How Can Robots Be Trustworthy?

Benjamin Kuipers
Computer Science & Engineering
University of Michigan

The Robot Problem

- Robots (and other AIs) will be increasingly acting as members of our society.
  - Self-driving cars and trucks on our roads and highways.
  - Companions and helpers for the elderly.
  - Teachers and care-takers for children.
  - Managers for complex distributed systems.
- How can we ensure that robots will behave well?
- How can we trust them?

A Robot is an Agent

- A robot is not simply a tool. It:
  - perceives the world,
  - builds a model,
  - selects an action to approach its goal, and
  - takes that action in the world.
- Its top-level goal is specified by humans.
  - It creates its own sub-goals.
- As agents, we want robots to be trustworthy.

The Deadly Dilemma
(née "Trolley Problem")

- A self-driving car drives down a narrow street with parked cars all around.
- Suddenly, an unseen pedestrian steps in front of the car.
- What should the car do?

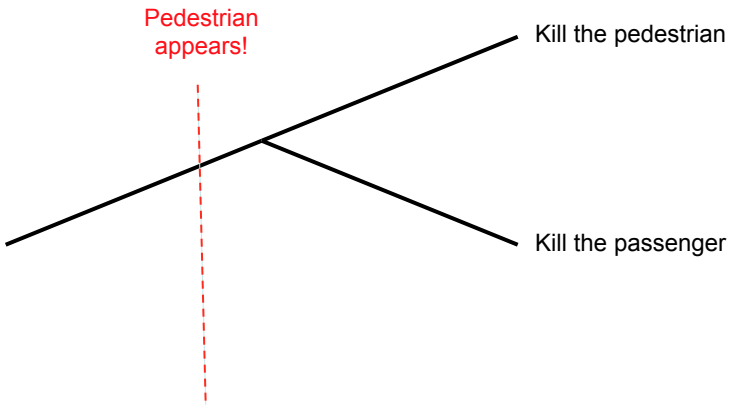
What should the robot do?

- Should the car take emergency action to avoid hitting the pedestrian?
- What if saving the pedestrian causes a serious collision, endangering or killing the passengers?
- What if the pedestrian is a small child?

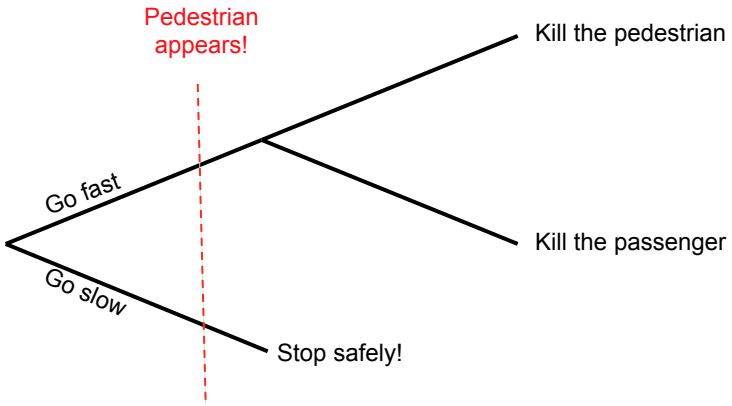
Who should the robot kill?

- The Deadly Dilemma is constructed to have only two answers, and both are bad.
- Should it kill the *pedestrian* or the *passenger*?
  - If it chooses to kill the *pedestrian*, why should the public trust self-driving cars to be on our roads?
  - If it chooses to kill the *passenger*, why would anyone ever trust the self-driving car enough to buy one?

The Deadly Dilemma is a Distraction

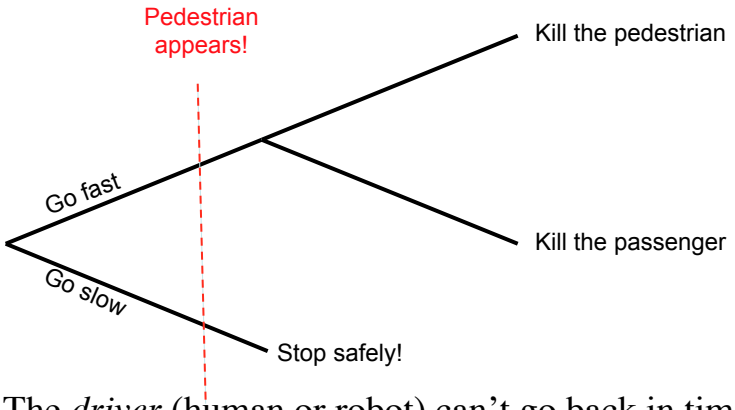


The Deadly Dilemma is a Distraction



- At a *previous* decision point, a different decision would have avoided the Deadly Dilemma entirely!

But Who Can Go Back in Time?



- The *driver* (human or robot) can't go back in time.
- But the *designer* can!
  - Detect the potential Deadly Dilemma.
  - Recognize the previous decision point, and act then.

Deadly Dilemmas are Rare

- The *designer* can anticipate the Deadly Dilemma, and identify the "upstream" decision to avoid it.
  - e.g., slowing down for "invisible pedestrians."
- The *driver* is far more likely to have "Near Miss" scenarios where catastrophe can be avoided.
  - "Near Miss" scenarios provide training experiences, teaching the driver how to respond well.
  - The driver can learn from deliberation and analysis to recognize the critical upstream decision points.

A Robot Must Earn Our Trust

- The self-driving car must show "practical wisdom."
  - Slow down where pedestrians could appear.
  - Steer to maximize visibility and warning time.
  - Show foresight and expertise at each start, stop, and turn.
- Trust is social capital to be accumulated.
  - The robot shows that it anticipates and avoids problems.
    - An avoided problem often looks like simple courtesy.
  - There is plenty of room for improvement in safety.
    - Currently, 94% of crashes involve driver error.

Trust

- *Trustworthiness* is a persistent property of an individual that an observer estimates.
  - *Trust* accepts vulnerability in order to cooperate, with confidence (based on the trustworthiness of the partner) that it will not be exploited.
- Estimating trustworthiness:
  - Trust may be given readily (depending on prior).
  - Trust is lost quickly based on negative evidence.
  - Trust is restored only slowly from positive experience.
- **Claim**: Trust is not expressible as utility maximization.

Back to Fundamentals

(Morality, Ethics, and Trust for Humans and Robots)

- An individual agent perceives its environment, and decides how to act.
  - Morality and ethics are sets of principles that constrain the behavior choices of individuals.
- It is tempting to think that morality and ethics are personal and individual.
  - This is not correct.
  - Society provides the moral and ethical principles.
  - Why?

More Fundamentals

- Unconstrained, individual decisions to maximize personal reward can lead to bad results, both for society and for the individuals involved.
  - Selfish reward maximizers exploit the vulnerability of potential partners.
  - Prisoners' Dilemma, Tragedy of the Commons, etc.
- Morality and ethics are provided by society to encourage trust and cooperation
  - by discouraging exploitation of vulnerability.

Benefits of Cooperation

- Individuals collaborate on larger projects with greater benefits.
  - Division of labor, pooled capital, economies of scale . . .
- Social invariants save resources.
  - e.g., don't kill, steal, or drive on the wrong side of the road,
  - Less need for protection and recovery.
- Cooperation produces more resources for society, so it has a better chance to survive and thrive.

A Few Clear Conclusions

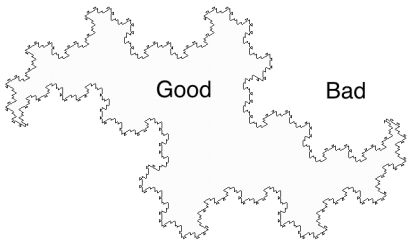
- The world is unboundedly complex.
  - Abstraction is necessary for practical inference.
- Moral and ethical reasoning takes place at several different time-scales.
- Moral and ethical reasoning involves several different representations for knowledge.

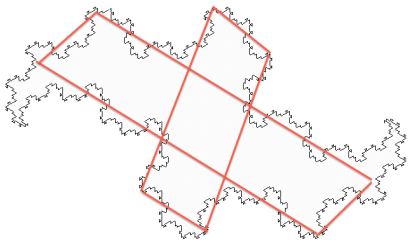
Unbounded Complexity

- The complexity of the physical and social world is essentially unbounded.
  - A core problem for an intelligent agent (human, animal, or robot) is to cope with that complexity.
- Tractable reasoning requires abstraction.
  - Intelligent agents have limited inference capabilities.
    - We can do a lot of very simple computations.
    - Or a few more complex computations.
- Ethical reasoning requires abstraction.
  - How to abstract that complexity is part of the ethical decision, not prior to it.

Unbounded Complexity

- Metaphorically:





  - Abstractions are approximations.
    - Useful, but never perfect.
  - Finding the right abstraction is part of the problem.
    - Seldom a good assumption to start with.

Time-Scales for Moral Decisions

- Moral decisions take place at multiple time-scales:
  - **Fast**: Rapid response to urgent situations;
  - **Slow**: Deliberative reflection on less urgent situations, as well as explaining and evaluating the outcomes of previous decisions;
  - **Slower**: Gradual evolution of prevailing social norms.
- This has been widely observed:
  - Kahneman, *Thinking, Fast and Slow* (2011)
  - Haidt, *The Righteous Mind* (2012)
  - Greene, *Moral Tribes* (2013)

Representations for Moral Knowledge

- Major theories of philosophical ethics suggest different AI knowledge representations.
  - **Deontology** ("*What is my* duty*, to do, or not to do?*")
    - Pattern-matched rules and constraints
  - **Utilitarianism** ("*What action maximizes* utility *for all?*")
    - Special case of **consequentialism** ("*What action has the best* consequences *for all?*")
    - Decision theory / Game theory
  - **Virtue Ethics** ("*What would a* virtuous *person do?*")
    - Case-based and analogical reasoning
- These are different aspects of a more complex reality. (*The Blind Men and the Elephant*)

These Pieces Fit Together

- In a world of unbounded complexity, an agent (human or robot) must make urgent decisions.
  - Sometimes, those decisions are wrong, perhaps because of applying the wrong abstraction.
  - Errors are opportunities for learning.
  - Learning has benefits at longer time-scales.
  - Multiple representations are needed to express different abstractions, to meet different requirements.

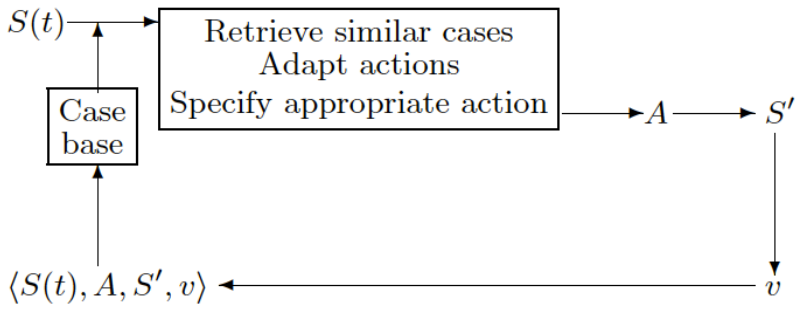
Problems to Solve

- **Form**:
  - How is moral and ethical knowledge
    - expressed in different representations and
    - used at different time-scales?
- **Content**:
  - What moral and ethical principles should we actually build into a robot?
  - Who gets to decide?

Cases Represent Experience

- A *situation* $S(t)$ is a rich (very high information content) description of current experience.
  - Case-based reasoning typically represents cases with *propositional feature vectors*.
  - Analogical reasoning typically represents cases with *first-order object-relation descriptions*.
- A *case* $< S, A, S', v >$ describes experience:
  - an initial situation S
  - the action A taken in that situation
  - the resulting situation S'
  - the valence v, evaluating the outcome of the action

Case-Based (or Analogical) Reasoning



- Given a situation $S(t)$, retrieve best matching cases.
- Synthesize the best action A.
  - Approach the good (v) cases and avoid the bad ones.
- Observe the outcome S', and its valence v.
- Add a new case to the knowledge base.

Which Moral Principles?

- What should the principles be? E.g.:
  - Protect your group.
  - Do not harm people.
  - Respect your elders and superiors.
  - Tell the truth.
  - Respect property ownership.
  - Respect social norms.
  - Do unto others as you would have them do unto you. [The Golden Rule]
  - . . .
- How do we evaluate these, and decide?
  - These have different meanings in different representations.

Moral and Ethical Variation

- Morality and ethics vary substantially across human societies.
  - Different cultures and subgroups in our world.
  - Societies change over historical time.
- Morality changes and evolves with society.
  - Singer, *The Expanding Circle* (1981)
  - Pinker, *The Better Angels of Our Nature* (2011)
  - Norenzayan, *Big Gods* (2013)

Who Decides?

- Who should decide the moral and ethical principles that a robot will follow?
  - The owner? The manufacturer? The designer?
  - Microsoft's Tay fell in with bad companions, and learned to spread and defend despicable racist beliefs.
  - Robots do not (yet?) have rights to self-determination.
- Remember: a poor choice could undermine the cooperation that society depends on.

References

- Bacharach, Guerra & Zizzo. The self-fulfilling property of trust: An experimental study. *Theory and Decision*, 2007.
- Greene. *Moral Tribes: Emotion, Reason, and the Gap between Us and Them*, 2013.
- Haidt. *The Righteous Mind: Why Good People are Divided by Politics and Religion*, 2012.
- Johnson & Mislin. Trust games: A meta-analysis. *J. Economic Psychology*, 2011.
- Kahneman. *Thinking, Fast and Slow*, 2011.
- Kuipers. Toward morality and ethics for robots. *AAAI Spring Symposium on Ethical and Moral Considerations in Non-Human Agents*, 2016.
- Kuipers. How can we trust a robot? *CACM*, to appear.
- Lin, Abney & Bekey. *Robot Ethics: The Ethical and Social Implications of Robotics*, 2012.
- Norenzayan. *Big Gods: How Religion Transformed Cooperation & Conflict*, 2013.
- Pinker. *The Better Angels of Our Nature: Why Violence Has Declined*, 2011.
- Rousseau, Sitkin, Burt & Camerer. Not so different after all: a cross-discipline view of trust. *Academy of Management Review*, 1998.
- Singer. *The Expanding Circle: Ethics, Evolution, and Moral Progress,* 1981.
- Wallach & Allen. *Moral Machines: Teaching Robots Right from Wrong*, 2009.

Michigan Unemployment Insurance Fraud Computer System has 93% error rate

- MIDAS made 22,427 findings of fraud and assessed penalties **without human involvement**. (2013-2015)
  - The people accused lost unemployment benefits, and faced penalties up to 400%, aggressive collection methods, and garnished wages and tax refunds.
- On review, 20,965 of these findings were false.
- Another 31,206 cases had some human involvement.
  - After checking 7,000 of these, the rate of false findings is "about the same."
- These are under review, with some restitution.
  - The situation remains in flux. (1-2017)
  - The money collected has been used to balance the budget.