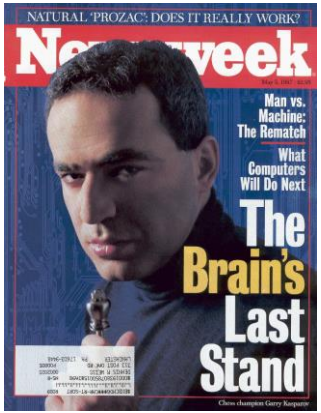
The limits of algorithms and implications for AI safety

AI in Asia, Korea University Law School, Seoul

DECEMBER 16, 2016

STEVE WILSON (@STEVE_LOCKSTEP)

PRINCIPAL ANALYST, DIGITAL SAFETY & PRIVACY

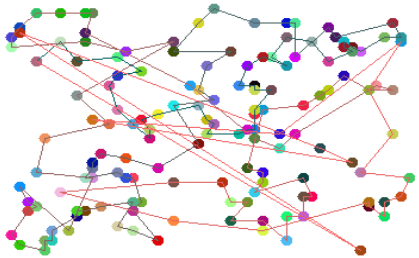


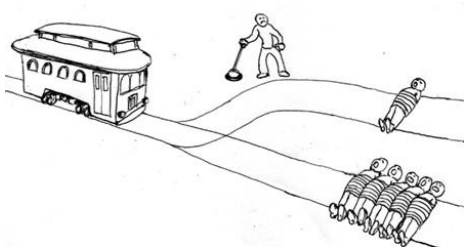
Algorithms

"Algorithm" is surprisingly ill-defined!

Repeatable set of instructions;
a *recipe*.

Closed set of inputs.
Deterministic outputs.





Algorithms

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

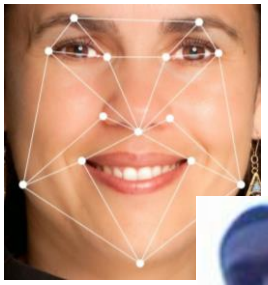
Self Driving Cars may decide life & death!

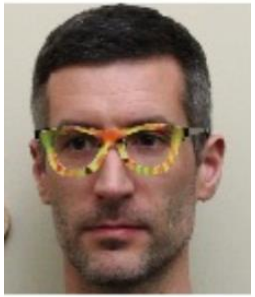
Consider court cases:

- *unbounded inputs*
- *unpredictable*.

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

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





Managing expectations

We're mired by simplistic assumptions about computers.

Machine vision illustrations are obsolete.

Tutorials mostly describe algorithms methodically mapping facial features, but it's not how computers "think" anymore. Neural networks work out for themselves what distinguishes faces.

Neural Networks: surprising failures.

- **Object recognition**

  Algorithms for labelling cloud photo services have infamously misclassified images of certain races.

- **Face Recognition**

  Researchers at Carnegie Melon University have fashioned goggles that trick neural algorithms into recognising celebrities. This type of failure mode is deeply counter intuitive. We humans have no idea what neural networks are seeing.

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

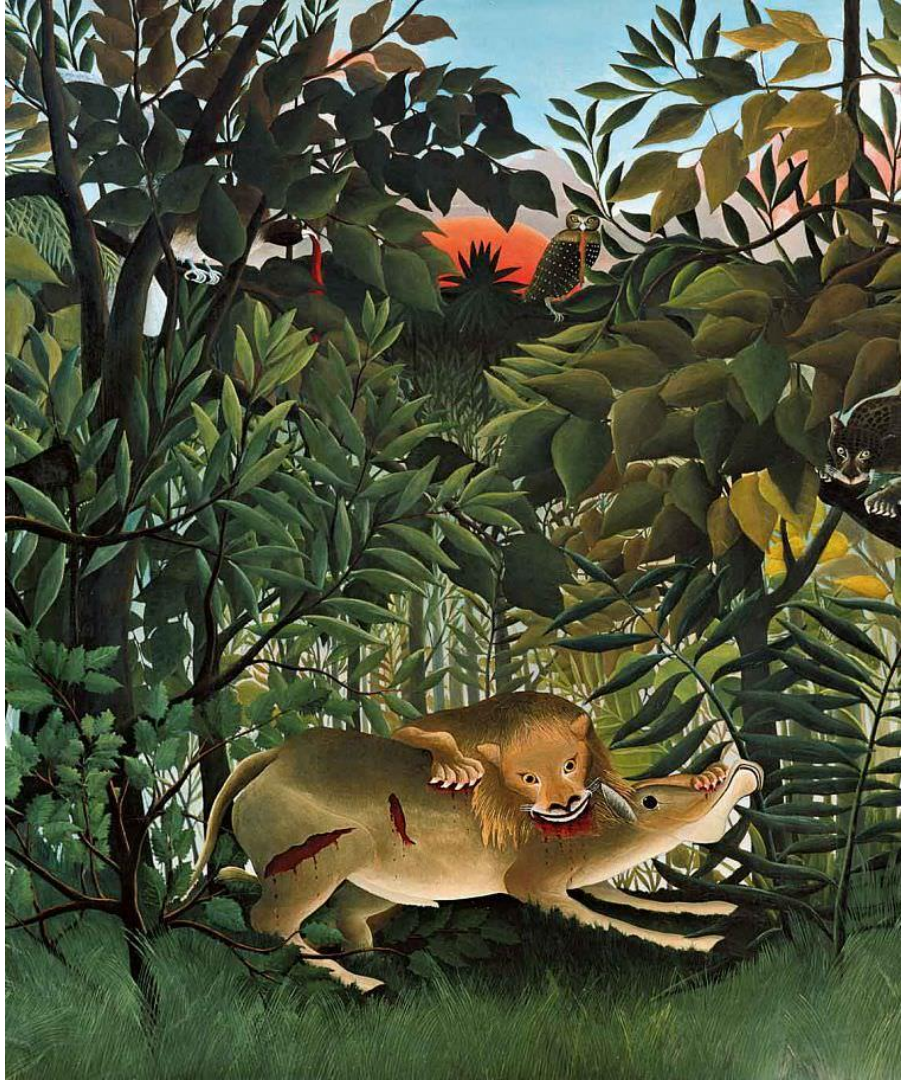
Some questions

How to interrogate a neural network?

Is it ethical to proceed on assumption that all tasks have algorithms?

What if a harm was not foreseeable
But *we knew it was not foreseeable!*

New types of random failure are coming.



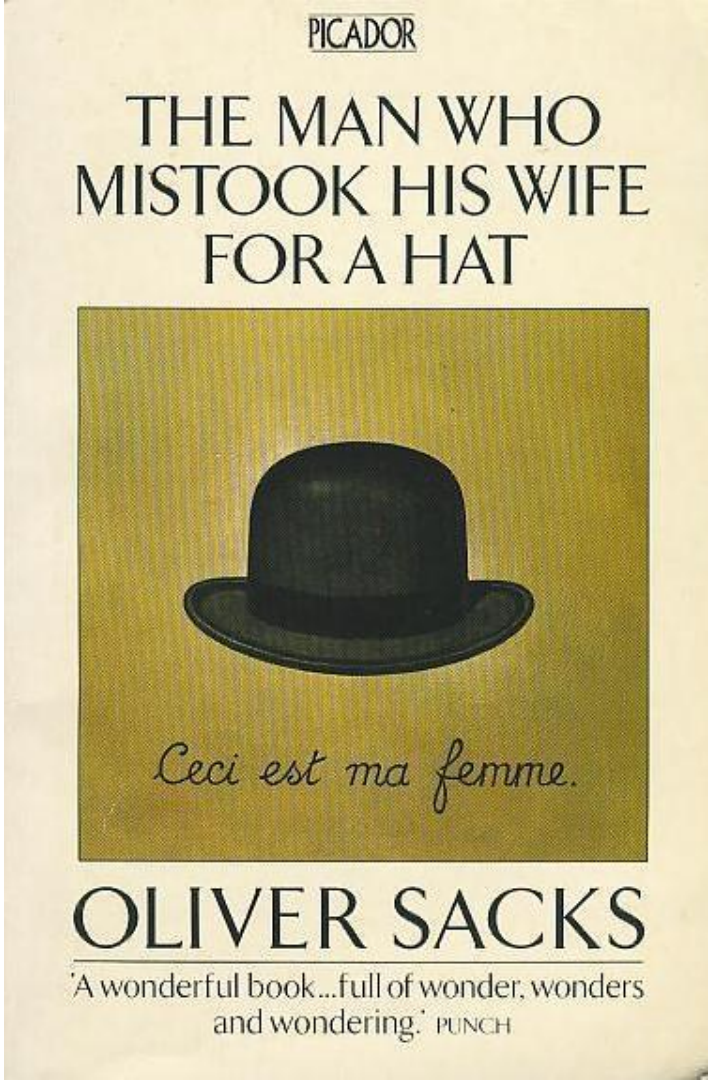
constellation RESEARCH

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

We will learn a lot about AI

from its failings!

Do we need a *pathology of AI*?



"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 tacit general knowledge about the world goes into making a good human driver?

We're still learning about learning.

> Think about your kid learning to ride a bike. Do you actually *teach* her how to ride? Or do you *let her learn* how to do it?

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

Steve Wilson

VP and Principal Analyst

+61 414 488 851

Steve@ConstellationR.com

@steve_lockstep

www.constellationr.com/users/swilson

www.ConstellationR.com

constellation RESEARCH