What's new in AI ethics since we last met

Alistair Knott

What we have so far. . .

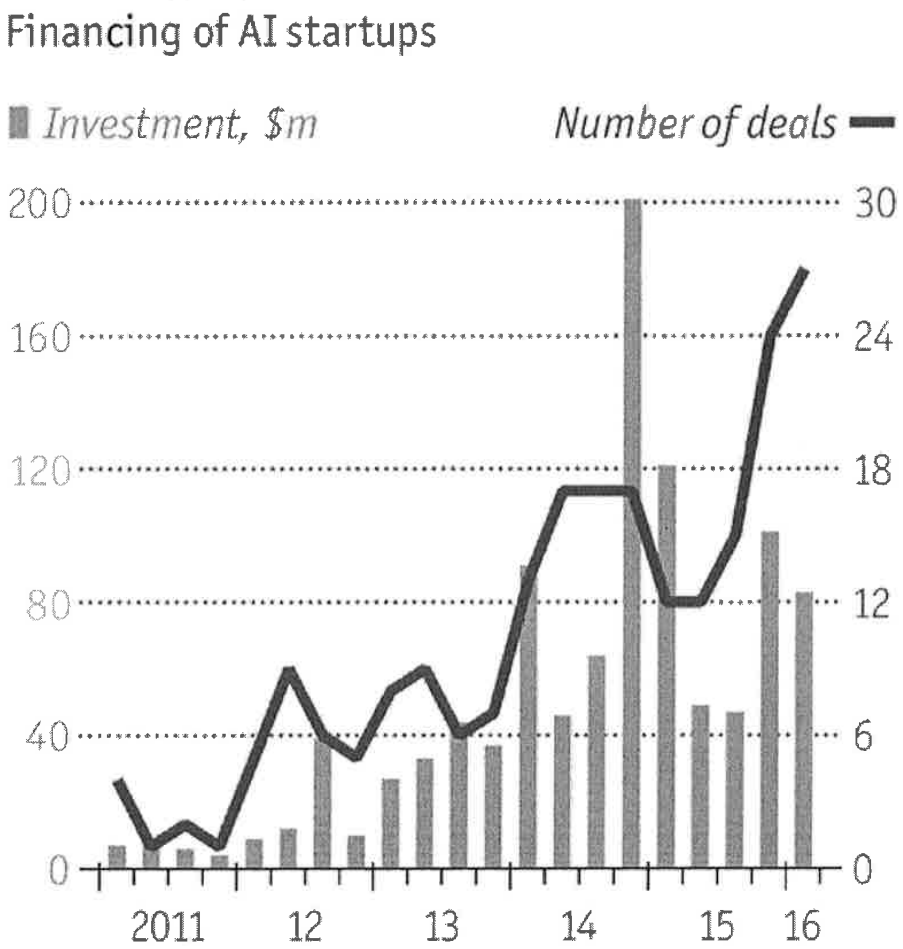
- Google/DeepMind's 'ethics board'
- Oxford's Future of Humanity Institute (Bostrom *et al.*)
  'AI safety' theme
- Cambridge's Centre for the study of Existential Risk
  (Tallinn, Huw Price)—'AI' theme
- Cambridge (MA)'s Future of Life Institute (Tallin again)
  Elon Musk funded a project to 'keep AI robust and beneficial'
- Leverhulme Centre for the Future of Intelligence (Cambridge,
  Oxford Martin School, UCL, Berkeley (Huw Price again)
- International Committee for Robot Arms Control (Noel Sharkey)
- Campaign to Stop Killer Robots

New Things

(1) The Economist's special report on AI

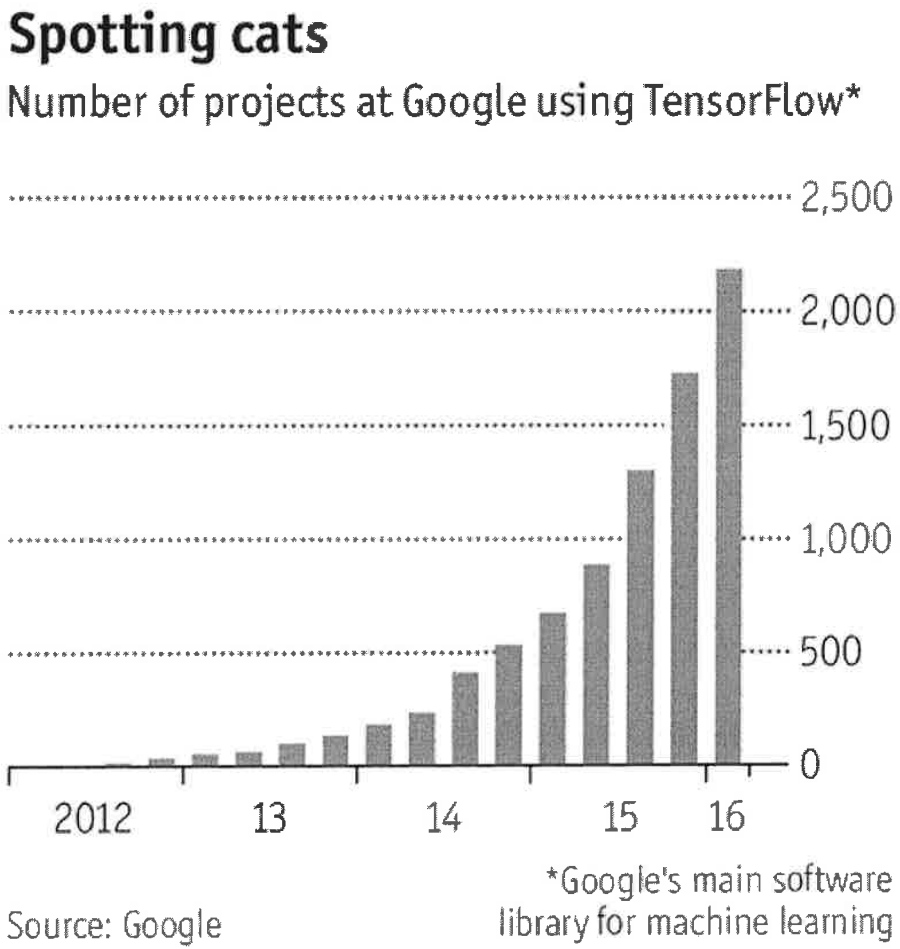
New Things

(1) The Economist's special report on AI



New Things

(1) The Economist's special report on AI



New things

(2) August 2016: Stuart Russell launches the Centre for Human-Compatible Artificial Intelligence.

- Based at UC Berkeley.

New things

(2) August 2016: Stuart Russell launches the Centre for Human-Compatible Artificial Intelligence.

- Based at UC Berkeley.
- Russell: quick to dismiss the prospect of ‘evil robots’.

New things

(2) August 2016: Stuart Russell launches the Centre for Human-Compatible Artificial Intelligence.

- Based at UC Berkeley.
- Russell: quick to dismiss the prospect of ‘evil robots’.
- The problems they study will be to do with what goals/motives we give to AIs. In particular: can we get robots to learn what *humans want*?

New things

(2) August 2016: Stuart Russell launches the Centre for Human-Compatible Artificial Intelligence.

- Based at UC Berkeley.
- Russell: quick to dismiss the prospect of 'evil robots'.
- The problems they study will be to do with what goals/motives we give to AIs. In particular: can we get robots to learn what *humans want*?
- 'In the process of figuring out what values robots should optimize, we are making explicit the idealization of ourselves as humans. As we envision AI aligned with human values, that process might cause us to think more about how we ourselves really should behave.'

New things

(3) Sep 2016: the first ‘AI100’ report was released.

- Launched in 2014 by Eric Horvitz (Stanford), as part of his presidency of the AAAI.
- Will report about the state of AI every 5 years, for the next 100yrs.

New things

(3) Sep 2016: the first ‘AI100’ report was released.

- Launched in 2014 by Eric Horvitz (Stanford), as part of his presidency of the AAAI.
- Will report about the state of AI every 5 years, for the next 100yrs.
- The reports will assess ‘the advances and influences of AI on people and society, and provide assessments and recommendations’, which will include ‘guidance on scientific, engineering, legal, ethical, economic, and societal fronts’.

New things

(3) Sep 2016: the first ‘AI100’ report was released.

- Launched in 2014 by Eric Horvitz (Stanford), as part of his presidency of the AAAI.
- Will report about the state of AI every 5 years, for the next 100yrs.
- The reports will assess ‘the advances and influences of AI on people and society, and provide assessments and recommendations’, which will include ‘guidance on scientific, engineering, legal, ethical, economic, and societal fronts’.
- Topics of interest include [everything].

New things

(3) Sep 2016: the first ‘AI100’ report was released.

- Launched in 2014 by Eric Horvitz (Stanford), as part of his presidency of the AAAI.
- Will report about the state of AI every 5 years, for the next 100yrs.
- The reports will assess ‘the advances and influences of AI on people and society, and provide assessments and recommendations’, which will include ‘guidance on scientific, engineering, legal, ethical, economic, and societal fronts’.
- Topics of interest include [everything].
- Overseen by a rotating standing committee. Current members: Barbara Grosz, Russ Altmann, Eric Horvitz, Alan Mackworth, Tom Mitchell, Deirdre Mulligan, Yoav Shoham.

New things

(4) Sep 2016: reports that researchers from the ‘big 5’ tech companies have been meeting to devise ethical guidelines relating to AI.

- That’s Google, Amazon, Facebook, IBM, Microsoft.

New things

(4) Sep 2016: reports that researchers from the ‘big 5’ tech companies have been meeting to devise ethical guidelines relating to AI.

- That’s Google, Amazon, Facebook, IBM, Microsoft.
- Details (& name) of this group ‘still to be worked out’.

New things

(4) Sep 2016: reports that researchers from the ‘big 5’ tech companies have been meeting to devise ethical guidelines relating to AI.

- That’s Google, Amazon, Facebook, IBM, Microsoft.
- Details (& name) of this group ‘still to be worked out’.
- The basic aim is ‘to ensure that A.I. research is focused on benefiting people, not hurting them’.

New things

(4) Sep 2016: reports that researchers from the ‘big 5’ tech companies have been meeting to devise ethical guidelines relating to AI.

- That’s Google, Amazon, Facebook, IBM, Microsoft.
- Details (& name) of this group ‘still to be worked out’.
- The basic aim is ‘to ensure that A.I. research is focused on benefiting people, not hurting them’.
- We know about this through four people involved ‘who are not authorized to speak about it publicly’.

Newly discovered things

(1) Dec 2015: Elon Musk founded ‘OpenAI’.

Newly discovered things

(1) Dec 2015: Elon Musk founded ‘OpenAI’.

- The idea: to recruit the best AI researchers out of existing companies, and get them to work in the public domain.

Newly discovered things

(1) Dec 2015: Elon Musk founded ‘OpenAI’.

- The idea: to recruit the best AI researchers out of existing companies, and get them to work in the public domain.
- ‘We could sit on the sidelines or we can encourage regulatory oversight, or we could participate with the right structure with people who care deeply about developing AI in a way that is safe and is beneficial to humanity.’

Newly discovered things

(1) Dec 2015: Elon Musk founded ‘OpenAI’.

- The idea: to recruit the best AI researchers out of existing companies, and get them to work in the public domain.
- ‘We could sit on the sidelines or we can encourage regulatory oversight, or we could participate with the right structure with people who care deeply about developing AI in a way that is safe and is beneficial to humanity.’
- Nick Bostrom: ‘If you have a button that could do bad things to the world, you don’t want to give it to everyone.’

Newly discovered things

Newly discovered things

(2) ‘Singularity University’ is organising a ‘summit’ in Christchurch in November.

Newly discovered things

(2) ‘Singularity University’ is organising a ‘summit’ in Christchurch in November.

- This is a Silicon Valley thinktank.
- ‘Our mission is to educate, inspire and empower leaders to apply exponential technologies to address humanity’s grand challenges.’

2nd proposed project: AI and employment

2nd proposed project: AI and employment

A worrying prospect is one where AI machines start to take people's jobs.

- Should there be legislation to control/prevent this?
- If it happens, should there be legislation to redistribute income?

2nd proposed project: AI and employment

Imagine some technologies that could replace humans in certain specific jobs:

- A dialogue-based computer tutor, in some area like 1st year health science.
- A dialogue-based language tutor, that can engage in open-ended L2 dialogues with a student.
- A dialogue-based ‘health consultant’.

2nd proposed project: AI and employment

Should there be special legislation governing the use of this technology by companies?

- Should it be restricted, if there are willing human participants? (On the lines of immigration law?)
- Should it be taxed? (E.g. using some analogue of social security for ‘electronic persons’?)
- What about Uni teachers’ role as the critic and conscience of society?