An Introduction To Machine Ethics

Concepts, Challenges and a critical outlook

Overview

- will follow for the final version!

Motivation

- Gilbert Ryles Distinction between Knowing how (procedural) vs. Knowing that (declarative)
  - A: Is to know how to do something just to know lots of facts (of the right sort)?
  - B: To know how to do something is not just to know the right facts about how to do it
- Without even defining Machine Ethics one Challenge will be
  - to translate procedural knowledge into propositional (declarative) knowledge
  - We need ethical theories in order to do that

What is (Machine) Ethics about?

Defining Ethics

- The English word "ethics" is derived from the Ancient Greek word ēthikós (ἠθικός), meaning "relating to one's character"

- Morality (from Latin: moralitas, lit. 'manner, character, proper behavior') is the differentiation of intentions, decisions and actions between those that are distinguished as proper and those that are improper

What is (Machine) Ethics about?

- Descriptive vs. Normative Ethics
- Meta-Ethics
- Applied Ethics
  - Computer Ethics
  - Machine Ethics
  - We want to explore the possibilty to endow machines with abilities to think/act morally
  - machine ethics is also a metaethics: what counts as a moral action, how can morals be implemented?

8 Criterias To Frame An Ethical Question

1. norms and values
2. generality
3. impartiality
4. universalizability
5. categorical
6. priorities between different moral arguments
7. sanctions
8. social function: morals regulate human affairs

Normative Ethics

Utilitarianism and Consequentialism

- The Classical Utilitarians, Jeremy Bentham and John Stuart Mill, identified the good with pleasure, so, like Epicurus, were hedonists about value. They also held that we ought to maximize the good, that is, bring about ‘the greatest amount of good for the greatest number’.
- Utilitarianism is also distinguished by impartiality and agent-neutrality. Everyone's happiness counts the same. When one maximizes the good, it is the good impartially considered.

Normative Ethics

Consequentialism

- Consequentialism = whether an act is morally right depends only on consequences (as opposed to the circumstances or the intrinsic nature of the act or anything that happens before the act).
- Any consequentialist theory must accept the claim that I labeled “consequentialism”, namely, that certain normative properties depend only on consequences. If that claim is dropped, the theory ceases to be consequentialist.

Normative Ethics

Deontology

- deontology falls within the domain of moral theories that guide and assess our choices of what we ought to do (deontic theories), in contrast to those that guide and assess what kind of person we are and should be.
- The most familiar forms of deontology, and also the forms presenting the greatest contrast to consequentialism, hold that some choices cannot be justified by their effects—that no matter how morally good their consequences, some choices are morally forbidden.

Normative Ethics

Deontology

- The most glaring one is the seeming irrationality of our having duties or permissions to make the world morally worse
- Second, it is crucial for deontologists to deal with the conflicts that seem to exist between certain duties, and between certain rights
- Kant's bold proclamation that “a conflict of duties is inconceivable” (Kant 1780, p.25) is the conclusion wanted, but reasons for believing it are difficult to produce.

Paradoxes

Moral Dilemmas as a Challenge for Machine Ethics

- The crucial features of a moral dilemma are these: the agent is required to do each of two (or more) actions; the agent can do each of the actions; but the agent cannot do both (or all) of the actions. The agent thus seems condemned to moral failure; no matter what she does, she will do something wrong.
- Ethicists who are concerned that their theories not allow for moral dilemmas have more than consistency in mind. What is troubling is that theories that allow for dilemmas fail to be uniquely action-guiding.

Paradoxes

- The principle that ‘ought’ implies ‘can’
  - Example

Can Machines Be Ethical?

- The fundamental idea of Kant's philosophy is human autonomy.
- "Autonomy" literally means giving the law to oneself
- On the compatibilist view, as Kant understands it, I am free whenever the cause of my action is within me. So I am unfree only when something external to me pushes or moves me, but I am free whenever the proximate cause of my body's movement is internal to me as an "acting being" (5:96).
- For transcendental idealism allows that the cause of my action may be a thing in itself outside of time: namely, my noumenal self, which is free because it is not part of nature.

Can Machines Be Ethical?

- If maxims in general are rules that describe how one does act, then imperatives in general prescribe how one should act.
- Finally, the only way to act freely in the full sense of exercising autonomy is therefore to act on formal principles or categorical imperatives, which is also to act morally.

Can Machines Be Ethical?

- In General: If we state that moral responsibility supposes at least four necessary conditions:
  - free will
  - causality
  - intentionality
  - knowledge
- It would be a categorical mistake to assume that machines can fulfill a 1:1 translation of hotly disputed general concepts that constitute a moral person
  - Problem: Often big philosophical question are undecidable

Can Machines Be Ethical? (Only Draft in German)

- die moralimplementation macht das kernstück der maschinenethik aus
- das system soll ein bestimmtes, hier ein moralisches, kognitives problem lösen und den lösungsweg explizit angeben
- woran lässt sich bemessen, ob eine funktional hinreichend große ähnlichkeit zu moralischem handeln vorliegt, um eine maschine als expliziten moralischen akteur zu begreifen?
- computer besitzen zwar keine mentalen zustände, aber funktional vergleichbare zustände.
- sie besitzen quasi-meinungen, quasi-wünsche etc.

Conclusion