Winter Term 21/22

Artificial Intelligence, Ethics & Engineering

Lecture-6: Normative Argumentation

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

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

Christian Zöllner (christian.zoellner@hpi.de)

Normative Argumentation

[dePoel&Royakkers2011]

- **Argumentation Theory**: An interdisciplinary study of analyzing and evaluating arguments.
- **Argumentation**: A set of statements, of which one (the **conclusion**) is claimed to follow from the others (the **premises**).
- **Valid Argumentation**: An argument whose conclusion follow with necessity from its premises: if the premises are true, the conclusion must be true.
  - E.g., **Modus ponens**: “q” follows from “if p then q” and “p”
- **Fallacy**: An error or deficiency in an argument.
  - E.g., **Modus tollens**: “not p” follows from “if p then q” and “not q”

Descriptive before Prescriptive Claims/Arguments

Descriptive: The autonomous system is equipped with a decision-making module that is trained on normal and exceptional situations.

Prescriptive: The autonomous system should not make decisions that endanger people's lives. When faces with a dilemma, it should try to delegate the decision to a human.

We do not want to make prescriptive claims without proper descriptive ones.

Because we do not want to suggest (prescribe) solutions to a problem without having understood (described) the problem very well.

Types of arguments

- Deductive (if premises are true, then the conclusion is true)
- Inductive
- Abductive
- Argument by analogy
- Reduction ad absurdum

Validity versus Truth

Validity tells me that “an argument is valid if conclusion follows logically from the premises”

However, an argument being valid does not imply that the conclusion is true.

Valid but not True Argument
Premise-1: Self-driving cars do not need a human to drive FALSE
Premise-2: Human drivers are the main causes of fatal accidents FALSE
Conclusion: Self-driving cars cause fewer fatal accidents TRUE

Valid but not True Argument
Premise-1: Robots follow the Asimov laws which prioritize the safety of humans TRUE
Premise-2: Autonomous weapons are robots TRUE
Conclusion: Autonomous weapons are safe FALSE

Deductive Soundness = validity + all premises are true

Similarities and Dissimilarities

If you have shown that two things are the same, can you later discover that they are distinct?

→ Yes, by discovering some hidden property, latent variable or relationship.

If you have shown that two things are distinct, can you later discover that they are the same?

→ Yes, via some abstraction or exogenous process.

Inductive Arguments

Premise-1: Most technology that evolves tends to become safer than older technology

Premise-2: Self-Driving cars are an evolving technology

Conclusion: Hence, self-driving cars will probably be safer than regular cars

Caveat
Future not always resembles the past

Abduction

Drawing a conclusion based on the explanation that best explains a state of events, rather from evidence provided by the premises. Also known as "inference to the best explanation".

> *“When you eliminate all the impossible, whatever remains, however improbable, must be the truth*.”
> Sherlock Holmes

Deductive and Non-Deductive Arguments

[dePoel&Royakkers2011]

- **Deductive Argument**: An argument which has a conclusion that is enclosed in (implied by) the premises.
- Valid arguments are deductive (monotonic), but in practice we often use non-deductive (non-monotonic) arguments where the conclusion is logically stronger than the premises.

- **Plausibility Principle**: The principle that enumeration and supplementary argumentation in a non-deductive argumentation can make the conclusion **plausible** (**acceptable**).

Critical Questions for Non-Deductive Arguments

[dePoel&Royakkers2011]

- **Inductive Argument**: A type of non-deductive argumentation. Arguments from a particular to the general.
- **Critical Questions**: Questions belonging to a certain type of non-deductive argumentation to check the degree of plausibility of a conclusion. Example **critical questions** for an **Inductive Argument**:

1. Were the experiments carried out relevant for the conclusion?
2. Were sufficient experiments carried out to support the conclusion?
3. Are there no counterexamples?

- **Sound Argumentation**: An argumentation for which the corresponding critical questions can be answered positively and which therefore makes the conclusion plausible if the premises are true.

Types of Arguments

[dePoel&Royakkers2011]

- **Argumentation by Analogy**: A type of non-deductive argumentation. An argumentation based on comparison with another situation in which the judgment is clear. The judgement is supposed also to apply to the analogous situation. **Critical questions**:

1. Are the two situations really comparable?
2. Is the judgement for the analogous situation really clear?

- **Means-end Argumentation**: A type of non-deductive argumentation. An argumentation in which from a given end the means is derived to realize the end. **Critical questions**:

1. Does the mean realize the end? 2. Can the means be carried out?

3. Are there any side effects? 5. Is the end acceptable?

4. Are there better means to achieve the end?

Types of Arguments

[dePoel&Royakkers2011]

- **Causality Argumentation**: A type of non-deductive argumentation. An argumentation in which an expected consequence is derived from certain actions. **Critical questions**:

1. Will the given situation or action indeed lead to the expected consequence?
2. Have no issues be forgotten, for example, with respect to the expected consequences?
3. How do you determine the expected consequences and can it be justified?

- **Proof of the absurd**: A type of deductive argumentation. An argumentation in which a certain proposition is proved by showing that the negation of the proposition leads to a contradiction. **Critical questions**:

   1. Does assuming the proposition indeed lead to an inconsistency?
   2. Is the considered negation the right negation?

Types of Arguments

[dePoel&Royakkers2011]

- **Characteristic-Judgement Argumentation**: A type of non-deductive argumentation. An argumentation based on the assumption that a certain judgement about a thing or person can be derived from a certain characteristics of that thing or person. **Critical questions**:

1. Does the characteristics mentioned justifies the judgement?
2. Are the characteristics mentioned all typical for the judgment?
3. Are there any other characteristics necessary for the judgement?
4. Does the thing or person posses characteristics that justify the negation of the judgement?
5. Does the thing or person posses the characteristics mentioned?

General Types of Fallacies

[dePoel&Royakkers2011]

- **Attack on the person (ad hominem)**: attempt to discredit an argument by bringing into question the presenter and not the argument itself.
- **Confusion of laws and ethics**: “if it isn’t illegal, it is ethical”
- **Straw person**: attempt is made to miss state a person’s actual position and conclude that the original argument is a bad argument.
- **Wishful thinking (fallacy of desire)**: interpret facts, reports, events, perceptions according to what he/she would like to be the case rather than according to the actual or rational evidence.
- **Naturalistic fallacy**: deriving ought from is.
- **The Privacy fallacy**: “If you have done nothing wrong, you have nothing to worry about.”
- **Ambiguity**: play with the meaning of words or phrases.

Fallacies of Risk

[dePoel&Royakkers2011]

- **The sheer size fallacy**: “X is accepted. Y is a smaller risk than X. So, Y should be accepted.” But X and Y may not be alternatives ...
- **The fallacy of naturalness**: “X is unnatural. So, X should not be accepted.” But what mean natural?
- **The ostrich’s fallacy**: “There is no scientific proof that X is dangerous. So, X does not give rise to any unacceptable risk.”
- **The delay fallacy**: “If we wait, we will know more about X. So, no decision about X could be made now.”
- **The technocratic fallacy**: “It is an engineering issue how dangerous X is. So engineers should decide whether or not X is acceptable.”
- **The fallacy of pricing**: “We have to weight the risk of X against its benefits. So, we must put a price on the risk of X.”

Moral Problems and Moral Dilemma

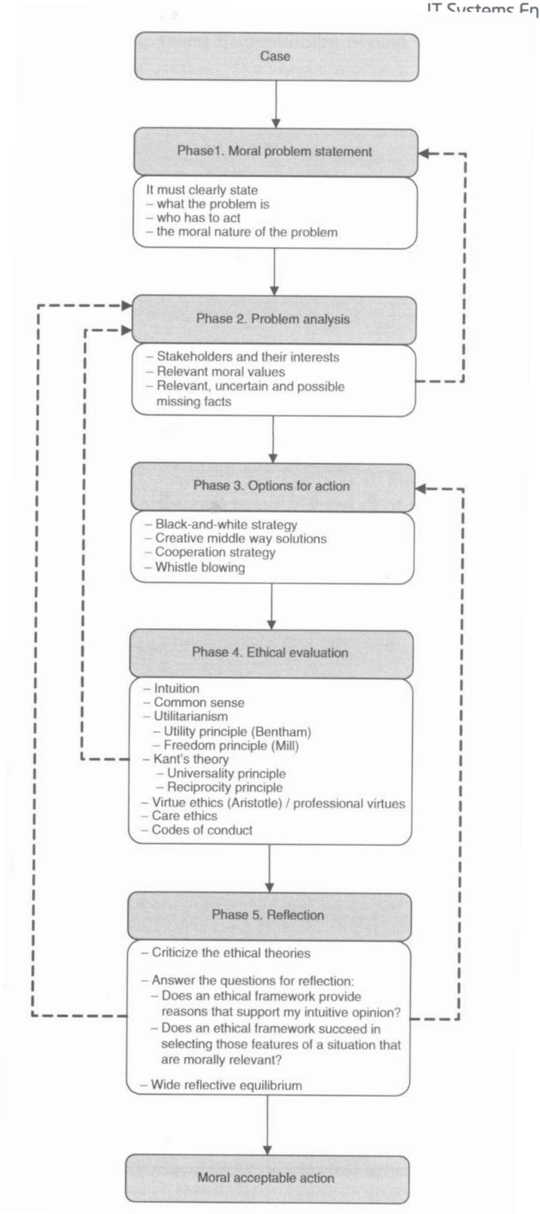
[dePoel&Royakkers2011]

- **Ill structured problem**: A problem that has no definitive formulation of the problem, may embody inconsistent problem formulations, and can only be defined during the process of solving the problem.
- **Moral problem**: A problem in which two or more positive moral values or norms cannot be fully realized at the same time.
- **Moral dilemma**: A **moral problem** with the crucial feature that the agent has only two (or a limited number of) options for action and that whatever he choses he will commit a moral wrong.

The Ethical Cycle

[dePoel&Royakkers2011]

- **Ethical Cycle**: A tool to structuring and improving moral decisions by making a systematic and thorough analysis of the moral problem, which helps to come to moral judgements and justify the final decision in moral terms.
    - Phase 1: Moral problem statement
    - Phase 2: Problem analysis
    - Phase 3: Options for action
    - Phase 4: Ethical evaluation
    - Phase 5: Reflection



Case

Phase1. Moral problem statement

It must clearly state
– what the problem is
– who has to act
– the moral nature of the problem

Phase 2. Problem analysis

– Stakeholders and their interests
– Relevant moral values
– Relevant, uncertain and possible missing facts

Phase 3. Options for action

– Black-and-white strategy
– Creative middle way solutions
– Cooperation strategy
– Whistle blowing

Phase 4. Ethical evaluation

– Intuition
– Common sense
– Utilitarianism
  – Utility principle (Bentham)
  – Freedom principle (Mill)
– Kant's theory
  – Universality principle
  – Reciprocity principle
– Virtue ethics (Aristotle) / professional virtues
– Care ethics
– Codes of conduct

Phase 5. Reflection

– Criticize the ethical theories

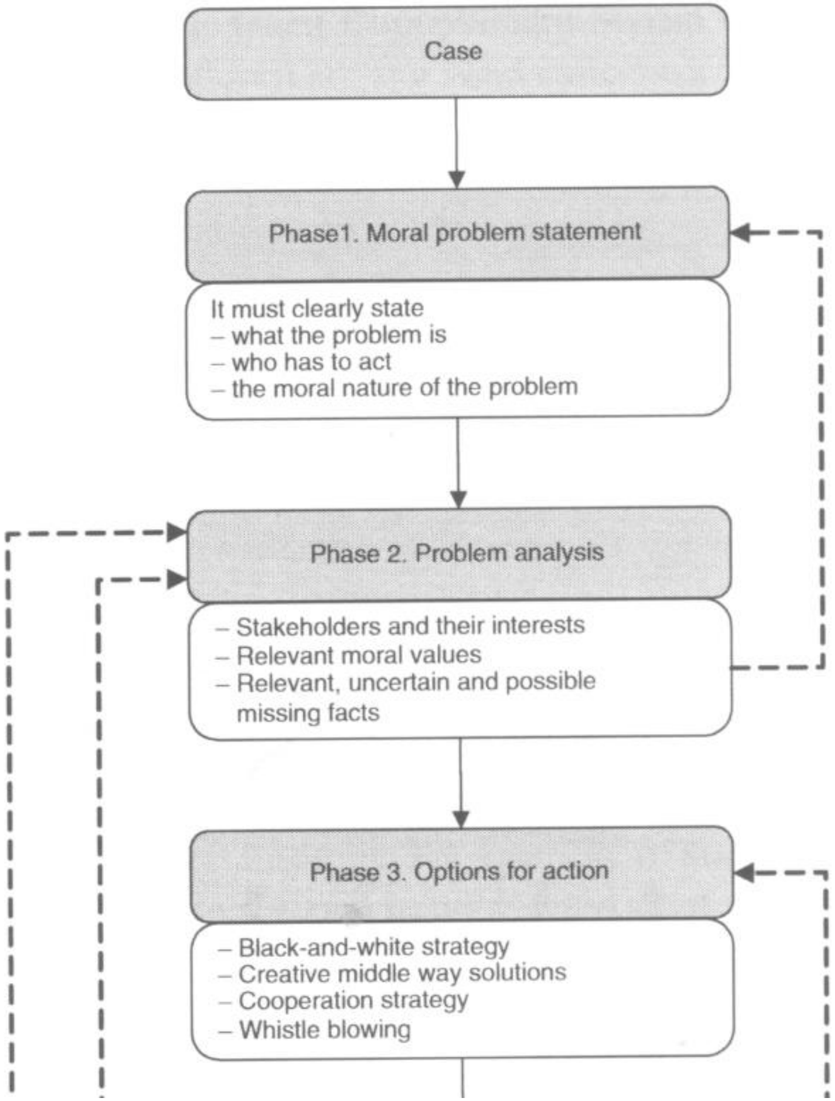
– Answer the questions for reflection:
  – Does an ethical framework provide reasons that support my intuitive opinion?
  – Does an ethical framework succeed in selecting those features of a situation that are morally relevant?

– Wide reflective equilibrium

Moral acceptable action

The Ethical Cycle

[dePoel&Royakkers2011]



Case

Phase1. Moral problem statement

It must clearly state
- what the problem is
- who has to act
- the moral nature of the problem

Phase 2. Problem analysis

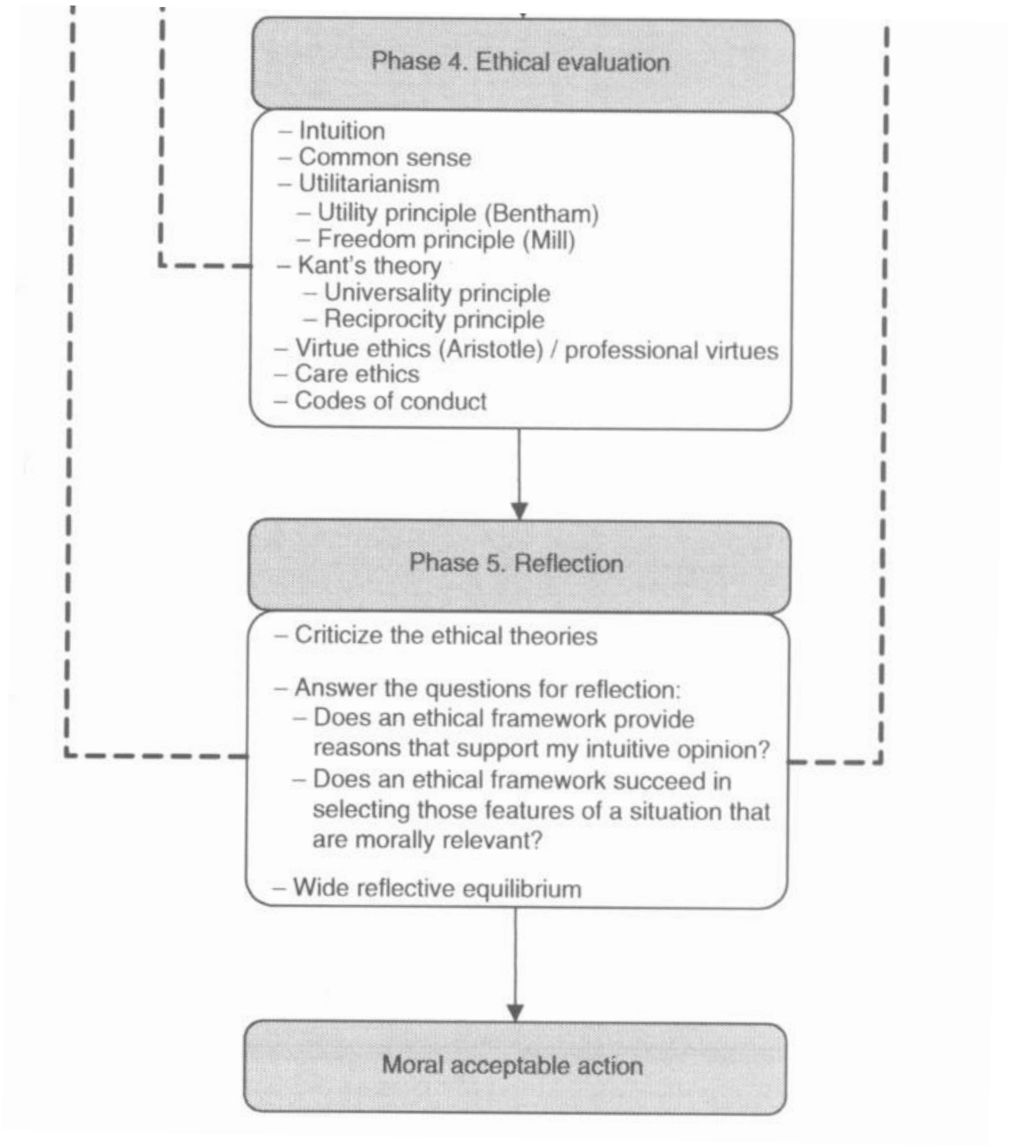
- Stakeholders and their interests
- Relevant moral values
- Relevant, uncertain and possible missing facts

Phase 3. Options for action

- Black-and-white strategy
- Creative middle way solutions
- Cooperation strategy
- Whistle blowing

The Ethical Cycle

[dePoel&Royakkers2011]



Phase 4. Ethical evaluation

- Intuition
- Common sense
- Utilitarianism
  - Utility principle (Bentham)
  - Freedom principle (Mill)
- Kant's theory
  - Universality principle
  - Reciprocity principle
- Virtue ethics (Aristotle) / professional virtues
- Care ethics
- Codes of conduct

Phase 5. Reflection

- Criticize the ethical theories
- Answer the questions for reflection:
  - Does an ethical framework provide reasons that support my intuitive opinion?
  - Does an ethical framework succeed in selecting those features of a situation that are morally relevant?
- Wide reflective equilibrium

Moral acceptable action

Ethical Questions in Design & Risk

[dePoel&Royakkers2011]

A moral problem during design relates to a value conflict with possible moral implications:

- **Value conflict**: A value conflict arises if
(1) a choice has to be be made between at least two options for which two values are relevant as choice criteria,
(2) at least two different values select at least two different options best, and
(3) the values do not trump each other.
- **Trumping (of values)**: If one value trumps another any(small) amount of the first value is worth more than any (large) amount of the second value.

Cost-Benefit Analysis

[dePoel&Royakkers2011]

- **Cost-Benefit Analysis**: A method for comparing alternatives in which the relevant advantages (benefits) and disadvantages (costs) of the options are expressed in monetary units and the overall monetary cost or benefit of each alternative is calculated.
- **Contingent Validation**: An approach to express values like safety or sustainability in monetary unity by asking people how much they are willing to pay for a certain level of safety or sustainability (for example, the preservation of a piece of beautiful nature).
- **Incommensurability**: Two (or more) values are incommensurable if they cannot be expressed or measured on a common scale or in terms of a common value measure.

Other Methods ...

[dePoel&Royakkers2011]

- **Cost-Benefit Analysis**: A method for comparing alternatives in which various decision criteria re distinguished on basis of which the alternatives score. On the basis of the score of each of th alternatives on the individual criteria, usually a total score is calculated for each alternative.
- **Threshold**: The minimal level of a (design) criterion or value that an alternative has to meet in order to be acceptable with respect to that criterion or value.
- **Reasoning**: An approach that aims at clarifying the values that underlie the conflicting design requirements by (1) identify relevant values, (2) specify the values, and (3) looking for common ground among values.
- **Value Sensitive Design**: An approach that aims at integrating values of ethical importance in a systematic way in engineering design.

Comparison of the Methods …

[dePoel&Royakkers2011]

Method	Weighting of values?	Main advantages	Main disadvantages
Cost-Benefit Analysis	All monetary	Options are made comparable	Values are treated as commensurable
Multiple Criteria Analysis	Trade-offs	Options are made comparable	Values are treated as commensurable
Thresholds	Minimal value	Selected alternatives meet thresholds	Are the thresholds independent?
Reasoning	Only related	Might solve value conflicts	Not all value conflicts can be solved this way
Value Sensitive Design	Not applicable	Can lead to better alternatives	May not solve the problem

When are Risks Acceptable?

[dePoel&Royakkers2011]

- **Informed consent**: Principle that state the activities (experiments, risks) are acceptable if people have free consented to them after being fully informed about the (potential) risk and benefits of these activities (experiments, risks).
- **Risk-Cost-Benefit Analysis**: This is a variant of regular cost-benefit analysis. The social costs for risks reduction are weighed against the social benefits offered by risk reduction, so achieving an optimal level of risk which the social benefits are highest.

When are Risks Acceptable?

[dePoel&Royakkers2011]

- **Best available technology**: As an approach to acceptable risk (or acceptable environment emissions), best available technology refers to an approach that does not prescribe a specific technology but uses the best available technological alternative as yardstick for what is acceptable.
- **Precautionary Principle**: Principle that prescribes how to deal with threats that are uncertain and/or cannot be scientifically established. In its most general form the precautionary principle has the following general format: If there is (1) a threat, which is (2) uncertain, then (3) some kind of action (4) is mandatory. This definition has four dimensions: (1) the threat dimension; (2) the uncertainty dimension; (3) the action dimension, and (4) the prescription dimension.

Summary

- Disaster or major failures of engineered systems result in the question of **(moral) responsibility**.
- **Normative Ethics** tries to formulate normative recommendations about how to act or live, but no ethic theory is generally accepted and depending on the case the consequences, actions, nature of the acting person, or the group may all matter.
- Therefore, it is suggested to use **normative argumentation** for specific moral problems and study possible actions rather than try to simply apply only **one** specific ethic theory.
- Normative argumentation can employ the **ethical cycle** and **valid** or **plausible** arguments and has to uncover **fallacies**.
- Several methods allow to study related **value conflicts** including the related moral problems (in particular risk) also during design.

Tasks

- Think about how to express the ethical dilemmas in projects into a set of arguments.
  - Ideally , try to identify argumentations with possible fallacies

- List the risks involved in your scenario
  - Discuss briefly how to tackle the risks (choose two methods)

END