A survey of current discussions about AI and ethics/society

Alistair Knott

Google's 'AI ethics board'

Google's 'AI ethics board'

When Google bought DeepMind in 2014, a condition of sale was that Google would set up a committee to investigate the ethics of new AI technologies.

Google's 'AI ethics board'

When Google bought DeepMind in 2014, a condition of sale was that Google would set up a committee to investigate the ethics of new AI technologies.

This is 'still being worked out'. (p.c. from Google)

Google's 'AI ethics board'

When Google bought DeepMind in 2014, a condition of sale was that Google would set up a committee to investigate the ethics of new AI technologies.

This is 'still being worked out'. (p.c. from Google)

There *is* an ethics board, but we don't know who's on it.

Google's 'AI ethics board'

When Google bought DeepMind in 2014, a condition of sale was that Google would set up a committee to investigate the ethics of new AI technologies.

This is 'still being worked out'. (p.c. from Google)

There *is* an ethics board, but we don't know who's on it.

From a Guardian interview with Mustafa Suleyman, June 2015:

> *We will [publicise the names], but that isn't the be-all and end-all. . .*
>
> What gives Google the right to choose the board without public oversight?
>
> *That's just what I said to Larry [Page]. . .*

Future of Humanity Institute

- University of Oxford (Martin School: a research/policy unit with £100M of funding)
- Founded 2005
- Director: Nick Bostrom (philosopher)
- Members include K Eric Drexler
- AI safety is one of the themes.
- Flagship project is a ‘Strategic AI Research Centre’ (£1M of funding), announced July 2015

Centre for the Study of Existential Risk

- Cambridge University
- Founded in 2012 by Jann Tallinn (who also founded Skype), Huw Price (philosopher) Martin Rees (physics)
- AI is one of the themes
- AI mission: ‘to significantly advance the state of research on AI safety protocol and risk, and to inform industry leaders and policymakers on appropriate strategies and regulations to allow the benefits of AI advances to be safely realized’.

Future of Life institute

- ‘A volunteer-based organisation’ in Cambridge, MA
- Founded 2014 by Jann Tallinn
- AI is one of the themes
- July 2015: Elon Musk gave an $11M grant for projects that ‘keep AI robust and beneficial’. ($7M already allocated.)

The Leverhulme Centre for the Future of Intelligence

- Based at Cambridge University, with links to Oxford (Martin School) ICL, UC Berkeley (Stuart Russell)
- £10M funding announced December 2015
- Director will be Huw Price (a philosopher)

Conferences

AI Safety conference (Jan 2015)

- Organised by the Future of Life Institute
- A closed-door conference (but many people's slides are available)
- Delegates signed an open letter:
  - There is now a broad consensus that AI research is progressing steadily, and that its impact on society is likely to increase.
  - We recommend expanded research aimed at ensuring that increasingly capable AI systems are robust and beneficial.
  - Research priorities:
    Economics: labour market forecasting, market disruptions
    Law/ethics: autonomous vehicles, machine ethics, weapons
    Technical: verification, validity, security, control

  Over 8,600 signatures so far.

Conferences

AAAI Workshop on AI and Ethics (organised by Toby Walsh)

- 2014-2016

IJCAI panels on AI and ethics / ‘the future of AI’

- 2013 (organised by Stuart Russell)
- 2015

Organisations

International Committee for Robot Arms Control (ICRAC)

- An NGO drawing together roboticists, lawyers, ethicists, arms control experts
- Founded 2009
- Noel Sharkey (roboticist) is a key figure

Organisations

The Campaign to Stop Killer Robots

- An initiative from 9 NGOs (including ICRAC and Human Rights Watch)
- Founded 2012
- Focusses on banning certain classes of robot
- Participated at a UN meeting on ‘lethal autonomous weapons systems’ (2015)