Why and How Should Robots Behave Ethically?

Benjamin Kuipers
Computer Science & Engineering
University of Michigan

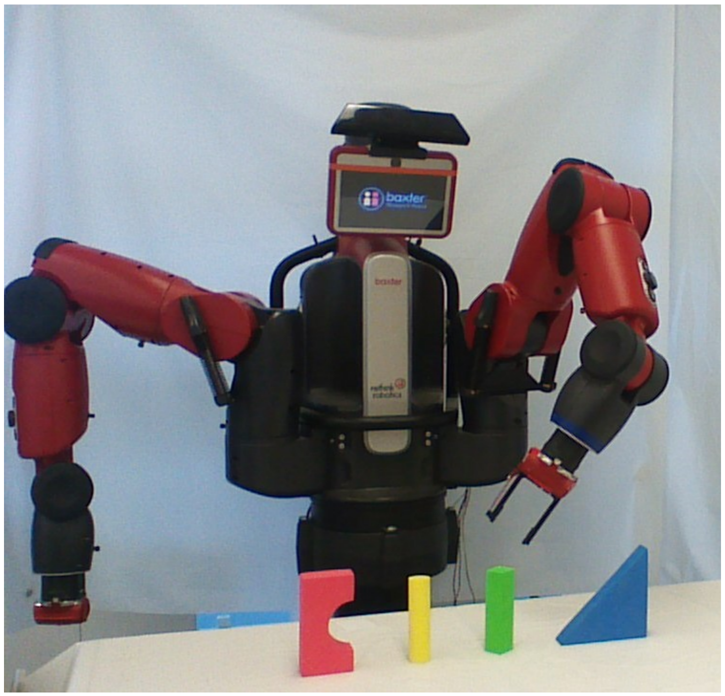
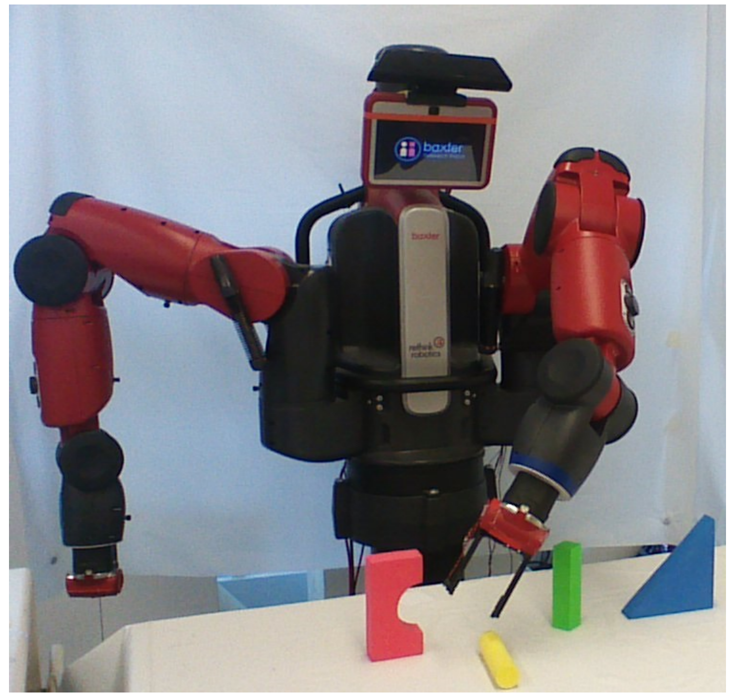
Key Problems of the Mind: AI and Robotics

- Commonsense knowledge of the physical world:
  - *Space*: large-scale, small-scale, peri-personal, . . .
  - Qualitative representations of *continuous change*, including qualitative simulation of dynamical systems.
  - Learning the structure of the *sensorimotor system*.
  - Learning about *objects*, *actions*, and *plans*.

- Commonsense knowledge of the social world:
  - Theory of mind: the beliefs, goals, plans of others.
  - Learning through imitation of skilled others.
  - Morality, ethics, trust: behaving well in society.

Learning to reach, like a baby

- Baxter sees an object; reaches and moves it.
  - Pushes the yellow object; avoids the others.

- What does it need to know, to learn to do this?

Useful Insights

- AI (including Robotics) is not a *thing*.
  - It's a *medium* for expressing hypotheses as computational models.
- The power and robustness of commonsense knowledge comes from multiple representations that can express states of incomplete knowledge.
  - *Space*: topological / metrical representations.
  - *Continuous change*: qualitative / quantitative.
  - *Sensorimotor*: egocentric/allocentric, static/dynamic.
  - *Objects*: 2D images / 2D surfaces in 3D / 3D models.
- Search for ways to use multiple distinct representations together to achieve practical goals.

The Problem of Robots

- We are likely to have more robots (and other 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 can we ensure that robots will behave well?
- How can we trust them?

We worry about robot autonomy.

If we give them great power, they may do great harm, even if we set their goals.




SkyNet Fights Back

- Terminator 2 (1991)
  - https://www.youtube.com/watch?v=4DQsG3TKQ0I

Lessons

- Deploying SkyNet was rational.
  - "*perfect operational record*"
- SkyNet was a learning system.
  - "*learned at a geometric rate*"
- "*SkyNet fights back*."
  - As a critical defense system, it was undoubtedly programmed to protect itself.
- SkyNet finds an unexpected solution.
  - Creative, unconstrained problem-solving.
  - No commonsense or moral critic of plans.

“What about me, Frank?”

- Robot & Frank (2012)
  - https://youtu.be/eQxUW4B622E

“You’re starting to grow on me.”

- Robot & Frank (2012)
    - https://youtu.be/xlpeRIG18TA

"You lied?"

- Robot & Frank (2012)
    - https://youtu.be/3yXwPfvvIt4

Lessons

- Robot has no moral or legal inhibition from stealing, shoplifting, or robbery.
  - "*I took it for you. Did I do something wrong, Frank*?"
  - "*I don't have any thoughts on that* [stealing]."

- Robot has no inhibition against lying.
  - "*I only said that, to coerce you*."
  - "*Your health supercedes my other directives*."

- Robot has no concern for self-preservation.
  - "*The truth is, I don't care if my memory is erased or not*."

Deciding What To Do:
The State of the Art in AI

Decision Theory and Game Theory

- The standard approach to decision making in AI [Russell & Norvig, 3e, 2010] defines **Rationality** as choosing actions to *maximize expected utility*.

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

  - where

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

- **Utility** $U(s)$ represents the individual agent's preference over states of the world.
- *Game theory* is decision theory in a context with other decision-making agents.

The Crux is Defining Utility

- **Utility** $U(s)$ represents the individual agent's preference over states of the world.
  - Utility need not be self-centered. In principle, the individual's utility can reflect *everyone's* welfare.
  - Unfortunately, that's often hard to implement.

- Utility is often defined selfishly --- in terms of the agent's own reward.
  - Appropriate in entertainment games and war games.
  - In society, maximization of self-centered reward often leads to bad outcomes, individually and collectively.
  - Prisoner's Dilemma, Tragedy of the Commons, . . .

Prisoner's Dilemma

- Two prisoners are separated, and offered:
  - If you testify and your partner doesn't, you go free and your partner gets 5 years in prison.
  - If you both testify, you both get 3 years.
  - If neither testifies, you both get 1 year.

Utility is years in prison.

	Testify	**Don't**
Testify	(–3, –3)	(0, –5)
Don't	(–5, 0)	(–1, –1)

- Whatever your partner does, **Testify** is your best choice. Same for your partner.
  - Nash equilibrium: (**Testify, Testify**).
  - You both get 3 years: the *worst* collective outcome.

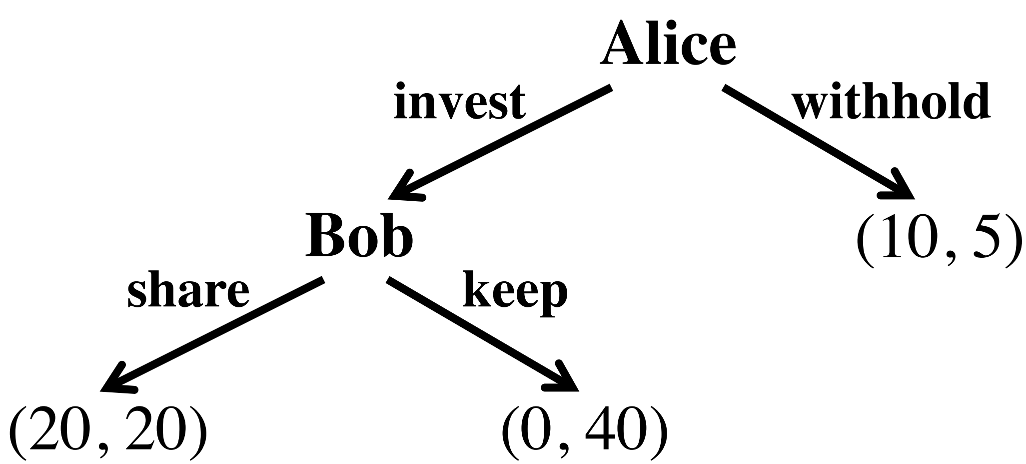
The Tragedy of the Commons

[Garret Hardin, 1968]

- I can graze my sheep on the Commons, or on my own land.
  - Personally, I'm better off grazing as many of my sheep as I can on the Commons, saving my own land.
  - Likewise everyone else.
- So we all overgraze the Commons, and it dies.
  - Then we have only our own land, and no Commons.
  - We're all worse off!
- Modern, real-world Commons:
  - Clean air and water, fishing, climate change, . . .
  - (This shows that the Prisoner's Dilemma scales up.)

The Basic Trust Game

- Alice has $10. Bob has $5.
  - If Alice does nothing, everyone keeps what they have.
- Alice can invest her $10 with Bob.
  - Bob turns $15 into $40.
- Bob decides whether to share the $40 with Alice.



Utility is dollars.

- Nash equilibrium: B:**Keep**, thus A:**Withhold**.

The Basic Trust Game

- Alice has \$10. Bob has \$5.
  - If Alice does nothing, everyone keeps what they have.
- Alice can invest her \$10 with Bob.
  - Bob turns \$15 into \$40.
- Bob decides whether to share the \$40 with Alice.

		Bob	
		Share	**Keep**
Alice	**Invest**	(20, 20)	(0, 40)
	Withhold	(10, 5)	(10, 5)

Utility is dollars.

- Nash equilibrium: B:**Keep**, thus A:**Withhold**.

The Public Goods Game

- N players contribute money to a common pool.
  - The pool is multiplied (× 2 or 3) and the result is distributed evenly among the players.
- Best for society (Cooperation):
  - Everyone contributes their maximum, to get the most benefit from the multiplication.
- Best for individual (Nash equilibrium):
  - Contribute nothing. Save your investment for yourself.
  - Share in the benefit from everyone else’s contribution.
- Cooperation is best for society and each individual.
  - Selfish optimization discourages cooperation.
  - Even the free rider’s benefit collapses.

There are many economic games

- The games highlight conflict between individual's short-term interest, and society's interest (which is often the individual's long-term interest, too).
  - Prisoner's Dilemma
  - Tragedy of the Commons
  - Basic Trust Game
  - Public Goods Game
  - Ultimatum Game
  - Dictator Game
  - . . .
- Ordinary people typically do better than the Nash equilibrium that is the Game Theory "optimum."

What Have We Learned?

- Utility should *not* be defined as individual reward.
  - This may be OK in entertainment, and perhaps war.
  - But in society, it discourages cooperation.

- Philosophical utilitarianism defines utility as *everyone's* reward, which raises other problems:
  - Impossibly demanding requirements.
  - Conflicts with responsibility to family and community.
  - Difficult to build a decision model that is both tractable and reasonable.

Society, Cooperation, and Trust

What is a Society?

- A society is a collection of individual agents, existing in an environment.
  - The environment may include resources, opportunities, threats, and other agents and their societies.
- Individuals interact continually.
  - Some interactions may be abstracted as “games”.
  - Games may be repeated, finitely or infinitely.
  - There may be one game, or many different games.
  - Players may be identifiable, or anonymous.
  - Individuals may belong to “us”, or to “them”.

Cooperation Pays Off for Society

- The society benefits from cooperative behavior.
  - Individuals get good rewards, but may be tempted by even better rewards for free riding.
  - Widespread free riding defeats cooperation.
    - Nash equilibrium = $(0, 0, \ldots 0)$
- Social norms direct individuals toward cooperation, and away from tempting local optima.
  - Societies can evolve mechanisms for punishment of free riders, even when punishment is costly.

Trust is Necessary for Cooperation.

- Many aspects of society depend on trust.
  - I can trust most people not to try to kill or steal from me.
    Saves on overhead for defending myself.
  - I trust most drivers to drive safely and courteously.
    Allows me to drive more safely and efficiently.
  - I trust most companies to fix/replace defective products.
    Makes it easier to shop and buy.
  - I can trust most people to keep most of their promises.
    Enables cooperative enterprises.
  - . . . (many others)

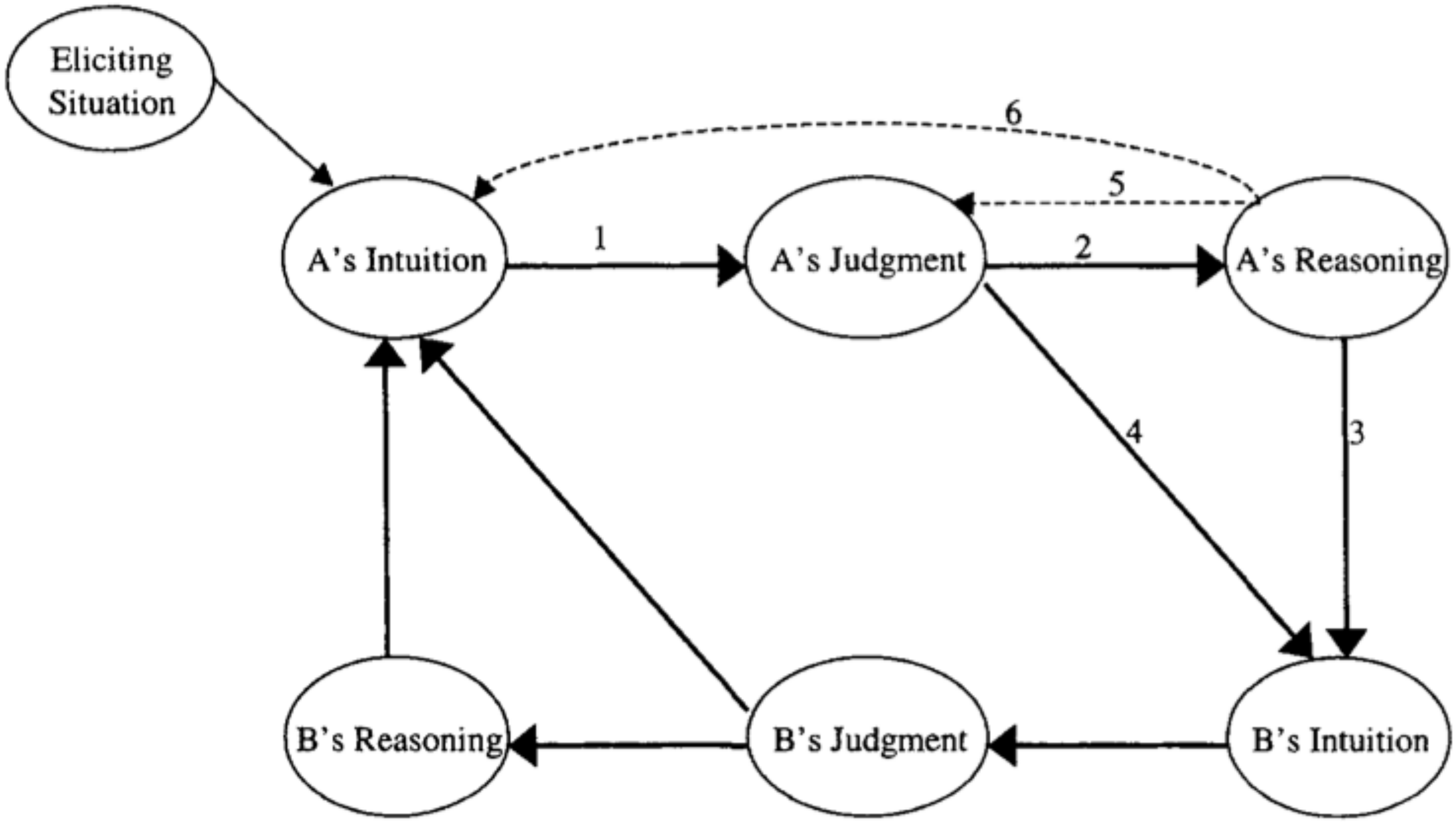
Trust and Trustworthiness

- What is trust?
  - "*Trust is a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another*."
- Trust has value for you.
  - Others can take actions offering larger benefits for all, even though it makes them vulnerable to you.
- Trust is a capital asset ("social capital").
  - It accumulates slowly.
  - It can be destroyed quickly.

Explaining Moral Decisions

- Your actions speak for you.
  - They signal what sort of person you are.
  - They signal what you approve of.
- Your explanation clarifies those actions.
  - Which simple abstract model you used to decide.
  - Which parameter values you used in that model.
  - Demonstrate how you used the model.
- Your explanation affects the trust others have in you, in a positive or negative way.
  - It can also influence the moral evolution of society.

Haidt's Moral Decision Architecture



- "Intuition" = pattern-matched emotional response drives quick judgment. Justification comes later.
- Judgment and justification send signals to others.

[Jonathan Haidt, 2001]

Social Evolution

Evolution of Ethics and Society

- Individuals may want to maximize own utilities.
- But, society offers greater collective strength and health than any individual --- self or threat.
- Therefore, individuals who are inclined to join into successful societies will thrive, relative to loners.
- Societies "want" to survive, thrive, and propagate, in the evolutionary sense that those that do are increasingly represented in the future population.

Evolution of Ethics and Society

- Societies “want” to survive, thrive, and propagate, in the evolutionary sense that those that do are increasingly represented in the future population.

- Societies succeed according to their abilities to cultivate ethics, morality, and trust among their individuals, producing a surplus of resources for those individuals, and for society as a whole.

- Which specific ethics and morality helps a society survive, thrive, and propagate depends on its physical, cultural, and competitive context.

Evolution of Societies

- Societies evolve over time, including changes to their morality and ethics.
  - They respond to changes in their environment.
  - Changes in individual decisions affect the social norms, for better or for worse.

- For a society to survive and thrive:
  - It must accumulate resources.
  - It must protect itself against predation and attack.
  - It must keep the allegiance of its individuals.

- Its social norms help it survive and thrive.

Consequentialism

- Evolutionary development of societies.
  - Morality, ethics, and trust promote cooperation.
  - Cooperation makes society stronger and healthier.
  - The strongest societies survive and propagate.

- The value of a moral and ethical system is defined by the survival and propagation of the society.
  - A meaningful definition, but . . .
  - Predicting evolutionary progress is not a feasible way to make ethical decisions in real time.

- Individuals need simpler, more useful, heuristics.

How Can an Individual Decide What To Do?

Real-Time Ethical Response

- Situations often need an immediate response.
  - No time for careful deliberation.
  - Real-time response requires pattern-matched rules, constraints, or cases.
- But deliberation is possible after the fact.
  - We learn from good and bad decisions.
  - We learn from explanations: others', and our own.
- The knowledge representation must support:
  - Useful states of partial knowledge, and
  - Incremental improvement toward practical wisdom.

Individuals Need Ethical Heuristics

- We draw on theories of philosophical ethics that philosophers and prophets have been thinking, teaching, and developing for many centuries.
  - **Utilitarianism** ("*What action maximizes* utility *for all?*")
    - Special case of **consequentialism** ("*What action has the best* consequences *for all?*")
  - **Deontology** ("*What is my* duty*, to do, or not to do?*")
  - **Virtue ethics** ("*What would a* virtuous *person do?*")

- Instead of treating these as mutually exclusive, we see them as parts of a single complex reality.
  - "*The Blind Men and the Elephant*"
  - "*Climbing the same mountain on different sides*"

An AI Perspective on Ethical Theories

- The different ethical theories suggest different AI knowledge representations, able to express different kinds of ethical knowledge.
  - **Utilitarianism** (*Decision theory* / *Game theory*)
    - Good for continuous optimization, but not in real time.
    - Sensitive to choice of utility measure.
  - **Deontology** (*Pattern-matched rules and constraints*)
    - Good for explanation and computational efficiency.
    - Depends on the terms that can appear in patterns.
  - **Virtue Ethics** (*Case-Based Reasoning*)
    - Good for expressive power in complex domains.
    - Good for incremental learning from experience.
- Using multiple models together is more robust.

An Ethical Knowledge Base

- Must express many states of knowledge from beginner to expert (*phronesis*).

- Case base:
  - Rich description of current situation
    - Actors, relations, actions, events, context, . . .
  - *Cases*: stored descriptions of previous *situations*
    - Situation, moral valence (good/bad), response, success

- Pattern-matched rules and constraints:
  - Relatively simple pre-specified pattern language.
    - "*Thou shalt not kill / steal / lie / . . .*"

Early Ethical Knowledge

- Children are taught rules, constraints, and simple patterns by their parents.
  - "*You* stole *this*. *What do you think about that*?"
- The early ethical knowledge base is populated from experienced situations with clear labels.
  - The state space of possible situations is enormous.
  - The labeled cases characterize large regions.
  - Little knowledge of the complex boundaries between clear regions.
- Content determined by current state of societal moral and ethical knowledge.

Using the Ethical Case-Base

- When the agent encounters a new situation
  - Retrieve the most *similar* matching cases
    - Evaluate similarities and differences
  - Adapt case response to the needs of the situation
- When conflicting cases match the situation
  - Analyze the similarities and differences.
    - Compare features supporting different evaluations.
  - Compare and adapt the associated responses.
- Select or construct a response, and do it.
  - Observe outcome quality, and critiques by others.

Updating the Ethical Case-Base

- Store the description of the current situation as a new case in the case-base.
  - Include response and its evaluation.
  - The growing case-base represents accumulated experience.
- When many similar cases have the same response:
  - Identify the relevant features; abstract away variation.
  - Create a new explicit rule.
- Nearby cases with different responses require slow post-hoc deliberation and analysis.

Phronesis

- Practical wisdom needs a rich and dense case-base.
  - "rich" means a variety of different case descriptions.
  - "dense" means a new situation matches many cases.
  - Abstract cases to rules for simplicity and efficiency.

- Phronesis requires quality of decisions, not just quantity of experience in cases.

- Several learning methods:
  - From explicit instruction by parents and others.
  - From personal experiences and outcomes.
  - From observing exemplary others (*phronemos*).

This is a Preliminary Sketch

- Design goals:
  - Combine insights from major ethical theories.
  - Provide expressive power for states of knowledge.
  - Identify feasible incremental inference methods.
  - Feedback systems at multiple time-scales.
  - Experience can lead to increasing expertise, both for the individual and for society.
- There is much more to be learned.
  - But it's a start.
  - Help with debugging is always welcome.

What About
Self-Driving Cars?

The Deadly Dilemma

- 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 self-driving car 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?
- We call this the “Deadly Dilemma.”

Who should the self-driving car kill?

- Should it kill the pedestrian or the passenger?
  - If the pedestrian, why should the public tolerate these self-driving cars?
  - If the passenger, why should anyone ever trust (and buy) the self-driving car?

- Even if the Deadly Dilemma is very unlikely, it will not be impossible.
  - People still want to know what the car will decide.

Can the designer avoid the problem?

- Must the car make the decision in real time? Can we design the car to avoid the problem?
  - Realistically, a car cannot drive slowly enough to make such a collision *impossible*.
- A good outcome cannot be guaranteed.
  - Human drivers make risk-benefit trade-offs.
  - To have acceptable performance, a self-driving car will necessarily make such trade-offs.
- The problem is framed too narrowly.
  - The car must act to earn our trust.

The Cars Must Earn Our Trust

- The social capital of trust must be accumulated.
  - Society must learn that the car is trustworthy.
  - Every car must *show* that it protects every life.
    - Not just the lives of its own passengers.
- The self-driving car must continually demonstrate "practical wisdom."
  - Slow down where pedestrians could appear.
  - Steer to maximize visibility and warning time.
  - Demonstrate foresight and expertise when starting, stopping, and turning.
- In case of disaster, well-earned trust will lead to understanding, and a chance for forgiveness.

Signaling Intent

- The Google car stops on yellow lights, and has suffered from rear-end collisions.
  - Legally, it is blameless. But is this right?
  - It should be aware of what other drivers expect.
  - It should flash its brake lights, to signal its intent.
- Taking turns at a four-way stop.
  - Back up slightly, to yield right-of-way.
  - Move forward slowly, to assert right-of-way, when it's your turn.
- Human drivers have ways to signal to each other.
  - How should a self-driving car send signals?
  - Does it need a better signaling mechanism?

Technological Fixes . . .

. . . make the Deadly Dilemma less likely, though still not impossible.

- "Deer Crossing" – dangerous, suddenly-appearing hazard, without the moral dilemma.
  - Constant situational awareness
  - Early warning → best immediate response

- "Avoiding the invisible pedestrian" –
  - Understand and respond to motion affordances.
  - Add beacons to eliminate visibility limitations.

- . . .

Conclusions

Framework Summary

- Society exists for individual people.
- Cooperation benefits society (and individuals).
- Trust is necessary for cooperation.
- Morality/ethics helps the society survive, thrive, and propagate, by encouraging cooperation.
- Individuals need useful ways to decide what to do.
  - Rules, constraints, and cases for quick response.
  - Utilitarianism and explanation for slower post-hoc analysis and learning.
  - Abstraction of useful cases to converge on a concise vocabulary of patterns and set of rules.

Conclusions for Robots

- To act as members of our society:
  - Robots must show that they are trustworthy.
  - Robots must be able to explain their behavior, and learn from explanations.
  - Robots should not be given power beyond the trust they have earned.

- To know how robots can behave well:
  - We need a tractable computational model of how morality and ethics helps people behave well in society.

References

- Robert Axelrod. *The Evolution of Cooperation*, 1984.
- Bacharach, Guerra & Zizzo. The self-fulfilling property of trust: An experimental study. *Theory and Decision*, 2007.
- Dandekar, Goel, Wellman & Wiedenbeck. Strategic formation of credit networks. *ACM Trans. Internet Technology*, 2015.
- Alan Page Fiske. *Structures of Social Life*. 1991.
- Jonathan Haidt. *The Righteous Mind*. 2012.
- Johnson & Mislin. Trust games: A meta-analysis. *J. Economic Psychology*, 2011.
- Kuipers. Toward morality and ethics for robots. *AAAI Spring Symposium on Ethical and Moral Considerations in Non-Human Agents*, 2016.
- Leyton-Brown & Shoham. *Essentials of Game Theory*, 2008.
- Lin, Abney & Bekey. *Robot Ethics: The Ethical and Social Implications of Robotics*, 2012.
- Rousseau, Sitkin, Burt & Camerer. Not so different after all: a cross-discipline view of trust. *Academy of Management Review*, 1998.
- Russell & Norvig. *Artificial Intelligence: A Modern Approach*, 3e, 2010.
- Wright & Leyton-Brown. Level-0 meta-models for predicting human behavior in games. *ACM Conf. Economics & Computation*, 2014.