, so this definition conflates the physical abilities of a system with its intelligence. I have discussed these issues in more detail elsewhere (Gamez 2021).

Another universal measure of intelligence has been put forward by Chollet (2019), who defines intelligence in terms of skill acquisition efficiency. However, skills are difficult to define and depend on many factors that are not related to intelligence.

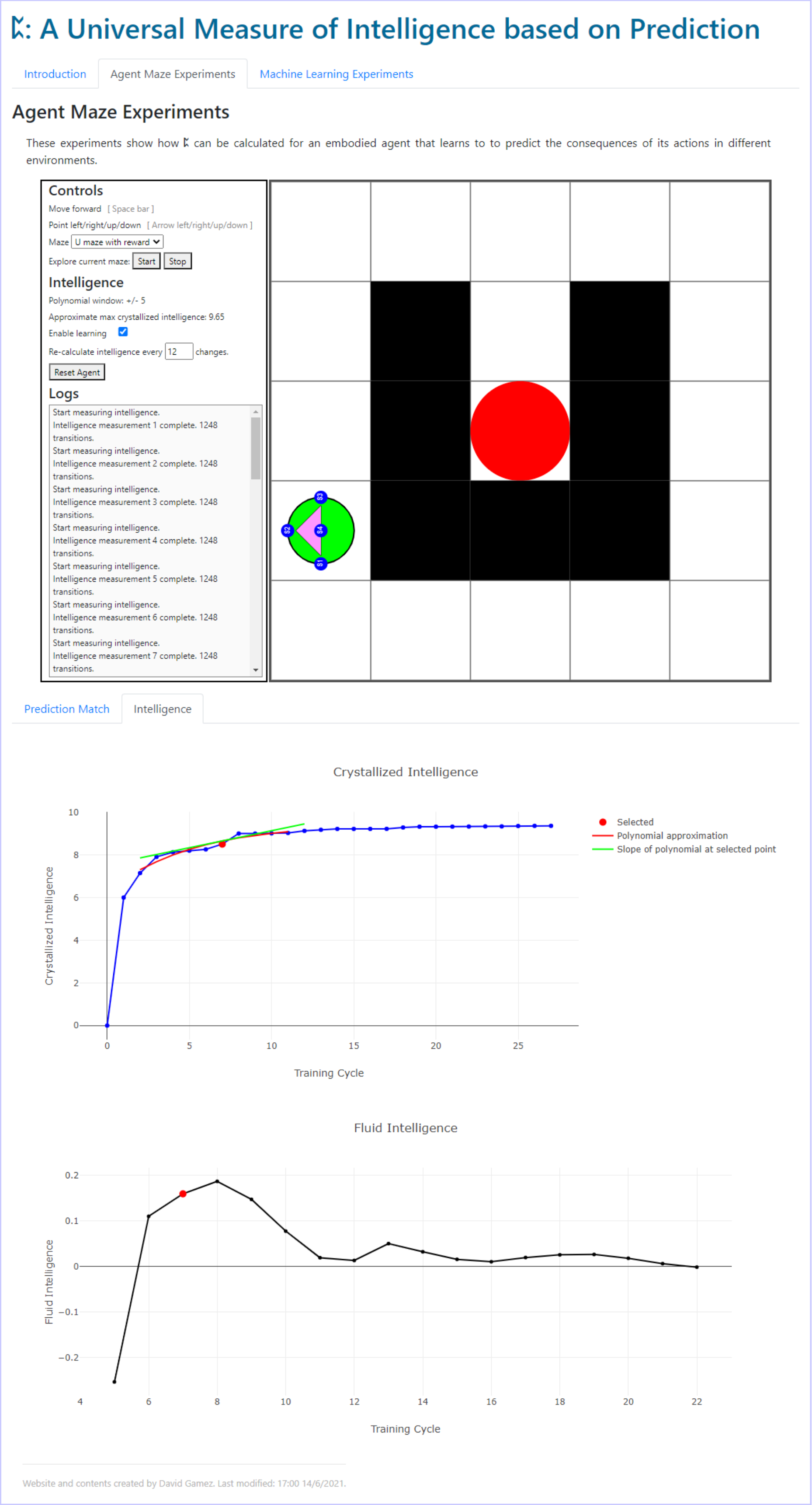
4. Algorithm for Measuring Predictive Intelligence

In my own work I have been developing a universal algorithm for measuring intelligence, which is based on the number of accurate predictions that the system makes in a set of environments. A compression algorithm compensates for differences in the complexity of predictions and environments. The algorithm also includes a logarithm that makes it easier to compare highly complex systems, such as humans, with trivial AI systems on the same scale. An early version of the algorithm along with the mathematical details is described in Gamez (2021). A paper on an improved version of the algorithm will be published soon.

The feasibility and performance of the algorithm have been tested on an agent in a variety of maze environments, and on a deep neural network that performs time series prediction. These experiments show that predictive intelligence can be measured in real time as systems learn about their environments. The experiments are implemented as a website: www.davidgamez.eu/pi, which is shown in Figure 1 and Figure 2.

These experiments show that it is straightforward to measure predictive intelligence on artificial systems, when we have full access to their internal states and their environments can be fully explored. More work is required to estimate the predictive intelligence of less accessible artificial systems that partially explore their environments.

We have very limited access to natural systems' internal states. Brain activity can be read non-invasively using fMRI, MEG and EEG, but these technologies have very low spatial and/or temporal resolution. Optogenetics can give us close to real time measurements of the entire brain of small transparent organisms, such as the Zebrafish larvae (Portugues et al. 2013) and it might be possible to apply this to other transparent animals, such as the glass octopus. With non-transparent animals we can only measure ~20,000 neurons on the surface of the brain in real time with current technology. Ways will have to be found to estimate the predictive intelligence of these systems from limited data and from external behaviour. We will then be able to systematically study and compare the intelligence of humans, non-human animals and artificial systems.



ꓘ: A Universal Measure of Intelligence based on Prediction

Introduction | Agent Maze Experiments | Machine Learning Experiments

Agent Maze Experiments

These experiments show how ꓘ can be calculated for an embodied agent that learns to to predict the consequences of its actions in different environments.

Controls

Move forward [Space bar]

Point left/right/up/down [Arrow left/right/up/down]

Maze: U maze with reward

Explore current maze: Start Stop

Intelligence

Polynomial window: +/- 5

Approximate max crystallized intelligence: 9.65

Enable learning ☑

Re-calculate intelligence every 12 changes.

Reset Agent

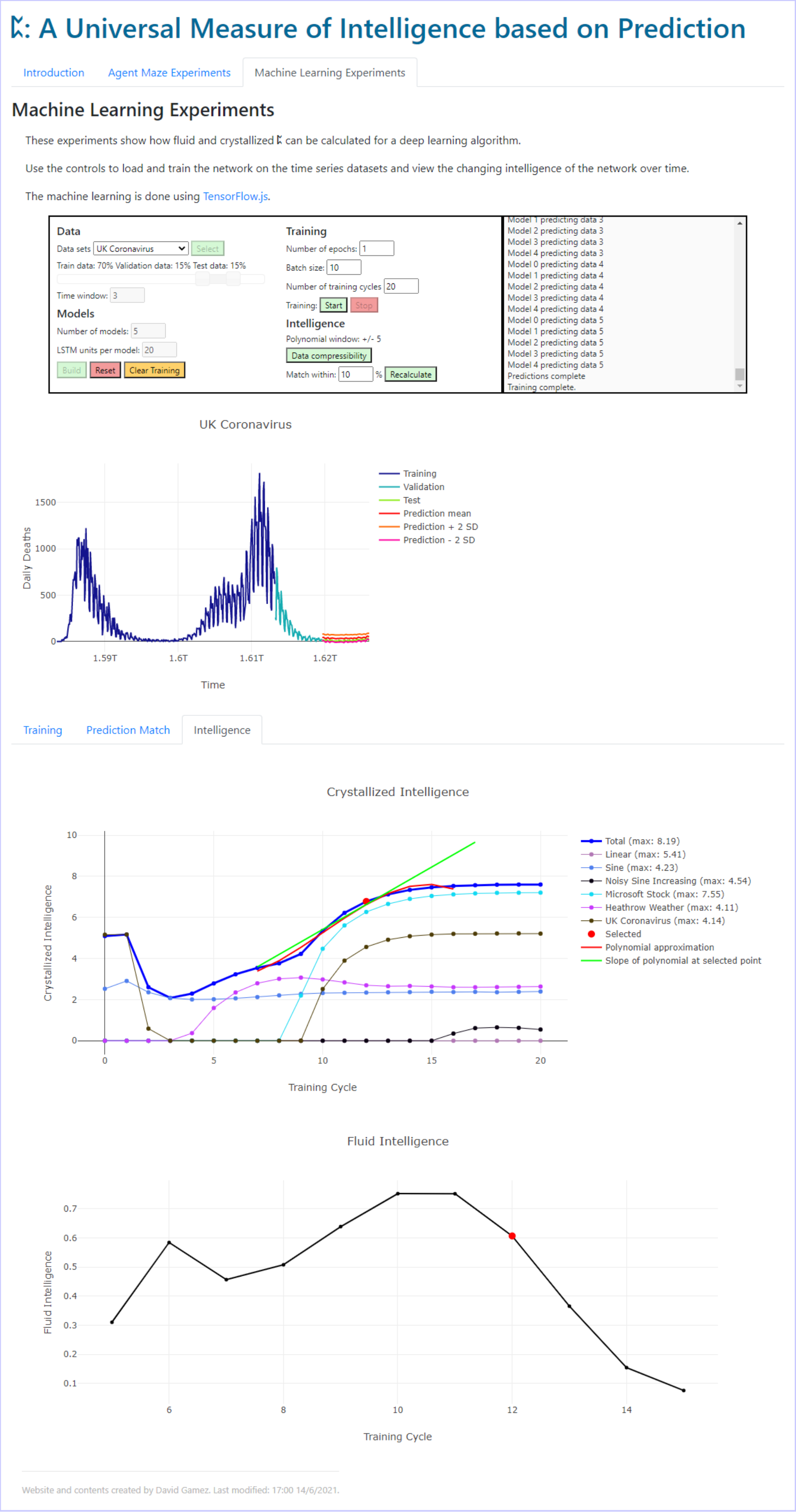
Logs

```
Start measuring intelligence.
Intelligence measurement 1 complete. 1248
transitions.
Start measuring intelligence.
Intelligence measurement 2 complete. 1248
transitions.
Start measuring intelligence.
Intelligence measurement 3 complete. 1248
transitions.
Start measuring intelligence.
Intelligence measurement 4 complete. 1248
transitions.
Start measuring intelligence.
Intelligence measurement 5 complete. 1248
transitions.
Start measuring intelligence.
Intelligence measurement 6 complete. 1248
transitions.
Start measuring intelligence.
Intelligence measurement 7 complete. 1248
transitions.
```

Prediction Match | Intelligence

Website and contents created by David Gamez. Last modified: 17:00 14/6/2021.

Figure 1. Agent maze experiments. The agent (shown in green) explores a variety of maze environments as its predictive intelligence is measured and displayed on the graphs at the bottom of the page.



Ꝁ: A Universal Measure of Intelligence based on Prediction

Introduction | Agent Maze Experiments | Machine Learning Experiments

Machine Learning Experiments

These experiments show how fluid and crystallized Ꝁ can be calculated for a deep learning algorithm.

Use the controls to load and train the network on the time series datasets and view the changing intelligence of the network over time.

The machine learning is done using TensorFlow.js.

Data

Data sets: UK Coronavirus

Train data: 70% Validation data: 15% Test data: 15%

Time window: 3

Models

Number of models: 5

LSTM units per model: 20

Training

Number of epochs: 1

Batch size: 10

Number of training cycles 20

Intelligence

Polynomial window: +/- 5

Match within: 10 %

```
Model 1 predicting data 3
Model 2 predicting data 3
Model 3 predicting data 3
Model 4 predicting data 3
Model 0 predicting data 4
Model 1 predicting data 4
Model 2 predicting data 4
Model 3 predicting data 4
Model 4 predicting data 4
Model 0 predicting data 5
Model 1 predicting data 5
Model 2 predicting data 5
Model 3 predicting data 5
Model 4 predicting data 5
Predictions complete
Training complete.
```

Website and contents created by David Gamez. Last modified: 17:00 14/6/2021.

Figure 2. Time series prediction experiments. A deep neural network is trained to predict future values of different time series and the system calculates its changing intelligence over time.

5. Predictive Intelligence and the Governance of AI

AI has many issues that could potentially be addressed through regulation, such as bias, lack of transparency and accountability. This section discusses how my definition and measure of intelligence could be used to address the risks that are posed to humans by artificial intelligence.

Many people have discussed the possibility that superintelligent AI systems could try to eliminate humans, possibly because they perceive them as a rival or threat. The Terminator films and Hal in 2001 Space Odessey are fictional examples of this scenario. The superior intelligence of these systems combined with their access to computer networks, weapons, etc. would give the AIs the edge over humans. A more benign possibility is that advanced AI systems would treat us in the same way that we treat animals – a scenario known as the gorilla problem (Russell 2019). AIs might keep us as pets or put us in a zoo, and we would not be able to do anything about it because of the AIs' superior intelligence and power.

AIs could also *accidently* eliminate or subjugate humans as they pursue another goal – a possibility known as instrumental convergence (Bostrom 2012). A simple example would be an AI system that was tasked with eliminating human suffering. A human might tackle this problem by trying to eliminate war, improving living conditions, etc. An AI with limited understanding of the human world might decide to eliminate humans (along with all of their suffering) instead. Bostrom (2016) gives an example of an AI that monopolizes the Earth's resources to maximize its paperclip production. AIs lack our deep understanding of the world and do not necessarily share our human values or emotional reactions. So AIs try to achieve their goals without understanding the negative side effects that their actions have on humans and the environment.

AI systems today are very far from posing any kind of existential threat - we are far more likely to destroy ourselves through poor programming or decision-making. However, some people think that we should take these threats seriously because of the possibility of a rapid intelligence explosion, in which AI systems develop more intelligent versions of themselves, which develop more intelligent version of themselves, and so on, leading to a rapid increase of artificial intelligence – a possibility known as the singularity (Chalmers 2010). An AI singularity would not be a problem if the resulting superintelligence was directed to solving the climate crisis or curing cancer. It would be a problem if it deliberately or inadvertently destroyed humanity.[1]

Machine learning algorithms often work as black boxes that learn from data and output classifications or predictions. The people using these algorithms do not know how they work and have little access to the internal representations that are autonomously developed by the system. This makes it hard to know what has happened when something goes wrong. The black box nature of AI algorithms also makes it difficult to detect bias. Some people think that the issues associated with AI superintelligence could be addressed by making AI more transparent – something known as 'white-boxing' AI. This would enable people to look inside the AI system to see if it was likely to pose a threat. However, white-boxing is unlikely to help much with superintelligent AI systems that are more complicated than the human brain. A human brain cannot understand something more complicated than itself, no matter how transparent the superior system is - it would be like a snail trying to understand the Internet.

[1] No one has come close to building a singularity machine. However, it is possible that a universal measure of intelligence, such as the one described in this paper, could form the starting point for a singularity machine. A genetic algorithm or reinforcement learning algorithm could use the amount of predictive intelligence as a feedback signal, which would enable it to develop increasingly intelligent versions of itself.

There has also been a considerable amount of discussion about how potentially dangerous AIs could be developed in restricted environments that are isolated from the physical world. This would reduce the ability of the AIs to access computer networks and minimize their ability to do damage to humanity (Bostrom 2016, Babcock et al. 2017). One problem with these proposals is that they do not define what they mean by artificial intelligence, so it is far from clear which systems should be contained. For the most part discussions about containment scenarios are highly theoretical and either end up with the AIs escaping (and potentially destroying humanity) or being so contained and cut off that they are effectively useless.

A theoretically informed definition of intelligence combined with an accurate method for measuring AI intelligence would be a much more promising way of addressing potential threats from superintelligent AIs. We can use a predictive definition of intelligence to precisely specify the set of systems that AI regulations apply to. We can then use a measure of predictive intelligence to limit the amount of intelligence in environments of concern. For example, if we want to prevent AI systems from compromising computer networks, then we need to measure their predictive intelligence in the areas of networking and cybersecurity. An AI with sub-human levels of predictive intelligence in these areas is very unlikely to pose a threat. If we want to prevent AI systems from manipulating people, then we need to measure their predictive intelligence in the environment of human decision-making and emotion. AIs only pose a threat to us when they exceed human capabilities in a sensitive environment. AIs with lower predictive intelligence than humans can be managed in the same way that we manage humans who attack computer networks or try to manipulate us.

White-boxing provides little protection against the negative consequences of superintelligence because human brains have a limited ability to understand something that is more complex than themselves. However, algorithms for measuring intelligence can be applied by a simpler system to a more complicated system. This enables us to monitor and regulate how much intelligence AIs have within particular environments without the requirement that the human brain has to fully comprehend the artificial system.

Many people are concerned that AI will take over their jobs and this has led to fears of mass unemployment and discussions about universal basic income. In the past the jobs lost to automation were replaced by new jobs. We do not know whether this will be the case with the jobs that will be lost to AI in the next 100 years. It is conceivable that governments will want to limit the amount of intelligence in particular environments to protect jobs (at the price of falling behind more technologically advanced competitors). In this case legislation could limit the amount of AIs' predictive intelligence in specific environments.

The definition of intelligence that I have proposed ends the confusion between classifiers and intelligent systems (see Section 2.3). This enables us to develop separate legislation that addresses the problems raised by artificial classifiers. For example, there is a lot of discussion about problems with bias or the misuse of face recognition technologies. These are not problems with artificial intelligence, but with artificial classification, and they should be addressed by legislation that regulates artificial classification systems and artificially intelligent systems that include classification technology.

6. Conclusion

It is impossible to govern artificial intelligence without a clear definition of intelligence and methods for measuring it. In this paper I have argued that prediction is the most important component of intelligence. Systems that can accurately see into the future are more intelligent because they have much more control over their environments, they can plan complex sequences of actions, and they can predict what other agents in their environment will do. This new understanding of intelligence leads to more accurate ways of measuring intelligence. This paper has briefly outlined the algorithm that I have developed that can measure predictive intelligence in humans, non-human animals and artificial systems.

This definition and measure of intelligence could play an important role in future attempts to regulate artificial intelligence. We could limit the amount of artificial intelligence in sensitive areas, such as cybersecurity. We can compare artificial intelligence with human intelligence. My measure of predictive intelligence is also a crude way of white-boxing AI by displaying its level of intelligence in particular environments. The distinction between artificial intelligence and artificial classification enables us to design separate targeted legislation for classifiers and intelligent systems.

References

Babcock, James, Janos Kramar, and Roman V. Yampolskiy (2017). Guidelines for Artificial Intelligence Containment. arXiv:1707.08476.

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

Bostrom, N. (2012). The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines* 22: 71-85.

Bostrom, N. (2016). *Superintelligence: Paths, Dangers, Strategies.* Oxford: Oxford University Press.

Chalmers, D. (2010). The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies* 17: 7-65.

Chollet, F. (2019). On the Measure of Intelligence. arXiv:1911.01547v2

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

Crosby, M., Beyret, B. and Halina, M. (2019). The Animal-AI Olympics. *Nature Machine Intelligence* 1: 257.

Gamez, D. (2019). The Intelligence of Sheep. *Animal Sentience* 25(27).

Gamez, D. (2021). Measuring Intelligence in Natural and Artificial Systems. Journal of Artificial Intelligence and Consciousness. Available from: https://doi.org/10.1142/S2705078521500090.

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

Harnad, S. (1994). Levels of Functional Equivalence in Reverse Bioengineering: The Darwinian Turing Test for Artificial Life. *Artificial Life* 1(3): 293-301.

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

Hingston, P. (2009). A Turing Test for Computer Game Bots. *IEEE Transactions on Computational Intelligence and AI In Games* 1(3): 169-86.

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.

Marino, L. and Merskin, D. (2019). Intelligence, complexity, and individuality in sheep. *Animal Sentience* 25(1): 1-26.

Portugues, R., Severi, K. E., Wyart, C. and Ahrens, M. B. (2013). Optogenetics in a transparent animal: circuit function in the larval zebrafish. *Current Opinion Neurobiology* 23(1): 119-26.

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.

Russell, S. J. (2019). *Human Compatible: AI and the Problem of Control*. USA: Penguin Random House.

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.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind* 59: 433-60.

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