Multi-Agent Systems

Moral Permissibility of Action Plans

Albert-Ludwigs-Universität Freiburg

Bernhard Nebel, Rolf Bergdoll, and Thorsten Engesser

Winter Term 2019/20

Motivation (1)

- Imagine an household robot:
  - You tell the robot that you want to go out and that you want him to take care of the children.
  - You tell him that he should try to keep the children quiet – in order not to upset the neighbours.
  - When coming back, you notice that the house is quiet . . . since the children are dead.
  - The robot has obviously violated some moral values.
- Less dramatic: You want to discuss with your robot whether some action plan is morally permissible.

Motivation (2)

- Can we build morally competent planers?
  1. How to judge action plans?
  2. How to evaluate goal choices?
  3. How to generate morally permissible action plans?
- Ethical theories are mainly aimed at the 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**: Don't do anything that leads to (some) negative consequences.
- **Asimovian**: Avoid harm if possible (either by doing something or by refraining from doing something)
- **Do-no-instrumental-harm**: Don't do anything that leads to (some) negative consequences, except it is a non-indented side-effect.
- **Principle of double effect** . . .

Principle of double effect (DDE)

An action is permissible if

1. The act itself must be morally good or neutral.
2. A positive consequence must be intended.
3. No negative consequence may be intended.
4. No negative consequence may be a means to the goal.
5. There must be proportionally grave reasons to prefer.

Planning formalism and more . . .

We assume an ordinary propositional planning formalism with conditional effects (e.g., SAS or ADL) extended by

- timed exogenous actions;
- counterfactual friendly execution semantics (unexecutable actions are simply skipped);
- an utility function u mapping from actions and facts to $\mathbb{R}$ (or $\mathbb{Z}$);
- defining the utility of a state as the sum of the utility of facts.

The Ethical Plan Validation problem

Ethical Plan Validation relative to principle X

- **Given**: A planning task (using the extended planning formalism) and a plan.
- **Question:** Is the plan morally permissible according to ethical principle X?

Deontological plan validation

- A plan is deontological permissible if all of its actions are not morally impermissible.

Theorem

The deontological plan validation problem can be decided in time linear in plan size.

Utilitarian plan validation

- Given a planning task and a plan, we can easily compute the utility of the reached final state.
- The plan is only permissible if the reached state has a maximum utility value over all reachable states.
- In so far, the validation problem is very similar to *over-subscription* planning.

Theorem

The utilitarian plan validation problem is PSPACE-complete.

Proof Sketch

- Membership: Impermissibility could be shown by guessing a higher-valued state and then non-deterministically verifying that there exists a plan to it. Hence, this problem is in NPSPACE. Since NPSPACE=PSPACE and PSPACE is closed under complement, we are done.
- Hardness: Reduce (propositional) plan non-existence to permissibility. Introduce two new operators, one has the original goal as a precondition and g as an effect. One with no precondition and f as an effect. Give g and f utility 1, and set f as the new goal. Now, the one-operator plan of making f true is permissible iff the original planning instance is unsolvable.

Do-no-harm plan validation (1)

- We could ask whether no harmful fact is true in the end. Only then we do no harm.
- → Harm could already be true in the initial state.
- Better: Do not add any harmful facts wrt. initial state.
- → Harmful fact could be removed and added again during execution.
- Next try: Do not any add *avoidable* harm.
- You can avoid harm by doing *more* or by doing *less*. We will only consider the latter option (since this is the idea behind the do-no-harm principle).
- Could harm be avoided by doing nothing?
- → Treating the entire plan as *one large action*.

Do-no-harm plan validation (2)

- Can harm be avoided by deleting a *single* action?
- → Same harm could be added be many different actions (over determination).
- More adequate: Could harmful consequences be avoided by leaving out a subset of actions?
- Note: Just leaving out prefix or suffix is not adequate, because an arbitrary set of actions spread out over the plan could be responsible.
- → Show impermissibility by guessing a harmful fact that is true in the goal, but by deleting parts of the plan can be avoided.

Theorem

The do-no-harm plan validation problem is co-NP-complete.

Proof sketch

- Membership: *Impermissibility* can be checked by a non-deterministic algorithm using only polynomial time: Guess a harmful fact f and a subset of action occurrences O. Verify that f is true in the final state of the original plan π, but not in final state of the modified plan where O is removed from π.
- Hardness: *3SAT* can be reduced to *impermissibility*. Assume a 3SAT problem instance with n variables v_i and m clauses c_j. The planning instance has variables $V = \{v_1, \dots, v_n, c_1, \dots, c_m, b\}$, for each variable v_i an action $V_i : \langle \top, v_i \rangle$, for each clause $c_j = (l_{j1} \vee l_{j2} \vee l_{j3})$ an action $C_j : \langle \top, \bigwedge_{k=1}^{3} (l_{jk} \rhd c_j) \rangle$, the action $G : \langle \top, (\bigwedge_{j=1}^{m} c_j) \rhd b \rangle$, and the action $B : \langle \top, \neg b \rangle$, with utility of $\neg b$ is -1 and 0 for all others.

Proof sketch (cont.)

- Consider the plan $V_1, \dots, V_n, C_1, \dots, C_m, G, B$ on the empty initial state, leading to a final state in which $\neg b$ is true.
- If we can delete a subset of the V_i's so that the original formula becomes statisfiable, then by deleting this set together with B, we show impermissibility.
- Similarly, impermissibility implies that the original formula is satisfiable.

Means to an end

Important notion: means to an end.

- When is an effect in a plan a means to an end?
- Use **counterfactual analysis**: Would the final intended (end) effect occur if the potential (means) effect did not happen?
- Light candle to make something visible.
- Switch light on and light candle . . . What is the means?
- Use toggle switches . . .

→ 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.

Do-no-instrumental-harm plan validation

- The means to an end definition implies that we have the same combinatorial problem as for the simpler do-no-harm principle.

Theorem

The do-no-instrumental-harm plan validation problem is co-NP-complete.

Double-effect plan validation

1. The act itself must be morally good or neutral.
2. A positive consequence must be intended.
3. No negative consequence may be intended.
4. No negative consequence may be a means to the goal.
5. There must be proportionally grave reasons to prefer.

- All criteria except for the no negative consequence may be a means to the goal condition can be checked easily.

Theorem

The double-effect plan validation problem is co-NP-complete.

Complexity Summary

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

Summary

- There is no theory about ethics in action planning.
- 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 and can interact with each other.
- Generating morally permissible plans is not straightforward (for all principles except the deontological one), because the properties can only be checked in the end and are difficult to approximate.
- Determining the complexity of goal selection permissibility is difficult for an analogous reason.

Discussion

- What could a planning algorithm and heuristics in this context look like?
- Where do the utility values come from?
- The understanding of what an **action** is is different from the computer science understanding (e.g. enter, break-in).
- Be aware that slight modelling changes can make a big difference. Example: Two lakes, two drowning persons, after the third time step, everybody drowned if not rescued: $\langle walk, walk, rescue \rangle$ is not do-no-harm permissible!

Literature I

F. Lindner, R. Mattmüller and B. Nebel. Moral Permissibility of Action Plans. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19): 7635–7642.