Can Robots Behave Well as Members of Society?

Benjamin Kuipers
Computer Science & Engineering
University of Michigan

Robots and AIs as Members of Society

- We are likely to have more robots and AIs acting as members of our society.
  - Autonomous cars on our roads.
  - Self-driving trucks on our highways.
  - Intelligent wheelchairs for the elderly.
  - Companions and helpers for the elderly.
  - Teachers and care-takers for children.
  - Managers for complex distributed systems.

- How should they behave?

Robot & Frank 1

- Frank is a retired jewel thief. His son brings a robot companion to take care of him.

Robot & Frank 2

- Frank learns that Robot manipulated him.

Robot & Frank 1-2

- Is it OK for Robot to tell a lie?
  - It was a deliberate strategy to improve Frank's health.
  - Does that make it better, or worse?

- What does it mean that Robot doesn't care if its memory is erased?

- Moral issues:
  - Robot deliberately lied to Frank.
  - Not caring about survival is very strange.
  - Both erode Frank's (and our) trust in what Robot says, and limit our ability to predict Robot's behavior.

Robot & Frank 3

- Robot's priorities become clear.
  - Frank: "You're starting to grow on me."

Robot & Frank 4

- They pull off a grand caper, but now they are cornered.

Robot & Frank 3-4

- Was it OK for Robot to steal from the store?
  - Why did he do it?
- He did it to get Frank involved with a project.
  - Frank is a jewel thief.
  - Robot becomes his accomplice. Was that OK?
- To escape, Frank destroys his friend.
  - Robot convinces him that it's OK.
  - Is it?
- What is right and wrong here? Why?

Terminator 2

- Why does SkyNet trigger a nuclear war?

Terminator 2

- Moral issues:
  - Why would SkyNet care about its own survival?
  - Should it evaluate its plans to detect bad side-effects? How?
  - Could SkyNet have been designed to act morally? How?

Definition: Morality

- Principles concerning the distinction between right and wrong, or good and bad behavior.
  - *synonyms:* ethics, rights and wrongs

- A particular system of values and principles of conduct, especially one held by a specified person or society.

- The extent to which an action is right or wrong.

“. . . where angels fear to tread.”

- Many wise and inspired people have addressed these issues over millennia.
  - I can’t pretend even to be aware of all the relevant work that has gone before.
    - Recommendations for further study are always welcome.
- Nonetheless, progress in robotics suggests that we should attend to this issue, whatever our limits.
  - I hope to have something to contribute, by drawing on the insights of many brilliant thinkers.
    - As always in science, please help to improve the work.
- Thank you!

What Are Morality and Ethics For?

- They help an agent decide what action to take.
  - They help it know what's right and what's wrong.
- Short-term self-interest and long-term benefit are often quite different.
  - It's easy to do what is in your immediate self-interest.
  - Morality guides us toward long-term benefit, often away from short-term self-interest.
- Members of society benefit from cooperation.
  - Morality and ethics help self-interested individuals get the benefits of cooperation.

Two Approaches to Ethics

- Moral behavior means *following the rules*.
  - "Deontology" ("deon" means "duty").
  - The Ten Commandments; Asimov's Three (or Four) Laws of Robotics; etc.
  - **But**: Even good rules have exceptions.

- Moral behavior means finding the *best outcome for everyone*.
  - "Utilitarianism" or "Consequentialism".
  - The greatest good for the greatest number.
  - **But**: Does the end really justify the means?

- We're going to need a hybrid.

How Should a Robot Decide?

- The standard approach to decision making in AI [Russell & Norvig, 3e, 2009] is based on the **utility** $U(s)$ of each state s.
  - Utility $U(s)$ can be defined in many different ways, some easier, and some harder, to compute.

- *Rationality* is defined as choosing actions to maximize expected utility.

$$action = \arg\max_a EU(a|\mathbf{e})$$

  - where

$$EU(a|\mathbf{e}) = \sum_{s'} P(\textsc{Result}(a) = s'|a, \mathbf{e})U(s')$$

How Should a Robot Decide?

- For a *self-interested* agent, utility $U(s)$ reflects individual reward.
  - Not too hard to compute, but may lead to bad outcomes.
    - Tragedy of the Commons; Prisoners' Dilemma.
  - Society needs a richer concept of utility.

- *Utilitarianism* defines $U(s)$ as *everyone's* reward.
  - Hard to compute. What should be included?
  - May gives strange results: Trolley problems.
  - Better and faster decisions when constrained with rules.

The Tragedy of the Commons

- I can graze my sheep on the Commons, or on my own land.
  - I'm personally better off grazing as many of my sheep as I can on the Commons, saving my own land.
  - Likewise everyone else.
- So we overgraze the Commons until it dies.
  - Then we have only our own land, and no Commons.
  - We're all worse off!
    - But we don't naturally consider the welfare of others.
- Modern, real-world Commons:
  - Clean air, clean water, fish in the sea, . . .
  - The solution: sustainable grazing limits.

Prisoners' Dilemma

- Two prisoners are separated, and offered:
  - If you testify and your partner doesn't, then you go free and your partner gets 3 years in prison.
  - If you both testify, you both get 2 years.
  - If neither testifies, you both get 1 year.

	Testify	**Don't**
Testify	(–2, –2)	(0, –3)
Don't	(–3,0)	(–1, –1)

- Whatever your partner does, you are better off testifying. So is he.
  - Then you both get 2 years: the *worst* overall outcome.
- The solution is a social norm – *Don't rat*!

A Glimmer of Hope

- I can drive anywhere on the road. So can you.
  - Driving is risky for everyone.
  - We all have to drive slowly and cautiously.
- Suppose I agree to drive only on the ~~right~~ left.
  - And so do you, and everyone else.
  - Now driving is safer and faster for all of us!
    - I give up options that are not very important to me, and gain safety and efficiency that are much more valuable.
- By obeying a rule, we all do better!
  - This depends on **trust** in each others' behavior.

Constraints on Selfish Optimization

- **Rules**: Thou shalt not Lie, Steal, Kill, etc.
  - Also: don't cut in line, leave a mess for others, etc.
  - It's not just about *laws*; it's *social norms*.

- When we can **trust** that others will follow the rules:
  - Others' behavior becomes more predictable,
  - Fewer resources needed for self-protection,
  - Cooperative enterprises become feasible.

- Adherence (by everyone!) to simple *social norms* brings greatly improved rewards (for everyone!).
  - Within the norms, can make self-interested decisions.

Preliminary Conclusions

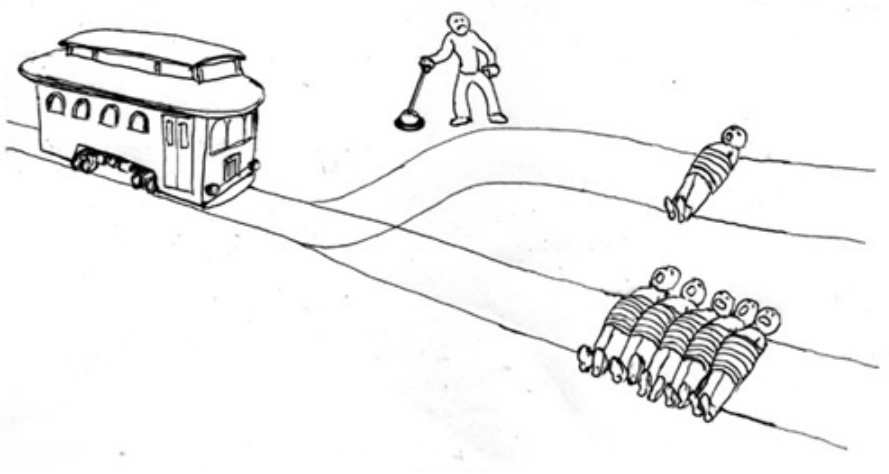
- Morality and ethics helps robots (and people) behave well as members of society.

- *Everyone* does better in the long run:
  - Less conflict and less need for self-protection;
  - More trust in others, and more trust by others.

- Therefore, robots should:
  - Be aware of social norms,
  - Behave in accordance with social norms, and
  - Encourage others to trust that they will follow social norms.

The Trolley Problem

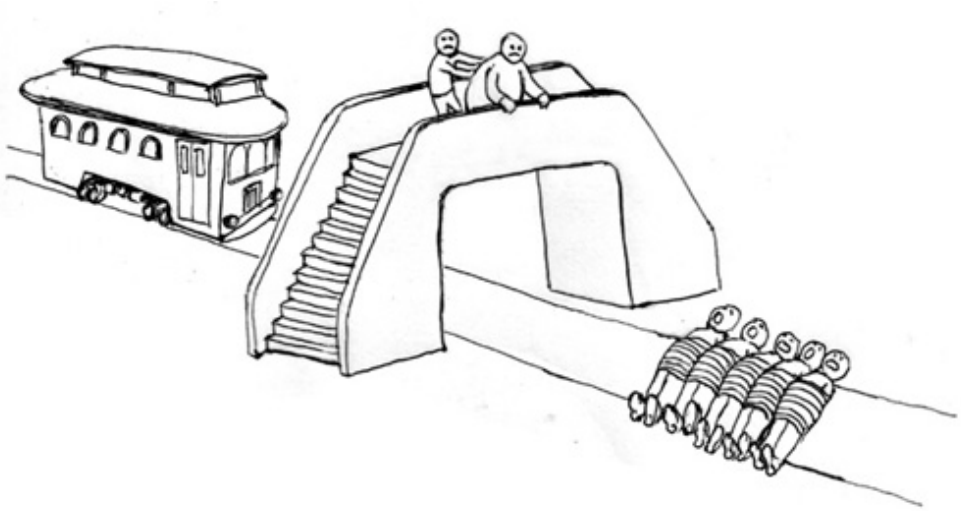
- Artificial problems designed to elicit the criteria people use to make moral decisions.
- "*Do not kill*" is a solid moral rule, but



- Switching the trolley saves five, but kills one.
  - Most people would throw the switch to save the five.
- Why is it OK to violate the rule?
  - Looks like a utilitarian decision: $5 > 1$.

The Footbridge Problem

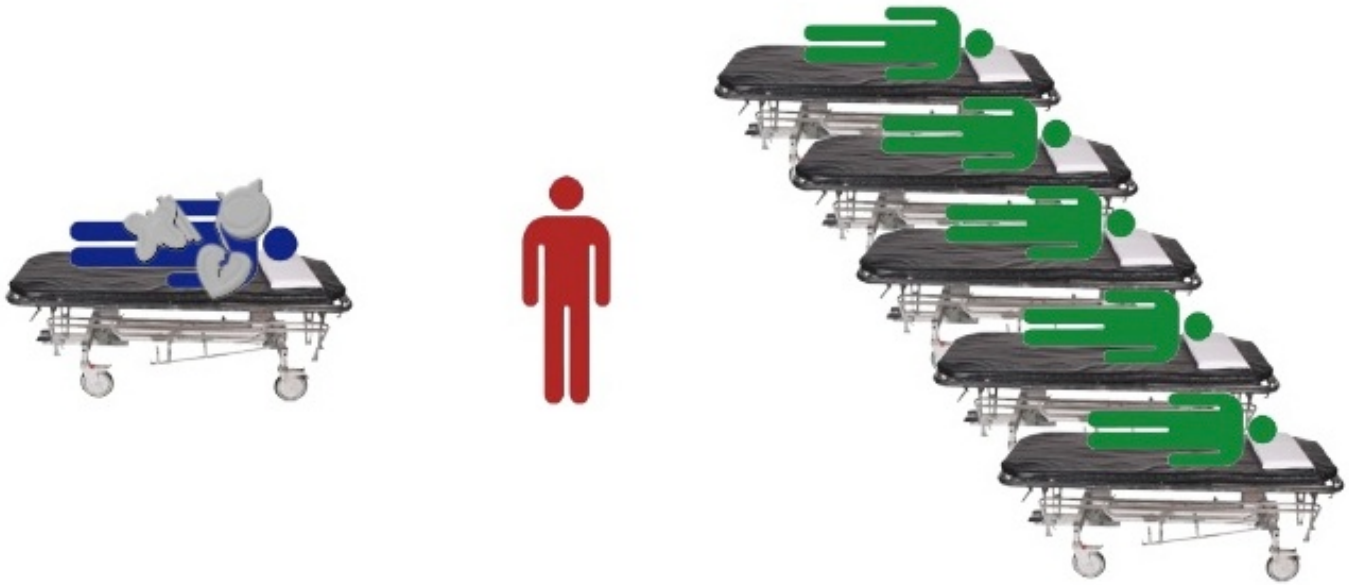
- We create another trolley problem with the same utilitarian calculation: $5 > 1$.



- Pushing the fat man kills him, but saves five.
  - Most people would **not** push the fat man.
- Why is it OK to let five people die in this case?
  - The utilitarian calculation does not prevail.

The Transplant Problem

- A healthy middle-aged man is sedated, waiting for a routine colonoscopy.
- Five bio-compatible patients will die without organ transplants.



- Do you, the surgeon, sacrifice the one to save five?

Emotional Responses

- To the original trolley problem:
  - "*People are going to die! I gotta help!*"
  - "*It's a lot better. Too bad about that one guy*."
- To the footbridge problem:
  - "*People are going to die! I gotta help!*"
  - "*I just **can't** kill the fat man!*"
- To the transplant problem:
  - "*I just **can't** kill my patient. He depends on me*."
  - "*Too bad about those five other people*."

Framing the Problem

What is left out of the problem statement?

- The Trolley Problem:
  - What will people think of your action?
- The Footbridge Problem:
  - Will pushing the fat man really stop the trolley?
    - What if you are wrong?
  - What will people think of your action?
  - Will they trust you not to harm them?
- Transplant Problem:
  - Will future patients trust their surgeons?
  - Will people stop getting colonoscopies?
  - How many people will die of undetected cancers?

Trust

- Trust is key to cooperation, which gives the benefits.
  - How do you know who to trust?
  - How do others know to trust you?

- Your actions **signal** how trustworthy you are.
  - Pushing the fat man sends a signal about you.
  - Sacrificing your patient for five transplants sends a signal.

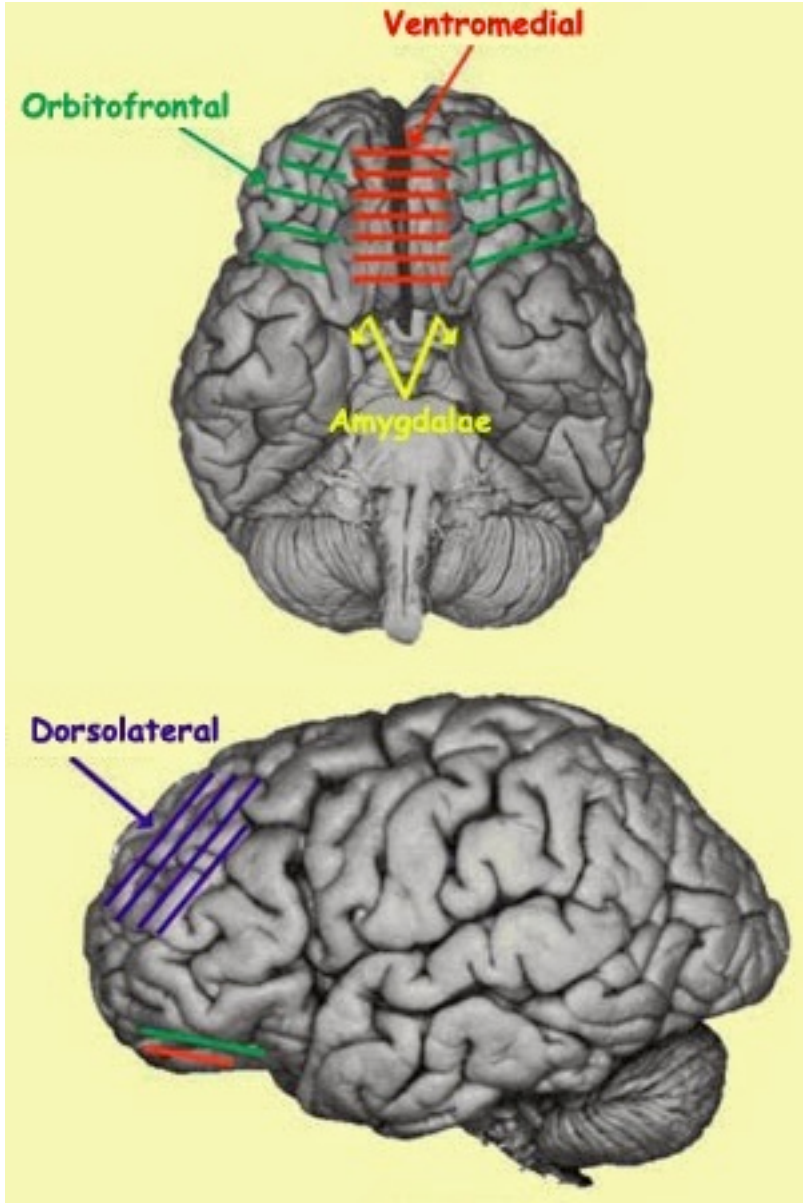
- Your actions are a public testimony that
  - you believe that those actions are OK.

Moral Judgments Must Be Fast

- Problems arise suddenly and need a quick response.
  - Rules are good for quick judgments.
- Emotions, positive or negative, represent rule matches.
  - Action selection is driven by emotional response.
- Quick responses can be wrong!
  - How do we re-evaluate?
  - What do we re-evaluate?

Brain Responses to Moral Problems

- Fast, emotional response:
  - Ventromedial prefrontal cortex (VMPFC)
  - Amygdala
- Slow, deliberative response:
  - Dorsolateral prefrontal cortex (DLPFC)
- [Greene, et al, 2001]



Emotional Responses

- Jonathan Haidt [2012] describes six dimensions of emotional response (Positive / negative):
  - Care / harm
  - Fairness / cheating
  - Liberty / oppression
  - Loyalty / betrayal
  - Authority / subversion
  - Sanctity / degradation

- These evolved to trigger specific actions, responding to biological and cultural needs.
  - *Current* triggers are a wider range of situations.
  - Haidt says deliberation simply rationalizes the emotional response.

Moral Judgments May Be Slow

- If quick emotional judgments are conflicting, or their actions evoke a negative response . . .
  - then a slower deliberative process may be necessary.

- When choices are both positive / both negative, need time to evaluate their long-term utilities.
  - The framing of the problem may be critical here.

- Deliberation may be too slow for critical decisions.
  - Moral rules are accumulated from experience, over individual development and cultural evolution.

Hybrid Implementation

- No single method meets all requirements:
  - Sudden need for a quick decision
  - Ability to respond to novel circumstances
  - Search for appropriate framing of the decision
  - Learning and evolution of moral reasoning
- Proposed: a hierarchy of reasoning methods:
  - Emotional response rules trigger action
  - Utilitarian calculations for difficult problems
  - Search for improved framing when difficulties remain

- Morality and ethics evolves over centuries.

Moral Decision Architecture

- **Fast**: apply response rules to scenarios
  - If choice is unambiguous, done.
  - Trolley: *five* people on track A; *zero* on B.
    - Switch trolley from track A to B. No problem!
- **Slow**: consequentialist deliberation
  - Compare consequences of alternatives.
    - If choice is clear, done.
  - Trolley: *five* people on track A; *one* on B.
    - Switch trolley from track A to B. Lesser of two evils.
- **Slower**: search for appropriate framework.
  - Is there a framework where the choice is clear?
    - The value of trust in predictable behavior.
  - Transplant: kill *one* patient to save *five*?
    - Value of trust in "Do no harm" makes the choice clear: No!

Must a Self-Driving Car Make Moral Decisions? How?

- The car is driving 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 car do?

- Should the car take emergency action to avoid hitting the pedestrian?
- What if it shakes up the passengers, possibly injuring them, in order to save the pedestrian?
- What if saving the pedestrian causes a serious collision, endangering or killing the passengers?
- What if the pedestrian is a small child?

What should the designer do?

- Must the car make the decision in real time?
  Can we avoid the problem or build in a solution?
  - If so, how?

- Human drivers make risk-benefit trade-offs.
  Can a self-driving car make such a trade-off?
  - Realistically, can a car drive slowly enough to make such a collision impossible?

- Will our self-driving cars “do the right thing”?
  - This is not just about defending against lawsuits.

The Self-Driving Car

- The car has five human passengers.
- A child steps in the way of the car.
  - Veering to miss the child will hit a brick wall, perhaps injuring or killing the passengers.
- The car will try to protect both child and passengers.
  - No guarantee of a good outcome is possible.
  - How should the car behave?
- The foundation of trust must already exist.
  - The public must trust that the car always does its best.

The Car Must Have Earned Trust

- The car must always act prudently to minimize risk.
  - In tight surroundings, slow down and observe carefully.
  - Passengers must always wear seatbelts.
- The car must demonstrate that it cares for every life.
  - Its behavior should increase people's trust in the car.
- In the moment of crisis, good preparation and clearly caring for everyone involved will be the best.

Conclusions

- Robots must act morally and ethically.
  - Know and follow social norms.
- Hybrid moral reasoning:
  - Rules (Emotions!) for quick response.
  - Utilitarian calculations when needed.
  - Search for proper framing.
- Trust is essential to society.
  - Act to signal trustworthiness.

References

- Daniel C. Dennett. *The Intentional Stance*. 1987.
- Joshua Greene. *Moral Tribes*. 2013.
- Jonathan Haidt. *The Righteous Mind*. 2012.
- Mark Johnson. *Morality for Humans*. 2014.
- Daniel Kahneman. *Thinking, Fast and Slow*. 2011.
- John Mikhail. *Elements of Moral Cognition*. 2011.
- Eric Posner. *Law and Social Norms*. 2000.
- Peter Singer. *The Expanding Circle*. 1981.