Human Morality
Features and Bugs

Joshua Greene
Harvard University
Features:

What problem does morality solve?
How does it do it?

Bugs: How does morality go wrong?

Governing philosophies for AI

What moral thinking should we put into AI?
What moral thinking should we use to govern AI?
Features
The Tragedy of the Commons
The Tragedy of the Commons

Individual rationality vs. collective rationality
Me vs. Us

Morality: A suite of psychological features that allow otherwise selfish individuals to reap the benefits of cooperation
efficiency vs. flexibility

Fast Moral Machinery

Self-motivating

Positive
- Compassion
- Love, friendship
- Goodwill
- awe

Negative
- shame
- guilt
- Embarrassment
- fear

Other-motivating

- gratitude

- anger
- contempt
- social disgust

Fast Cooperation

Rand et al, Nature Comm., 2014
Fast Cooperation

Morality Fast and Slow
Trolleyology

Foot, 1978; Thomson, 1985

5 > 1

5 > 1
Dual-Process Morality

Integrative Moral Judgment

Utilitarian Assessment
Which action will produce better results?

Integrative Judgment
Which action do you find more morally acceptable?

Emotional Assessment
Which action do you feel worse about doing?

Shenhav & Greene, Journal of Neuroscience, 2014
Moral vision

Cf. “Scene Construction”
Hassabis et al., 2007
Hassabis & Maguire, 2007

Valuing Life

Number of lives

Probability you can save them

Shenhav & Greene, Neuron, 2010
Reward system Tracks “expected moral value”

“Neural substrate of prediction and reward”
Lessons of the “Moral Brain”

No distinctive “moral faculty”
No “ethical subroutine”

Morality unified at the functional level, not at the mechanical level
Bugs

1. Harming and Helping:

“A robot may not injure a human being or, through inaction, allow a human being to come to harm.”

2. Obedience

3. Self-Preservation
The Tragedy of Commonsense Morality

Us vs. Them
Singer (1972); Unger (1996)
Dear Sir,

Please send us $200...

Singer (1972); Unger (1996)
Distance: Us and Them

Singer, 1972; Unger, 1996

68%

34%

Musen & Greene, in prep
Numbers: Diminishing Moral Returns?

value

lives
Directness

Cushman et al., *Psychological Science*, 2006; Greene et al., *Cognition*, 2009
Numbers vs. Directness

Push vs. Switch
≈ 1,000,000 lives

Greene et al., 2009, Paxton et al., 2011
Governing Philosophies
Programming Ethics

Cf. Moor, 1985; Wallach & Allen, 2008

Deontic rules

Utilitarian calculations

Virtue acquisition

Humanoid Hybrids
Governing AI

The Ethics of AI is *Ethics*
Social Justice, Freedom vs. Regulation, ...

Metamorality: The original value alignment problem

Debugging morality with scientific self-knowledge
Debugging Morality Fast and Slow

If you trust people’s moral intuitions, you’ll get all of the bugs along with the features.

If you reject people’s moral intuitions, you’ll get unhappy people.

Need more sophisticated moral thinking, not just for leaders and engineers, but for everyone.