Introduction to Machine-Ethics

Philip Wälde, M.A
Freie Universität Berlin, Institute of Computer Science
Seminar: Normative Reasoning and Machine Ethics [Winter Term 2019/20]

Draft Presentation

I. What is (Machine) Ethics about?

- basic concepts

a. knowing how vs. knowing that

- transforming implicit forms of reasoning into explicit propositional knowledge

b. ethic and ethos

- short remark on the basic semantic meaning
- morals vs. ethics
- descriptive vs. normative ethics
- applied ethics e.g. machine ethics
- we want to explore the possibilty to endow machines with abilites to think/act morally
- but: machine ethics is also a metaethics: what counts as a moral action, how can morals be implemented etc.

c. difference between law and morals, cannot be equated

- example for interconnection between those two domains

d. 8 criterias in order to establish something as an ethical question

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

e. utilitarism and denotology

- basic definitions from Bentham and Kant
- explain both ethical perspectives in the context of the aforementioned criterias

II. The free will debate (Leibniz and Frankfurt) -optional

- this would be a case for strong AI and what machines (probably) still can't do
- look at two similiar approaches from Leibniz and Harry Frankfurt
    - what is a free will
    - free will + causality + intentionality + knowledge as necessary components to establish moral responsibility

III. On the notion of Pseudoethics

- traditional approaches in philosophy presuppose disputed terms in building their theories e.g. consciousness, (free) will, personhood, intentionality, emotions etc.

- explain the notion of concepts denoted as "pseudo-X" as a heuristic to model an ethics that can be implemented
    - e.g. moores ethical impact agent, implicit ethical agents explicit ethical agents
    - what machines can do

- paradoxes/dilemmata as a challenge + concluding remarks

***I&III should be enough for a 20minute talk,**

****II would shed a light on a heavy armed philosophical debate as an example of what should not be taken as a point of reference, when doing machine ethics.**

*** sources and quotes on the upcoming slides, also a small bibliography