BRAINTREE
nature of intelligence

The Three Laws of Artificial Intelligence

Dispelling Common Myths of AI

We've all heard about it and watched the scary movies. An artificial intelligence somehow develops spontaneously and ferociously like some exponentially brilliant cancer. We might start with something simple, but the intelligence improves itself out of our control. Before we know it, the whole human race is fighting for its survival.

It all sounds absolutely terrifying (which is why it makes good movie plotlines). But despite earnest commentators, philosophers, and people who should know better spreading these stories, the ideas are pure fantasy. The truth is the opposite: artificial intelligence – like all intelligence – can only develop slowly, under arduous and painful circumstances. It's not easy becoming clever.

There have always been two types of Artificial Intelligence: reality and fiction. Real AI is what we have all around us – the voice-recognising Siri or Echo, the hidden fraud detection systems of our banks, even the number-plate reading systems used by the police. Then there's the fictional ideas of AI that we watch movies about and hear commentators warn about – the super-intelligent general AIs that are mysteriously going to emerge from computer science labs and take over. Every research scientist is working on the former kind of AI. But because I am asked endlessly about the latter rather more fanciful ideas of AI, I'll focus on them here. There are several myths on this topic worth dispelling and I'll do this using what I shall call the three laws of AI.

Myth 1: A self-modifying AI will make itself super-intelligent.

Some commentators believe that there is some danger of an Artificial Intelligence "getting loose" and "making itself super intelligent". The first law of AI tells us why this will never happen.

FIRST LAW OF AI: Challenge begets intelligence.

Intelligence only exists in order to overcome urgent challenges. Without the right kinds of problems to solve, intelligence cannot emerge or increase. Intelligence is only needed where those challenges may be varied and unpredictable. Intelligence will only develop to solve those challenges if its future relies on its success.

To make a simple Artificial Intelligence, we create an algorithm to solve one specific challenge. To grow its intelligence into a general Artificial Intelligence, we must present ever-more complex and varied challenges to our developing AI, and develop new algorithms to solve them, keeping those that are successful. Without constant new challenges to solve, and without some reward on success, our AIs will not gain another IQ point.

AI researchers know this all too well. A robot that can perform one task well, will never grow in its abilities without us forcing it to grow. For example, the automatic number plate recognition system used by police is a specialised form of AI designed to solve one specific challenge – reading car number plates. Even if some process were added to this simple AI to enable it to modify itself, it would never increase its intelligence without being set a new and complex challenge. Without an urgent need, intelligence is simply a waste of time and effort. Looking at the natural world this is illustrated in abundance – most challenges in nature do not require brains to solve them. Only very few organisms have needed to go to the extraordinary efforts needed to develop brains. Even fewer develop highly complex brains.

The first law of AI tells us that Artificial Intelligence is a tremendously difficult goal, requiring exactly the right conditions and considerable effort. There will be no runaway AIs, there will be no self-developing AIs out of our control. There will be no singularities. AI will only be as intelligent as we encourage (or force) it to be, under duress.

Myth 2: With enough resources (neurons/computers/memory) an AI will be more intelligent than humans.

Commentators claim that "more is better". If a human brain has a hundred billion neurons, then an AI with a thousand billion simulated neurons will be more intelligent than a human. If a human brain is equivalent to all the computers of the Internet, then an AI loose in the Internet will have human intelligence. In reality, it is not the number that matters, it is how those resources are organised, as the second law of AI explains.

SECOND LAW OF AI: Intelligence requires appropriate structure.

There is no "one size fits all" for brain structures. Each kind of challenge requires a new design to solve it. To understand what we see, we need a specific kind of neural structure. To move our muscles, we need another kind. To store memories, we need another. Biology shows us that you do not need many neurons to be amazingly clever. The trick is to organise them in the right way, building the optimal algorithm for each problem.

Why Can't We Use Maths to Make AIs?

We do use a lot of clever maths and because of this some Machine Learning methods produce predictable results, enabling us to understand exactly what these AIs can and cannot do. However, most practical solutions are unpredictable, because they are so complex and they may use randomness within their algorithms meaning that our mathematics cannot cope, and because they often receive unpredictable inputs. While we do not have mathematics to predict the capabilities of a new AI, we do have mathematics that tells us about the limits of computation. Alan Turing helped invent theoretical computer science by telling us about one kind of limit – we can never predict if any arbitrary algorithm (including an AI) will ever halt in its calculations or not. We also have the "No Free Lunch Theorem" which tells us there is no algorithm that will outperform all others for all problems – meaning we need a new AI algorithm tailored for each new problem if we want the most effective intelligence. We even have Rice's Theorem which tells us that it's impossible for one algorithm to debug another algorithm perfectly – which means that even if an AI can modify itself, it will never be able to tell if the modification works for all cases without empirical testing.

To make an Artificial Intelligence, we need to design new structures/algorithms that are specialised for each challenge faced by the AI. Different types of problem require different structures. A problem never faced before may require the development of a new structure never created before. There is no universal structure that will suit all problems – the No Free Lunch Theorem tells us this (see box). Therefore, the creation of ever greater intelligence, or the ability to handle ever more different challenges, is a continual innovation process, with the invention of new structures required that are tailored to every new challenge. A big problem in AI research is figuring out which structures or algorithms solve

which challenges. Research is still in its infancy in this area, which is why today all AIs are extremely limited in their intelligences.

As we make our AIs cleverer (or if we ever manage to figure out how to make AIs that can keep altering themselves) we encounter yet more problems. We cannot design the intelligence in one go, because we have no mathematics to predict the capabilities of a new structure, and because we have insufficient understanding of how different structures/algorithms map to which challenges. Our only option in designing greater intelligences is an incremental, try and test approach.

For each new structure, we need to incorporate it into the intelligence without disrupting existing structures. This is an extremely difficult thing to achieve, and may result in layer upon layer of new structures, each carefully working with earlier structures – as is visible in the human brain. If we want an even cleverer brain like ours, we can also add in the ability of some structures to repurpose themselves if others are damaged – changing their structures until they can at least partially take over the role of lost functions. We have little idea how to achieve this, either.

The second law of AI tell us that resources are not enough. We still have to design new algorithms and structures within (and in support of) the AIs, for every new challenge that the AI faces.

Myth 3: As the speed of computers doubles every 18 months, AIs will exploit this computing power and grow exponentially cleverer.

Commentators claim that sheer brute calculating speed will overcome all challenges in the creation of AI. Use computers that are fast enough and an AI will be able to learn and out-think us. Since the speed of computer processors has been doubling approximately every 18 months for decades, this is surely an inevitability. Sadly, this point of view fails to recognise the impact of an opposing exponential that works as a significant brake on the development of AIs: testing.

THIRD LAW OF AI: Intelligence requires comprehensive testing.

Higher intelligence requires the most complex designs in the universe. But every tiny change made in an attempt to improve the design of an intelligence has the potential to destroy any or all of its existing capabilities. It doesn't help that we have no mathematics capable of predicting the capabilities of a general intelligence (see box). For these reasons, every new design of intelligence needs complete testing on all the problems that it exists to solve. Partial testing is not sufficient - the intelligence must be tested on all likely permutations of the problem for its designed lifetime otherwise its capabilities may not be trustable.

All AI researchers know this hard truth only too well: to make an Artificial Intelligence, it is necessary to train the AI and test all its capabilities comprehensively in its intended environment at every stage of its design. More than any other aspect, it is the process of testing that requires the most time. This time constraint produces a brake to the process of designing intelligence. At worst, the level of testing is exponential for each incremental gain in intelligence.

To understand why, imagine an intelligence that can distinguish two types of object using one feature – say colour – where it can recognise 10 different colours. In this case, the AI can understand at most 10 different kinds of item and classify them into two classes. If its capabilities are expanded to handle two features – say colour and 10 shades of brightness, then it can understand at most 100 different kinds of item. If it could handle 100 features, it can understand 10, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000 items. How do we know that the AI will classify them all correctly? This becomes a very important question for safety critical applications. Imagine the AI is driving your car and the two classes of object are "road" and "obstacle".

In reality it's much more complex. Most real intelligences will distinguish countless millions of items, using many millions of features, each of which could take thousands or millions of values. The testing of these intelligences takes a very long time indeed. This should be no great surprise. Evolution creates its biological brains by testing billions of brains in parallel in every possible scenario for thousands of millions of years, only permitting the successful brains to contribute to future brain designs. This is a highly efficient and effective method of testing that we can only dream of performing.

The tremendous need to test AIs has significant implications. We cannot design better AIs without testing them at each stage. We cannot make use of AIs in any safety-critical application until appropriate testing is performed. We need certification so that we know exactly how well an AI performs for well-defined tasks. We also cannot assume that an AI that has passed one test will continue to do so – like a human pilot, any AI that continues to learn must be continuously retested to ensure it remains certified. Finally, any human interaction with AIs also implies training and new certification for us. An automobile with an AI that takes control in some circumstances becomes a liability when the AI reaches the limits of its capabilities and the driver has not been trained to remain alert enough to take back control.

The third law of AI tell us that as intelligence increases, the time required for testing may increase exponentially. Ultimately, testing may impose practical limits to achievable artificial intelligence, and trustable artificial intelligence. Just as it becomes harder and harder to go faster as we approach the speed of light, it becomes harder and harder to increase intelligence as we build cleverer brains.

Conclusions

Artificial Intelligence has amazing potential to improve our lives, helping us live healthier, happier and generating new jobs. The creation of AI is one of the greatest scientific and engineering feats that we will ever undertake. It will be a new technological revolution. But this revolution will not magically happen on its own. The three Laws of AI tell us that we must slowly give more challenges to our AIs, carefully design new intelligent structures so that they can overcome these challenges, and perform massive testing to confirm that they can be trusted to solve the challenges. Thousands of skilled scientists and engineers are tirelessly following exactly these steps to bring us every tiny incremental improvement, for this is our design process and our scientific method. Do not be fearful of AI – marvel at the persistence and skill of those human specialists who are dedicating their lives to help create it.

February 2017

www.braintree.com