Please check-in

Winter Term 21/22

Artificial Intelligence, Ethics & Engineering

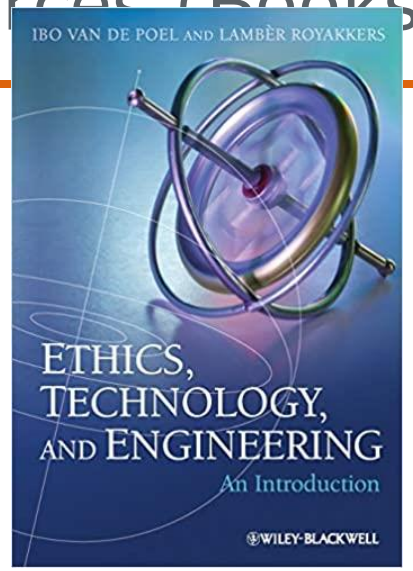
Lecture-4: Introduction to Ethics Theories

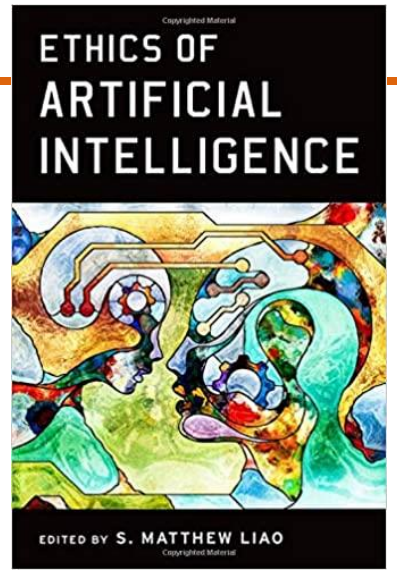
Prof. Dr. Holger Giese (holger.giese@hpi.uni-potsdam.de)

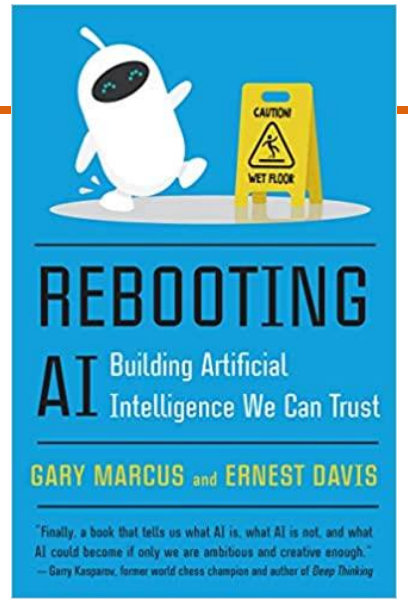
Christian Medeiros Adriano (christian.adriano@hpi.de) - **"Chris"**

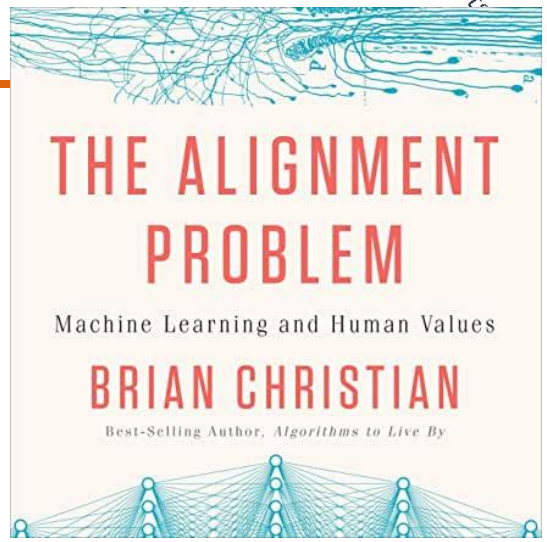
Christian Zöllner (Christian.zoellner@hpi.de)

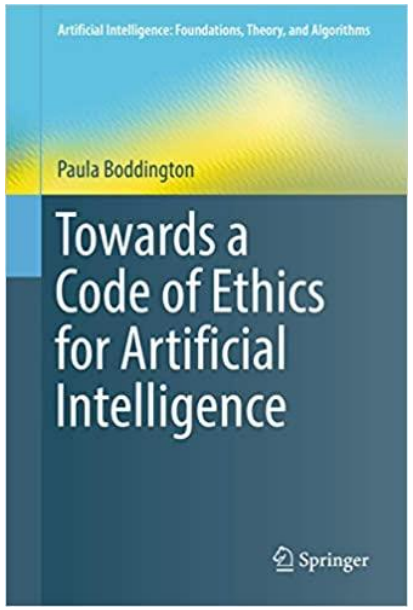
Sources (Books)

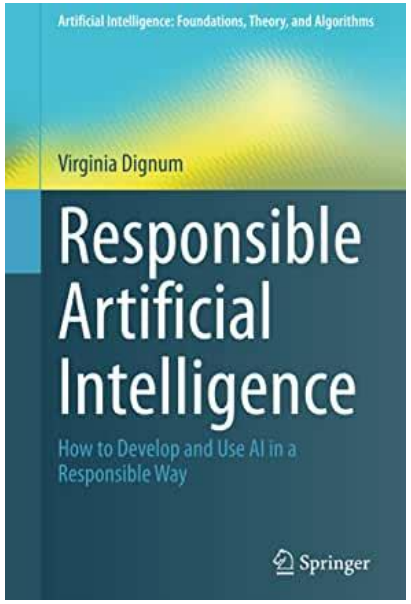


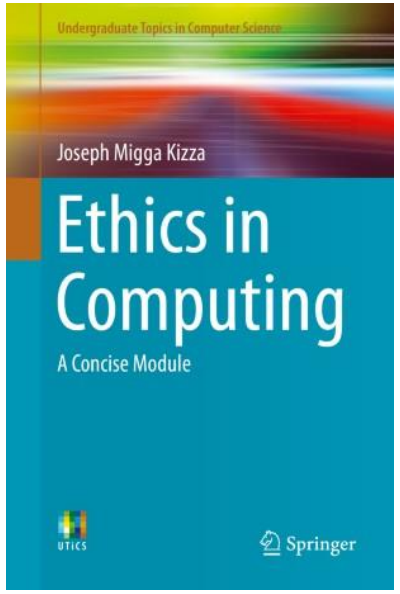


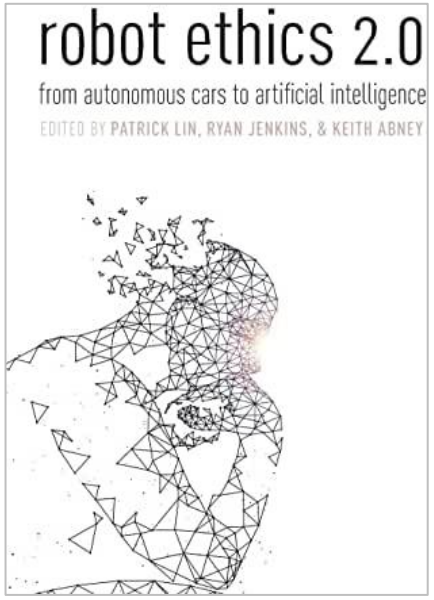












III.2 Normative Ethics

[dePoel&Royakkers2011]

Some Terminology:

- **Ethics**: The systematic reflection on morality.
- **Morality**: The totality of opinions, decisions, and actions with which people express, individually or collectively, what they think is good and right.
- **Descriptive Ethics**: The branch of ethics that describes existing morality, including customs and habits, opinions about good and evil, responsible and irresponsible behavior, and acceptable and unacceptable action.
- **Normative Ethics**: The branch of ethics that judges morality and tries to formulate normative recommendations about how to act or live.

Values, Norms, and Virtues

[dePoel&Royakkers2011]

Normative judgements can be based on:

- **Values**: Lasting convictions or matters that people feel should be strived for in general and not just for themselves to be able to lead to a good life or to realize a just society.
  - **Intrinsic Values**: Values in and of itself.
  - **Instrumental Values**: Something that is valuable in so far as it is a means to or contribute to something else that is intrinsically good or valuable.
- **Norms**: Rules that prescribe what actions are required, permitted, or forbidden.
  - **Hypothetical Norm**: A condition norm, that is, a norm which only applies under certain circumstances, usually of the form "if you want X do Y".
- **Virtues**: A certain type of human characteristics or quality.

Relativism, Universalism, and Absolutism

[dePoel&Royakkers2011]

Problematic positions concerning **normative judgements**:

- **Normative relativism**: An ethical theory that argues that **all moral points of views** – all values, norms, and virtues – are equally valid.

- **Universalism**: An ethical theory that states that there is a system of norms and values that is **universally applicable** to everyone, independent of time, place, or culture.
  - **Absolutism**: A rigid form of **universalism** in which no exception to rules are possible.

Ethical Theories

[dePoel&Royakkers2011]

Most prominent classes of theories:

- **Consequentialism**: The class of ethical theories which hold that the consequences of actions are central to the moral judgement of those actions.
  - Example: **Utilitarianism**
- **Deontological Ethics / Duty Ethics**: The class of approaches in ethics in which an action is considered morally right if it agrees with a certain moral rule (law, norm, principle).
  - Example: **Kantian Theory**
- **Virtue Ethics**: The class of ethical theories that focus on the nature of the acting person. This class of theories indicates which good or desirable characteristics people should have or develop to me moral.
  - Example: **The Good Life / *Eudaimonia (Aristotle)***
- **Care Ethics**: The class of ethical theories that emphasize the importance of relationships, and which holds that the development of moral does not come by learning general moral principles.
  - Example: **Social Ethics of Engineering**

Moral Questions and Dilemma: Examples

Moral Question:
Develop Self-Driving Car:

1)Is it morally proper to replace cars by self-driving cars?

2)Which evidence is required to make it morally acceptable to replace cars by self-driving cars?

…

Moral Dilemma:
"Trolley problem" for Driver and Self-Driving Cars:

- The cars will kill 5 persons on the road ahead, but it may also turn right where it will only kill one person.

1)Is it morally proper to turn right or not?

2)Which additional aspects influence such a decision?

(adjusted versions of the trolley problem; cf. [Foot1967, Pereira&Saptawijaya2016])

Utilitarianism

[dePoel&Royakkers2011]

- **Utilitarianism**: A type of **consequentialism** based on the utility principle. In utilitarianism, actions are judged by the amount of pleasure and pain they bring about. The action that brings the greatest happiness for the greatest number should be chosen.
  - **Hedonism**: The idea (value theory) that pleasure is the only thing that is good and to which all other things are instrumental.
  - **Utility principle**: The principle that one should choose those actions that result in the greatest happiness for the greatest number.
  - **Moral balance sheet**: A balance sheet in which the costs and benefits (pleasure and pain) for each possible action are weighted against each other. Bentham (1748-1832) proposed the drawing up of such balance sheets to determine the utility of actions. Cost-benefit analysis is a more modern variety of such balance sheets.

Moral Questions and Dilemma: Utilitarianism

Moral Question:
Develop Self-Driving Car:

1)Is it morally proper to replace cars by self-driving cars?
Yes, if the pain is less (less fatalities) and the pleasure is more (???).

2)Which evidence is required to make it morally acceptable to replace cars by self-driving cars?
Balance sheet!

Moral Dilemma:
"Trolley problem" for Driver and Self-Driving Cars:

- The cars will kill 5 persons on the road ahead, but it may also turn right where it will only kill one person.

1)Is it morally proper to turn right or not?
Yes!

2)Which additional aspects influence such a decision?
Balance sheet!

Critique of Utilitarianism (1/2)

[dePoel&Royakkers2011]

Critique: the position of the individuals cannot always be protected if the calculation of the majority outweighs the unhappiness of a few. Extensions by Mills (1806-1873):

- **Freedom principle**: The moral principle that everyone is free to strive for his/her own pleasure, if they do not deny or hinder the pleasure of others.
- **No harm principle**: The principle that one is free to do what one whishes, if no harm is done to others. Also known as the **freedom principle**.

Critique of Utilitarianism (2/3)

[dePoel&Royakkers2011]

- **Critique**: consequences cannot be foreseen objectively and often are unpredictable, unknown, or uncertain.
  ➪ "fix": consider expected consequences
- **Critique**: utilitarianism can lead to unjust division of costs and benefits
  - **Distributive justice**: The value of having a just distribution of certain important goods, like income, happiness, and career.
  - **Marginal utility**: The additional utility that is generated by an increase in a good or service (income for example). **Argument:** as an increase is more effective for poor then rich people a too unjust division of costs and benefits is avoided

Critique of Utilitarianism (3/3)

[dePoel&Royakkers2011]

- **Critique**: relations are ignored, and only individual happiness is considered
- **Critique**: in **act utilitarianism** breaking even human rights or falsify measurements can be justified and therefore **rule utilitarianism** looking at rules (in contrast to actions) is considered.
  - **Act utilitarianism**: The traditional approach to utilitarianism in which the rightness of actions is judged by the (expected) consequences of those actions.
  - **Rule utilitarianism**: A variant of utilitarianism that judges actions by judging the consequences of the rules on which these actions are based. These rules, rather than actions themselves should maximize utility.

Kantian Theory (Duty Ethic)

[dePoel&Royakkers2011]

Foundations of the **Kantian Theory** from Kant (1724-1804):

- **Autonomy**: a person oneself should be able to determine what is morally correct through reasoning (**enlightenment**: a range of ideas centered on reason as the primary source of authority and legitimacy)
- **Good will**: According to Kant, we can speak of good will if our actions are led by the categorical imperative. Kant believes that the good will is the only thing that is unconditionally good.

Categorical Imperatives

[dePoel&Royakkers2011]

The **Kantian Theory** is based on two (equivalent) **categorical imperatives**:

- **Universality principle**: Act only on that maxim which you can at the same time will that it should become a universal law.
- **Reciprocity principle**: Act as to treat humanity, whether in your own person or in that of any other, in every case as an end, never as a means.
- **Equality postulate**: The prescription to treat persons as equals, that is, with equal concern and respect. (implied by the categorical imperatives)

Moral Questions and Dilemma: Kantian Theory

Moral Question:
Develop Self-Driving Car:

1)Is it morally proper to replace cars by self-driving cars?
Moral autonomy: The view that a person himself or herself should (be able to) determine what is morally right through reasoning.
➪ **provide information about the involved risk and enable choice!**

Moral Dilemma:
"Trolley problem" for Driver and Self-Driving Cars:

■The cars will kill 5 persons on the road ahead, but it may also turn right where it will only kill one person.

1)Is it morally proper to turn right or not?
Yes, as the killed people are not used as means.

Moral Dilemma:
Variant of the "Trolley problem" for Driver and Self-Driving Cars:

■The cars cannot turn right but may steer to hit a fat person such that it will slow down and therefore not kill the 5 persons.

1)Is it morally proper to steer to hit a fat person?
No, as the fat person is used as means (and not an end).

Remark: Yes, for act utilitarianism and possibly no for rule utilitarianism

Critique of the Kantian Theory

[dePoel&Royakkers2011]

- **Critique:** According to Kant all moral laws can be derived from the categorical imperative. But are all these laws form an unambiguous and consistent system of norms?
  - **Prima facie norms**: Prima facie norms are the applicable norms, unless they are overruled by other more important norms that become evident when we take everything into consideration.
- **Critique:** Rigid adherence to moral rules can make people blind to the potential very negative consequences of their action.
  - Unitarianism and Kantian theory can become diametrically opposed concerning the moral correctness of an action.

The Good Life / *Eudaimonia* (Care Ethics)

[dePoel&Royakkers2011]

- **The Good Life / *Eudaimonia (Aristotle 384-322 BC)***: The highest good or *eudaimonia*: a state of being in which one realizes one's unique human potential. According to Aristotle, the good life is the final goal of human action.
- Each moral virtue (also referred to as character virtue by Aristotle) holds a position of equilibrium according to Aristotle. A moral virtue is the middle course between two extremes of evil.
- **Practical wisdom:** The intellectual virtue that enables one to make the right choice for action. It consists in the ability to chose the right mean between two vices.

Critiques of The Good Life

[dePoel&Royakkers2011]

- **Critique-1**: Virtue ethics are similar to the duty ethics as each virtue relates to a moral rule. However, it appears that not all obligations to act can be reduced to virtues and vice versa.
- **Critique-2**: Virtue ethics often does not give a clear clue about how to act while solving a case, in contrast with Unitarianism and Kantian ethics. But it can be argued that having the right virtues does **facilitate** responsible action.
- **Critique-3**: Can we declare a moral virtue to be good without any reservation? Kant's example for this is a cold psychopath whose virtues (e.g., self-control) make him much more terrible than he would be without those virtues.

However: professional codes of conduct often refer to some virtues.

Social Ethics of Engineering

[dePoel&Royakkers2011]

Social Ethics of Engineering: An approach to **care ethics** for engineering that focusses on the social arrangements in engineering rather than on individual decisions. If these social arrangements meet certain procedural norms the resulting decisions are considered **acceptable**.

- Norms of engagement for the participation of engineers in groups, processes, involving both engineers and non-engineers:
  - Competency,
  - Cognizance (requiring interdisciplinary skills and breath built in the group)
  - Democratic information flow,
  - Democratic teams,
  - Service-orientation,
  - Diversity,
  - Cooperativeness,
  - Creativity, and
  - Project management skills.

- Emphasize care (e.g., for safety and sustainability)
- Norms of engagement are rather similar to virtues, but the are understood at the **level of the group processes** and **social engagement** and not as individual character traits.

Critiques of Care Ethics

[dePoel&Royakkers2011]

- **Critique-1**: Care ethics are philosophically vague since is is unclear what “care” exactly entails. Therefore, care ethics are not very normative.

- **Critique-2**: Care ethics assume that caring is good, thus it can tell us neither what makes a particular attitude or action right, nor what constitutes the right way to pursue them.

- **Critique-3**: Care ethics judges a situation by means of “good care” and not according to principles. But the question is what turns “care” into “good care”?

- **Critique-4**: Care ethics like virtue ethics does not give concrete indications how one has to act in a particular situation in contrast to utilitarianism and Kantian ethics.

Comparison and Limitations

[dePoel&Royakkers2011]

Observations:

- No ethic theory is generally accepted.
- Application of theories to cases is not straightforward.
- Different ethical theories stress different aspects of a situation:
  - **Consequentialism**: focus on consequences
  - **Deontological Ethics / Duty Ethics**: focus on actions that are considered morally right or not
  - **Virtue Ethics**: focus on the nature of the acting person
  - **Care Ethics / Social Ethics of Engineering**: emphasize the importance of relationships and focus on the groups and group processes
- Depending on the case the consequences, actions, nature of the acting person, or the group may all matter!

Focus on **Normative Argumentation** for specific cases rather than try to simply apply only **one** specific theory

Project Example: Engineering support of Design, Verification & Validation of Ethical Argumentation

Outcome: Methodology, Models, Data, and Tool Prototype
Topic: Perception of Competing Argumentations for Ethical Dilemmas on <Fairness x Trustworthiness, Privacy x Safety, etc.>
Domains/Ethical Dilemma: Recommender Systems, Social Networks, Surveillance and Identification Systems, Medical diagnosis

Possible Project Tasks:

- Describe the specific ethical dilemma with examples (domain-specific or multiple-domains)
- Describe the arguments that cover the various aspects of the dilemma (use multiple definitions/understandings of an ethical principle)
- Codify arguments using some model from a methodology or tool
- Use the model to generate a critique on the dilemma, e.g., identify fallacies, false assumptions, mis-understandings, etc.
- Design an experiment to evaluate the human subjective perception of these arguments (use threats to validity to check your assumptions)
- Revise the design after executing a pilot of the experiment (colleagues)
- Run the large-scale experiment

Project Example: Engineering support of Design, Verification & Validation of Ethical Argumentation

Discussion of Findings:

- Which fallacies or false assumptions were successfully identified? Which were not?
- Which counterarguments for these fallacies are well perceived? By whom?
- Are there external factors (demographics) correlated with the findings?
- Is there any detectable confounding? Or plausibly hidden?

Implications to Engineering:

- What variations in the argumentations could be credited/blamed for a more positive/negative perception?
- Are there possible improvements in the model that codifies the arguments? Fallacies that were not identified?
- Are there plausible ways to intervene in the argumentation to improve its positive perception or to increase the chances of fallacy identification?
- Is there any pre-validation/verification of the model that could have been done to improve the results?

Suggested task (Tuesday, Nov. 16.)

How would you express and address (mitigate) the ethical dilemma according to each theory?

- Consequentialism Ethics/Utilitarianism
- Deontology Ethics/Kantian Theory
- Virtue Ethics/The Good life, Aristotle Eudaimonia
- Care Ethics/Social Ethics of Engineering
  - For the social engineering ethics, associate your comments with the answers from the previous tasks. Add references to your claims or definitions.

END