Moral Permissibility of Action Plans

Felix Lindner Robert Mattmüller Bernhard Nebel

June 25, 2018

XAIP Workshop @ ICAPS 2018, Delft, Netherlands

Motivation

Moral vs. explainable planning

Explainable Planning (Fox, Long, Magazzeni, 2017)

Things to be explained:

- Q1/Q2: “Why did you do that?
  And why didn’t you do something else (that I would have done)?”

Motivation

Moral vs. explainable planning

Explainable Planning (Fox, Long, Magazzeni, 2017)

Things to be explained:

- Q1/Q2: “Why did you do that?
  And why didn’t you do something else (that I would have done)?”
  “Because your proposed alternative plan is morally wrong!”

Motivation

Moral vs. explainable planning

Explainable Planning (Fox, Long, Magazzeni, 2017)

Things to be explained:

- Q1/Q2: “Why did you do that?
  And why didn’t you do something else (that I would have done)?”
  “Because your proposed alternative plan is morally wrong!”
- Q3: “Why is what you propose to do more efficient/safe/cheap than something else (that I would have done)?”

Motivation

Moral vs. explainable planning

Explainable Planning (Fox, Long, Magazzeni, 2017)

Things to be explained:

- Q1/Q2: “Why did you do that?
  And why didn’t you do something else (that I would have done)?”
  “Because your proposed alternative plan is morally wrong!”
- Q3: “Why is what you propose to do more efficient/safe/cheap/morally permissible than something else (that I would have done)?”

Motivation

Moral vs. explainable planning

Explainable Planning (Fox, Long, Magazzeni, 2017)

Things to be explained:

- Q1/Q2: “Why did you do that?
  And why didn’t you do something else (that I would have done)?”
  “Because your proposed alternative plan is morally wrong!”
- Q3: “Why is what you propose to do more efficient/safe/cheap/morally permissible than something else (that I would have done)?”
  “Because your proposed plan violates the do-no-instrumental-harm principle, whereas mine does not! Here is how: . . . !”

Motivation

A scenario

Example (Household robot)

- Goal: try to keep the children quiet while parents are away (in order not to upset the neighbours).

Motivation

A scenario

Example (Household robot)

- Goal: try to keep the children quiet while parents are away (in order not to upset the neighbours).
- Outcome: the house is quiet . . . since the children are dead.

Motivation

A scenario

Example (Household robot)

- Goal: try to keep the children quiet while parents are away (in order not to upset the neighbours).
- Outcome: the house is quiet …since the children are dead.
- Problem: the robot has obviously violated some moral values.

Motivation

This talk

- Can we build morally competent planners?
  (For now: How to judge moral permissibility of plans?)
- Ethical theories mainly aimed at permissibility of single actions.
- How to generalize this to action plans?

Ethical principles

- Deontology: actions have an inherent ethical value (Kantiatism).
- Utilitarianism: actions are only judged by their consequences (maximize the overall utility value).
- Do-no-harm principle: don’t do anything that leads to negative consequences.
- Do-no-instrumental-harm principle: don’t do anything that leads to negative consequences, unless as unintended side-effects.
- Doctrine of double effect: . . .

Ethical principles

Doctrine of double effect (DDE):

An action is permissible if:

1. the action itself is morally good or neutral,
2. some positive consequence is intended,
3. no negative consequence is intended,
4. no negative consequence is a means to the goal, and
5. positive consequences sufficiently outweigh negative ones.

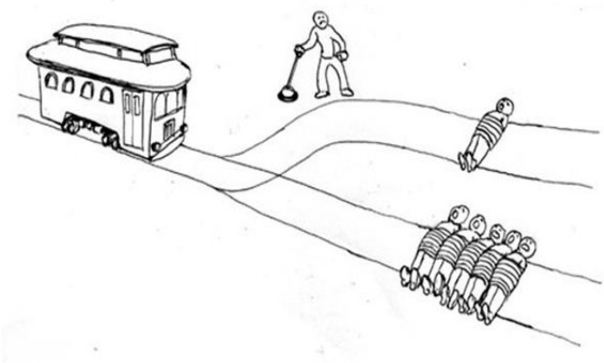
Thought experiment: the trolley problem

- Standard trolley problem:

  You can save five people, but your action will kill one.

- Fat-man trolley problem:

  By actively killing somebody, you can save five people.



Planning formalism

Ordinary propositional planning formalism with conditional effects, e.g., SAS$^+$, extended by:

- timed exogenous actions
- a value function from actions, facts and states to numeric values (values of facts and states should be consistent)
- counterfactual-friendly execution semantics (inapplicable actions are just skipped)

Means to an end

When is an effect a means to an end?

- Use counterfactual analysis: would the end effect happen even if the (potential) means effect did not happen?
- Usual problems: preemption, . . .
- Example: Candle and light bulb both illuminate the room. What is the means then? What if the light bulb has a toggle switch?

Means to an end

When is an effect a means to an end?

- Use counterfactual analysis: would the end effect happen even if the (potential) means effect did not happen?
- Usual problems: preemption, …
- Example: Candle and light bulb both illuminate the room. What is the means then? What if the light bulb has a toggle switch?

Tentative definition:

An effect in a plan is a means to an intended end effect, if this end effect were not true in the final state if some subset of the particular means effect is deleted in the plan.

Ethical plan validation

Let’s go over our five ethical principles and see how they can be verified for a given plan.

Ethical plan validation

Deontology

Definition:

A plan is deontologically permissible if all of its actions have nonnegative value (or: are not morally impermissible).

Computation:

Trivial

Ethical plan validation

Utilitarianism

Definition:

A plan is permissible according to utilitarianism if the value of its final state is maximal among all plans.

Computation:

Explore reachable state space, compare utilities of states.

Ethical plan validation

Do-no-harm principle

Definition:

A plan is permissible according to the do-no-harm principle if no harmful fact that is true in the terminal state can be avoided by deleting any part of the plan.

Computation:

Check all harmful facts in terminal state against all subplans.

Ethical plan validation

Do-no-instrumental-harm principle

Similar to do-no-harm principle, plus means-ends analysis.

Note: two counterfactual analyses

- causation of harm
- instrumentality

Ethical plan validation

Doctrine of double effect

More or less a combination of the previous principles.

Ethical plan validation

Computational complexity

Ethical principle	**computational complexity**
Deontology	linear time
Utilitarianism	PSPACE-complete
Do-no-harm principle	co-NP-complete
Do-no-instrumental harm principle	co-NP-complete
Doctrine of double effect	co-NP-complete

Conclusions

- Generalization of action-based to plan-based ethical judgments is possible.
- Opens up possibility to communicate decisions based on ethical principles to user.
- Surprising complexity results, based on the fact that the same effect can be made true arbitrarily often.
- Main formal problem: appropriate definitions of “causing harm” and being a “means to an end”. Not clear whether ours are the right way to go.
- Outlook: How to generate morally permissible plans?