On autonomous cars, autopilots and the Anthropocene: ethics and AI verification

Dr Anders Sandberg
Future of Humanity Institute
Oxford Martin School
Oxford University
Robots have to live in the anthropocene

• Autonomous systems that matter will be interacting with the human world
• Verification and validation in this context involves not only understanding how the robot will interact, but how humans and their institutions will interact back.
• The human world is an ethical minefield. Just walking in a straight line is ill-advised.
• A hierarchy of “moral verification” challenges.
Challenge 1: safe and beneficial AI

- Creating systems that behave in such a way that human aims are fulfilled beneficially and safely
  - Requirements need to embody this, but experience and the SAI debate show it to be fundamentally hard
- Traditionally: safety, controllability
- Also: inspectability, technical debt
- Long run: corrigibility, inverse reinforcement learning
Challenge 2: “Moral” behaviour of machines

- Creating systems that act in the same way as a moral human would
- Engineering hesitancy for encoding morality in specifications
- “Malicious autonomy”: mislearning and misbehaving without intention
- The value learning and alignment problem
Challenge 3: keeping humans behaving/thinking right

• Creating systems so that humans are not driven to detrimental actions or belief states
• Autopilot bias
• Emergent accidental deception
• Cannot assume humans certified or know what they do.
Challenge 4: reducing systemic risk

• Avoiding systemic risks that emerge from the overall process
• The joint human-machine system can misbehave
• System design of open systems: always underspecified
  • Incentive design, principle design
Verifying ethics and the ethics of verifying

• Smart systems do not exist in a vacuum: the human context is an active part of what is going on.

• Autonomous, adaptive and open systems: strict verification and validation will be limited.

• Good-enough verification depends on loss function.

• Good-enough-ethics *also* depends on understanding the loss function.