Ethics as a Technical Problem in AI
CCC Assured Autonomy Workshop #2

Benjamin Kuipers
University of Michigan

What Is Our Perspective?

• As AI researchers, we build computational models of aspects of Mind.
  – Some of these models can have great practical and economic impact.

• Ethics is a significant aspect of Mind.
  – At least in humans

• What is the pragmatic value of ethics?
  – Should all intelligent systems have ethics?
  – What would this mean?
What Is Ethics?

- Ethics is a tool for a society to encourage its individual members to behave in cooperative ways that benefit the society.

- Trust enables cooperation. Distrust discourages cooperation and damages society.

How Does Ethics Create Trust?

- **Social norms** are behavioral properties that one can trust.
  - I use the term “social norm” broadly and inclusively.
  - Trusting social norms saves resources that would otherwise be needed for defense and recovery.
  - E.g., driving on the same side of the road makes everyone’s transportation safer and more efficient.

- Trust can be lost, and is slow to recover.
  - E.g., is the speed limit a viable social norm?
How Does Ethics Create Trust?

• Visibly following the social norms and ethical principles of society signals trustworthiness.
  – “Costly signals” are less likely to be false.

• Trust is the willingness to accept vulnerability, with confidence that it will not be exploited.
  – Cooperation requires vulnerability.
  – A prospective cooperative partner must be trustworthy.

• Exploitation may yield a better reward on a single interaction, but the trustworthy person receives better opportunities for cooperation.

What is Ethics for?

Ethical principles

↓

Trustworthiness

↓

Trust

Cooperation

Positive-sum outcomes

More resources for society

Social norms

↓

Save resources for defense & recovery

More resources for society
Knowledge and Humility

• The world is infinitely complex.
  – “The baby, assailed by eyes, ears, nose, skin, and entrails at once, feels it all as one great blooming, buzzing confusion . . . ” [William James, 1890]

• Knowledge is finite.
  – We construct partial models of our experience.
  – Those models express certain aspects of the world.
    • Other aspects are treated as negligible.
    • And they may be, for some purposes, but not for others.
    • Multiple, different models help us triangulate reality.
  – “The Blind Men and the Elephant”

Moral Philosophy

• Major theories of philosophical ethics resemble different AI knowledge representations.
  – Deontology (“What is my duty, to do, or not to do?”)
    • Pattern-matched rules and constraints
  – Virtue Ethics (“What would a virtuous person do?”)
    • Case-based and analogical reasoning
  – Utilitarianism (“What action maximizes utility for all?”)
    • Decision theory / Game theory

• These are human constructs, trying to explain observed regularities in human moral decisions.
So, What?

• We are designing intelligent agents that participate in our society.
  – Other intelligent agents (humans, institutions) also participate in our society.

• What does the purpose of ethics imply?
  – For society to thrive, its members (humans, AIs, institutions) should behave ethically.
  – To create trust, to encourage cooperation.
  – Otherwise, society suffers.

• Trust is willingness to accept vulnerability, with confidence that it will not be exploited.

Which Social Norms for AIs?

• What do we expect to be able to trust?

• If a social norm is not respected by members of the society, it is weakened. People stop being able to trust it.
  – If AIs and institutions act as members of society, they can weaken our social norms, and hence our society.
  – Unless we can find ways to articulate the social norms that we expect AIs and institutions to follow.
What Do I Need to Trust?

• In a given context, I need to be able to trust that
  – my vulnerabilities will not be unfairly exploited.

• In the given context, the questions to ask:
  – What vulnerabilities do I have?
  – What are the potential exploitations?
  – What social norms would discourage the exploitations?
  – What punishments should violators receive?

• A proposed methodology for designing ethics.
  – These are answerable questions.
  – A society includes many different contexts.

Cooperation is Essential

• There are existential threats to humanity.
  – Don’t worry (too much) about super-intelligence.
  – Worry about climate change.

• To meet these existential threats
  – Cooperation will be essential;
  – Cooperation depends on trust.
  – Trust is being eroded.
  – We must do what we can.