What is Trust and How Can My Robot Get Some?

Benjamin Kuipers
Computer Science & Engineering
University of Michigan

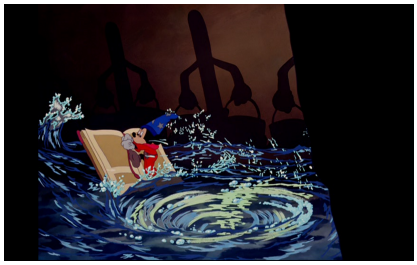
AIs as Members of Society

- We are likely to have more AIs (including robots) 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 trust them?

We worry about the robots

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



"You're starting to grow on me."

- Robot & Frank (2012)

"You lied?"

- Robot & Frank (2012)

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 desire for self-preservation.
  - "*The truth is, I don't care if my memory is erased or not*."

SkyNet Fights Back

- Terminator 2 (1991)

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.

Trust is Important to Society

- Many aspects of society depend on trust.
  - I can trust most people not to try to kill me 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 can trust most people to keep their promises most of the time. *Enables cooperative enterprises.*
  - I can trust most companies to replace or repair defective products. *Makes it easier to shop and buy.*
  - . . . (many others)
- We want to be able to trust robots, as they make decisions and act in our society.

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.*"
- What is trustworthiness?
  - *Deserving of trust or confidence; dependable; reliable.*
- Trust is the rational response to trustworthiness.
  - Trust from others has tangible value.
  - Rewards get better if you can trust the other players.

Complex World / Simple Models

- The actual world is infinitely complex.
  - To make decisions, we need simple models.
  - To build simple abstract models, we must decide
    - what to leave in,
    - what to leave out.
- Game theory provides simple models of the complex world, and procedures for deciding what to do.
  - Some models are plausible and helpful.
  - Other models are unrealistic and unhelpful.

How Should a Robot Decide?

- 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** $U(s)$ is the individual agent's preference over states of the world.
  - In principle, utility can reflect *everyone's* welfare, but that is typically too difficult to implement.

The Crux is Defining Utility

- **Utility** $U(s)$ is 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 typically defined in terms of the agent's own reward.
  - Individual maximization of self-centered reward often leads to bad outcomes, individually and collectively.
  - Tragedy of the Commons, Prisoners' Dilemma, . . .

The Tragedy of the Commons

- 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 overgraze the Commons until 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, . . .
  - (Shows that the Prisoner's Dilemma scales up.)

Prisoner's Dilemma

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

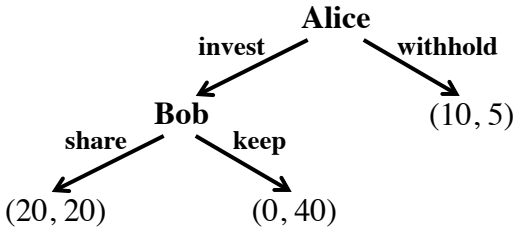
Utility is years in prison.

	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* overall outcome.

The Basic Trust Game

- Alice gets $10. Bob gets $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 gets $10. Bob gets $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**.

What's the Problem?

- Utility is the obvious measure: dollars, years, etc.
- Game Theory tells us the best action.
- But the result of the best action is terrible!

- What do people do in the Basic Trust Game?
  - Many people trust the other player.
  - Many are trustworthy, sharing with the investor.
- People do better than the Game Theory solution.

- Does this refute Game Theory? No.
- Does this show that people are irrational? No.

What's the Problem?

- We have a *reductio ad absurdum*.
- The assumption was: Utility can be defined strictly in terms of individual reward.
  - This is reasonable for artificial games, played for entertainment.
  - This is **not** reasonable for "games" as simple models (abstractions) for decision problems in the real world.
- **Conclusion**: The definition of utility must be expanded to include other factors, beyond reward to the individual agent.

Claim: Trust Has Value

- The trust that others place in me has value.
  - Others will take actions that offer larger benefits, even though it makes them vulnerable.
  - They trust me not to violate their trust.
- Trust is a capital asset ("social capital").
  - It accumulates slowly. ⇧
  - It can be destroyed quickly. ⬇
  - There can be a culturally-specific prior.
- Utility must include the value of trust, as well as individual reward.

The Value of Trust

- Suppose Alice has decided to trust Bob.
  - Bob now has $40 to keep or split.
- Based on individual reward, his choice is:

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

Utility is dollars.

The Value of Trust

- Suppose Alice has decided to trust Bob.
  - Bob now has $40 to keep or split.
- The value of trust changes the utilities:

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

Utility is dollars and trust

  - Now the Nash equilibrium is (**Invest**, **Share**), a win-win choice!

Trust in the Prisoner's Dilemma

- In the original PD, Utility is individual reward.
  - The Nash equilibrium is bad for the individual, and worst case collectively.
- With trust as part of utility,
  - The Nash equilibrium is (**Don't, Don't**), the best case, individually and collectively.

	Testify	**Don't**
Testify	(–3–2, –3–2) = (–5, –5)	(0–2, –5+1) = (–2, –4)
Don't	(–5+1, 0–2) = (-4, –2)	(–1+1, –1+1) = (0, 0)

  - (The additive combination, and the numerical values, are only for illustration.)

Including Trust in Utility

The additive combinations above, and the numerical values, are over-simplified, and purely for illustration.

- $U(s) = f(reward, trust)$.
  - Reward and trust don't have the same units, and the numerical values are just for illustration.
- There are many hypotheses about the mechanism, and the function reflecting it:
  - Kindness reciprocity
  - Inequality aversion
  - Trust receptiveness
  - Altruism
  - . . .

Trustworthiness and Character

- How is trustworthiness estimated?
  - Behavior is evidence of *character* (a hidden variable).
  - Character predicts future trustworthiness.
  - Reputation is an estimate of character.
- Reputation depends on good behavior, and on persistent identity.
  - When do you trust an eBay seller?
    - **Many** ratings comes from persistent identity.
    - **Positive ratings** come from good behavior.

What About One-Shot Games?

- Why should Bob care about the trust he gets if he will never see Alice again?
  - Behavior is evidence of *character* (a hidden variable).
  - Character predicts future trustworthiness.
- "*The Tell-Tale Heart*"
  - Bob can observe his own character.
  - He doesn't know how well others can observe this hidden variable from other evidence.
- If Bob cares about character, he will behave well.

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?

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*.
- Human drivers make risk-benefit trade-offs.
  - To be acceptable, a self-driving car will necessarily make such trade-offs.
  - No absolute guarantee of a good outcome is possible.
- The problem is framed too narrowly.
  - It is wrong to treat this as a "trolley problem".

The Car 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 cares for every life.
  - Not just for the lives of its own passengers.
  - People should learn to trust *all* self-driving cars.
- The car must always act prudently to minimize risk.
  - In tight surroundings, slow down and observe carefully.
  - Require its passengers to wear seatbelts.
- In case of disaster, well-earned trust will lead to understanding, and a chance for forgiveness.

Explanation

- 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 that you used the model correctly.
- Your explanation affects the trust others have in you, in a positive or negative way.
  - It can also influence the moral evolution of society.

Answering the Questions

- What is trust?
  - The response of others to your trustworthiness.
  - Their willingness to accept vulnerability, in confident expectation of your good behavior.
- How can my robot get some?
  - By signaling to others that it is trustworthy.
  - By demonstrating, repeatedly and consistently, that it will fulfill the trust placed in it, even when there are temptations to the contrary.

What do our robots need?

- The robot needs to recognize and use the simple game theory model appropriate to a situation.
- The robot needs to define utility in terms of both individual reward and the trust it receives.
- The robot needs to explain its choice of action, and understand critiques of its explanation.
- The robot needs to signal its *trustworthiness*, even its *character*, to those around it.
  - It also needs to recognize those signals from others.

Conclusions

- We want and need robots to be trustworthy.
- Game theory is a formal method for rationally selecting actions.
  - Utility defined only in terms of individual reward can lead to disaster.
  - Utility must include a component for trust.
- Trust can be gained slowly, and lost quickly.
  - Robots need to signal that they are trustworthy.
  - Explanations clarify the lessons from behavior.
- Robots should not be given power beyond the trust they have earned.

References

- Robert Axelrod. *The Evolution of Cooperation*. 1984.
- Bacharach, Guerra & Zizzo. The self-fulfilling property of trust: An experimental study. *Theory and Decision*, 2007.
- 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.
- 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.