The limits of algorithms and implications for AI safety

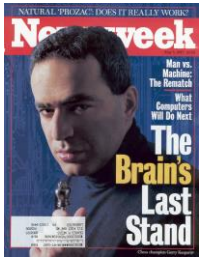
AI in Asia, Korea University Law School, Seoul

DECEMBER 16, 2016

STEVE WILSON (@STEVE_LOCKSTEP)

PRINCIPAL ANALYST, DIGITAL SAFETY & PRIVACY

constellation RESEARCH © 2010-2016 Constellation Research, Inc. All rights reserved.



Algorithms

"Algorithm" is surprisingly ill-defined!

Repeatable set of instructions; a *recipe*.

- Closed set of inputs.
- Deterministic outputs.

They say the world is "being eaten by software" and now that software seems to be coming alive. At long last, there is a broad debate about algorithms and modern society's dependence on them. Algorithmic trading. Web search. Machinery of government. Cultural biases. Transparency. Ownership. Have we put too much trust in algorithms?

"Algorithm" is surprisingly poorly defined but the first year computer science definition is fine: a repeatable set of instructions, like a recipe. The crucial thing is that an algorithm has a fixed set of inputs and it does a fixed set of things (creates a fixed set of outputs). One cannot "surprise" an algorithm. If a computer is programmed to make stock forecasts based on historic financial data, it cannot suddenly factor in the makeup of the board of directors.

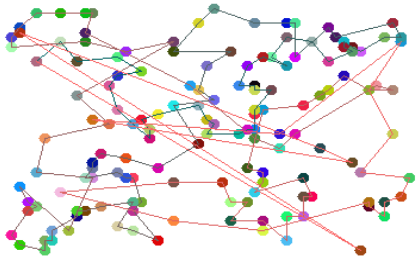
And as we shall see, first year computer scientists are taught some of the fundamental limits of algorithms – yet these lessons seem to get forgotten.

Game playing algorithms have long been a benchmark for intelligence for AI researchers. One of the most historically significant programs in the 1960s played backgammon, and beat the human world champion. Interestingly, the programmer had never played the game. Automating chess was always a much more serious challenge and it took decades before world class players were beaten. But when eventually (indeed inevitably) a World Champion lost to IBM's Deep Blue, some saw it as the beginning of the end of humanity.

But no, we did what humans do and what AI's can't: *we changed the rules*. We shifted our concepts of intelligence so that playing chess was no longer quite so special. A harder test turned out to be the more intuitive game of Go. But inevitably, that particularly human game was also automated, and recently history repeated itself, with Google's AlphaGo defeating the best human

The technical significance of AlphaGo for neural network technology and machine learning is great but the broader impact should not be exaggerated, for two reasons. First, we will change the rules again. And second, keep in mind how specialised these artificial intelligences really are.

A child could beat AlphaGo at chess. And IBM Watson can't drive a car.



Algorithms

Some problems *do not have algorithms*. E.g. The *Halting Problem*.

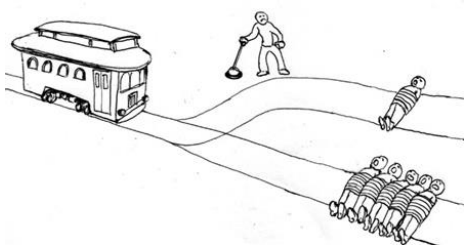
It has been known for decades with mathematical certainty that some very simply stated problems have no algorithmic solution. That is, we will never program a computer with a single method to solve such problems. This is not a limit of today's programming languages or architectures – it's fundamental. One example is the Halting Problem: it is not possible to create a algorithm that can examine any given program and tell if it will eventually halt. A whole branch of mathematics is devoted to the limits of *formal systems* – that is, processes that can be codified.

Self Driving Cars may decide life & death!

If certain really simple problems cannot be resolved by computers, then surely we must temper our expectations of mechanisms like driverless cars, which will be called upon to make life and death decisions.

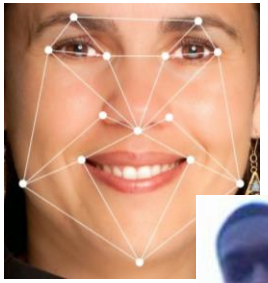
Consider court cases

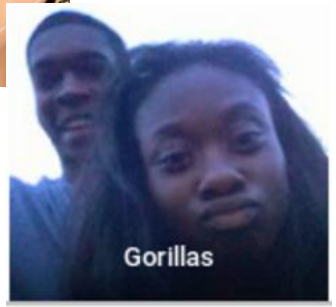
I don't say that computers can't drive – clearly they can and several now do. But we cannot expect a computer to deal with every eventuality in the real world. Many human decisions – especially when it comes to life and death – are unbounded; we cannot list all the important inputs in advance. Consider that court cases are *unpredictable*. No major legal action is ever the same. *Precedent* is a mainstay of the law and always relates to an unexpected input, something no algorithm could ever handle.

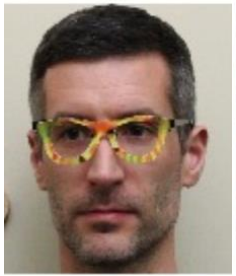


Some problems don't have answers! The Trolley Problem.

The Trolley Problem is a mainstay of driverless car reporting, but its true significance is often overlooked. The thing is, there is no resolution to the Trolley Problem. Reasonable people can disagree about the proper course of action in difficult situations. This style of dispute is rare in software development to date with its relatively simple problem spaces, but will increasingly feature in AI designs. The Trolley Problem is just the tip of a methodological iceberg.







Managing expectations

Lay people make simplistic assumptions about computers.

The way policy makers, lawyers and legislators are thinking about AI could be based on naïve assumptions about how these systems work.

Machine vision illustrations are obsolete. It's not how computers "think".

Consider object and face recognition, which lie on the cutting edge of commercial AI. Most of us have seen explainers which suggest computers interpret the world by extracting features from images, superimposing lattices and making measurements. But the latest machine vision and machine learning systems don't work like this anymore. The inner working of state-of-the-art AI is uncomfortably hard to trace. Deep Neural Networks mimic aspects of human mental machinery, but they are becoming inscrutable.

Neural Networks: surprising failures.

Neural network machine vision exhibits some startling failure modes.

- Object recognition in cloud photo service has scandalously misinterpreted certain races.
- Recent work at Carnegie Mellon has created artificially patterned goggles which fool face recognition algorithms into matching target celebrities, despite the patterns being nothing like facial features we recognise as such.
- A fatal crash of a Tesla was attributed to the vision system being confused by bright sunlight.

These cases of course point to shortcomings – including cultural biases – in the way AI systems are designed and tested but they also highlight just how alien AI can be. Yet the community is being asked to trust the behaviour of AIs as they become enmeshed in our human world. Now, human vision and human judgement are hardly perfect, and machines on average might eventually do better than us, but given the qualitative differences, measuring the improvement will be challenging, and the exceptions will continue to surprise us. We are not able to ask a machine to "Explain what you saw" much less "What were you thinking?"

Some questions

Is it ethical to proceed as if all tasks have algorithms?

If we know for sure some simple tasks have no algorithmic solution, how can we be sure that complex human behaviours can be neatly automated? Is there unexamined and unwarranted hubris at the heart of, for instance, self driving cars?

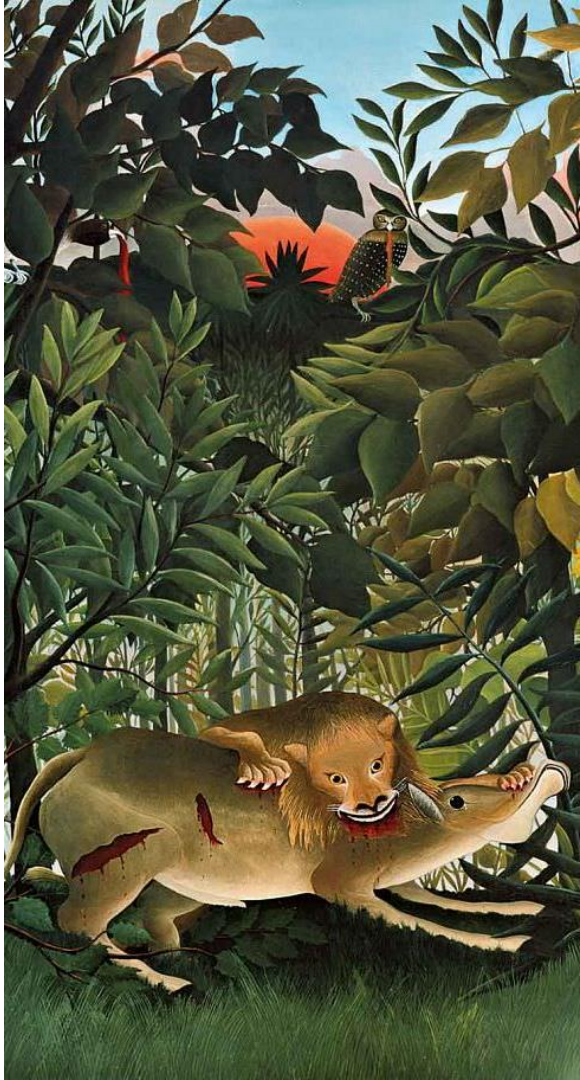
New types of random failure are coming.

We are accustomed to random failures of certain types. Mechanical components degrade over time in statistically predictable ways yet exceptions always occur, and material flaws can go undetected until it's too late. There are also "acts of god" for which no one is to blame. On the other hand, software errors are conventionally always due to a systematic design flaw, which in principle might have been detected or avoided. But complex algorithms running in complex environments might fail in truly unpredictable ways. Our paradigmatic ways of dealing with forensics and liability are likely to be stretched thin by the novelty.

How to interrogate a neural network?

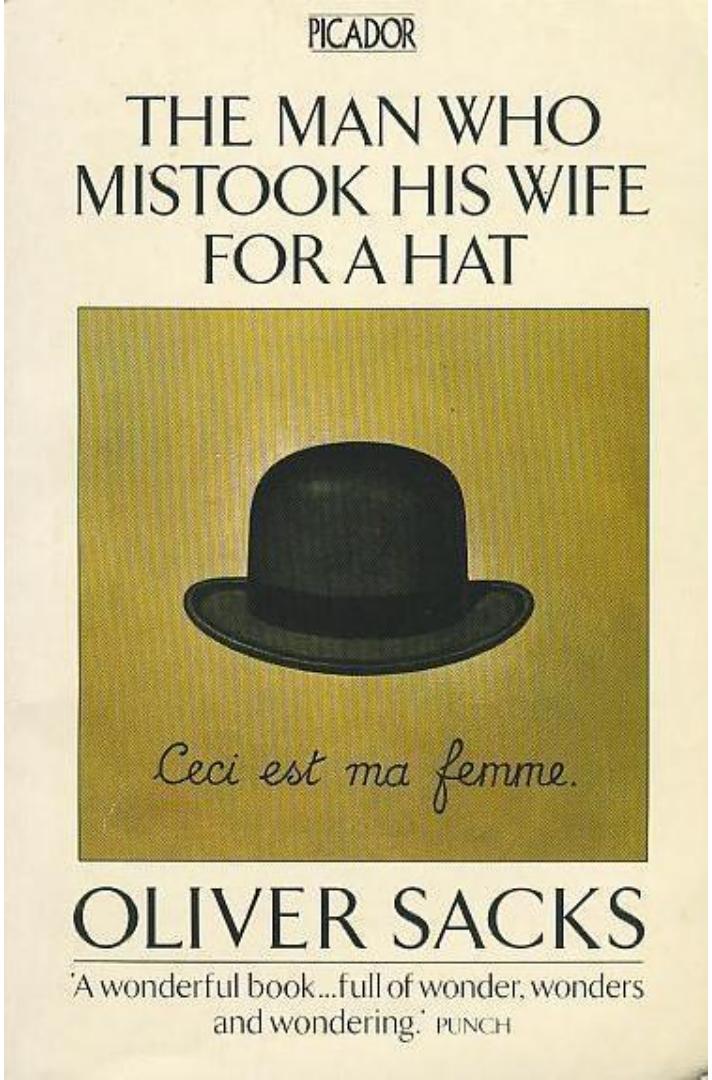
Neural networks do not proceed step by step to arrive at a decision. Instead they process an array of inputs through several layers of interconnected weights, and produce a more or less instantaneous result. There is no natural audit log or breakdown of why a particular result emerged. In the event of an unexpected answer, it is not clear if a rationale will be discoverable.

Moreover, artificial brains are unable to reflect or self-examine. They have no inner language with which they can explain their actions. Indeed, they have no inner life to examine as we understand it.

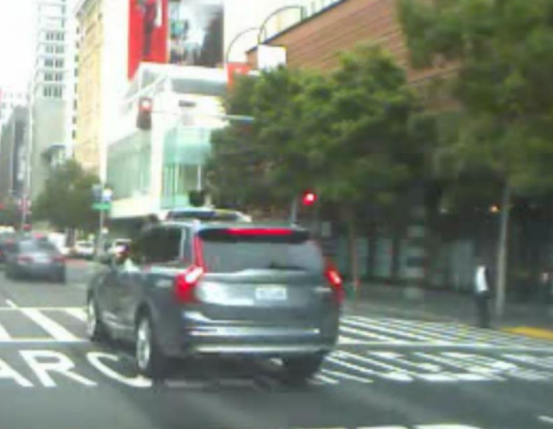


For a hundred years or more, most of what we knew about the workings of the brain came from studying rare, tragic and sometimes bizarre cases of brain damage. Cognitive researcher Oliver Sacks documented many of these in his compelling essays.

Similarly, we might learn a lot about AI from its failings, especially if neural networks are difficult to analyse. So should we plan for a pathology of AI?



Steve Wilson "The limits of algorithms and implications for AI safety"
AI in Asia, December 16, 2016,
Korea University Law School, Seoul



The future is bright but …

Do not expect machines to behave like us.

Regulators must prepare for surprises.

There isn't an algorithm for everything.

What if some tasks cannot be coded?

A self-driving Uber recently ran a red light in San Francisco and came close to striking a pedestrian. When you're teaching a teenager to drive, you will tell them not to run red lights but you probably won't have to say "don't hit people". How much general knowledge about the world goes into making a good human driver? Where does sound general knowledge come from?

We're still learning about learning.

Watch closely when a parent is teaching a child to ride a bicycle. Beyond the very basics like "hold onto the handle bars with both hands" there isn't much in terms of procedure that is spoken. Instead of explicitly teaching anything much at all, the parent instead *lets the child learn* to ride, by protecting them from harm and encouraging persistence.

It seems likely that most complex social tasks are learned rather than expressly taught. The way humans do this is not yet understood.

Machine learning will be one of the greatest developments in the digital world, but let's not underestimate the challenges, not oversimplify the regulatory paradigm shift. We cannot simply transfer centuries old societal compacts and ethical norms as if robots will behave just like us.

Steve Wilson

VP and Principal Analyst

+61 414 488 851

Steve@ConstellationR.com

@steve_lockstep

www.constellationr.com/users/

www.ConstellationR.com

constellation RESEARCH