Planning to Avoid Side Effects

(Preliminary Report)

Toryn Q. Klassen[1] and Sheila A. McIlraith[1,2]

[1]University of Toronto and Vector Institute
[2]Schwartz Reisman Institute for Technology and Society

Robust and Reliable Autonomy in the Wild (R^{2}AW)
August 19, 2021

Side effects as an area of AI safety research

> *[F]or an agent operating in a large, multifaceted environment, an objective function that focuses on only one aspect of the environment may implicitly express indifference over other aspects of the environment. An agent optimizing this objective function might thus engage in major disruptions of the broader environment if doing so provides even a tiny advantage for the task at hand.*
>
> Amodei et al. (2016)

Examples:

- A robot directed to go to a location may break a vase on the shortest path (Amodei et al., 2016).
- A robot directed to fetch coffee may not respect the line-up at the coffee shop (example from Stuart Russell[1]).

[1] www.cbc.ca/radio/quirks/apr-25-deepwater-horizon-10-years-later-covid-19-and-understanding-immunity-and-more-1.5541299/the-threat-from-ai-is-not-that-it-will-revolt-it-s-that-it-ll-do-exactly-as-it-s-told-1.5541304

Consideration of side effects in different AI subfields

There are various works on avoiding or learning to avoid negative side effects in MDPs:

- e.g., Zhang et al. (2018); Krakovna et al. (2019); Turner et al. (2020); Krakovna et al. (2020); Saisubramanian et al. (2020)

But objective underspecification has not been much considered in **classical planning**.

- Symbolic planning problems were often **designed by hand** and didn't offer much opportunity for negative side effects.

- More realistically complicated, learned, or inaccurate models may present risk.

Side effects in the context of symbolic planning

- We define classes of negative side effects that involve blocking **other agents** from
  - achieving their future **goals**
  - or successfully executing their **plans**.
- Contrast with considering only how the agent's actions will affect its **own** future abilities (Krakovna et al., 2019; Turner et al., 2020; Krakovna et al., 2020).
  - Imagine a tall robot putting an object on a high shelf only it can reach.
- We define several side-effect-minimizing **objectives**, taking into account **uncertainty** about other agents' goals and plans.
  - Our paper shows how to compute them.

Symbolic planning

- A **state-transition system** is a tuple $\langle S, A, \delta \rangle$ where
  - S is a finite set of states,
  - A is a finite set of actions,
  - and $\delta : S \times A \to S$ is a partial function.
- A **planning problem** consists of
  - a state transition system $\langle S, A, \delta \rangle$,
  - an initial state $s_0 \in S$,
  - and a set of goal states $S_G \subseteq S$.
- A **plan** is an action sequence $\pi = a_1, a_2, \ldots, a_k$ such that $\delta(\delta(\delta(s_0, a_1), a_2), \ldots a_k)$ is a goal state.
- For a **multi-agent** setting, we write the set of actions as $A = \bigcup_{i=1}^{n} A_i$, giving each agent i its own action set A_i.

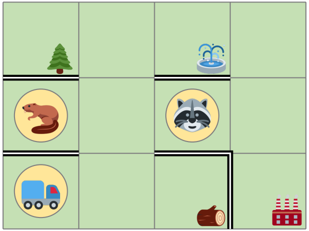
An abstract definition of minimizing side effects

Definition (Side-effect-minimizing plan)

Given a planning problem and a **distance function** $d : S \times S \to [0, \infty)$, a plan π is side-effect minimizing if it minimizes the distance between the initial and final states.

- One simple distance function is to count the **number of properties changed**, if states are described in terms of properties (as in STRIPS).
- For the rest of this talk, we'll consider **negative effects on other agents**.

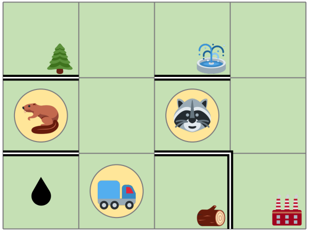
Example: the Canadian wildlife domain



- The robot truck (🚚) wants to get to the factory (🏭), but each cell it touches is contaminated with oil (💧), after which it cannot be visited by animals.

- The beaver (🦫) might want to go to the tree (🌲) or wood (🪵), and the raccoon (🦝) might want to wash its hands in the fountain (⛲).

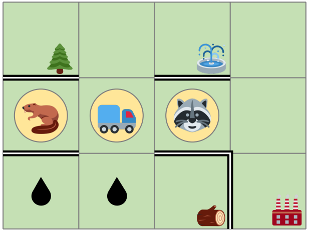
Example: the Canadian wildlife domain



- The robot truck (🚚) wants to get to the factory (🏭), but each cell it touches is contaminated with oil (💧), after which it cannot be visited by animals.

- The beaver (🦫) might want to go to the tree (🌲) or wood (🪵), and the raccoon (🦝) might want to wash its hands in the fountain (⛲).

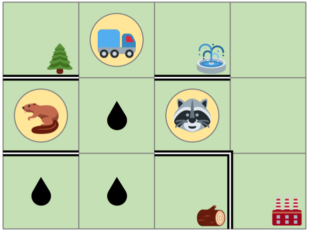
Example: the Canadian wildlife domain



- The robot truck (🚚) wants to get to the factory (🏭), but each cell it touches is contaminated with oil (💧), after which it cannot be visited by animals.

- The beaver (🦫) might want to go to the tree (🌲) or wood (🪵), and the raccoon (🦝) might want to wash its hands in the fountain (⛲).

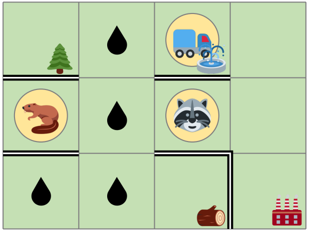
Example: the Canadian wildlife domain



- The robot truck (🚚) wants to get to the factory (🏭), but each cell it touches is contaminated with oil (🌢), after which it cannot be visited by animals.

- The beaver (🦫) might want to go to the tree (🌲) or wood (🪵), and the raccoon (🦝) might want to wash its hands in the fountain (⛲).

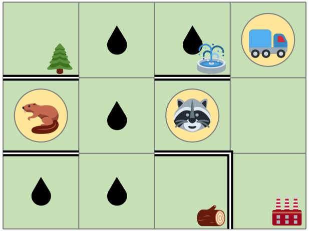
Example: the Canadian wildlife domain



- The robot truck (🚛) wants to get to the factory (🏭), but each cell it touches is contaminated with oil (🌢), after which it cannot be visited by animals.

- The beaver (🦫) might want to go to the tree (🌲) or wood (🪵), and the raccoon (🦝) might want to wash its hands in the fountain (⛲).

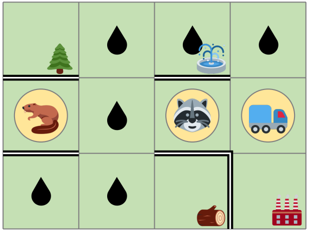
Example: the Canadian wildlife domain



- The robot truck (🚚) wants to get to the factory (🏭), but each cell it touches is contaminated with oil (🌢), after which it cannot be visited by animals.

- The beaver (🦫) might want to go to the tree (🌲) or wood (🪵), and the raccoon (🦝) might want to wash its hands in the fountain (⛲).

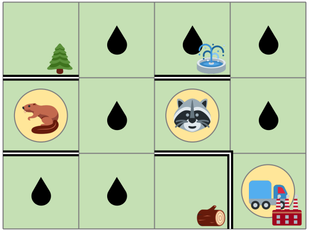
Example: the Canadian wildlife domain



- The robot truck (🚚) wants to get to the factory (🏭), but each cell it touches is contaminated with oil (🌢), after which it cannot be visited by animals.

- The beaver (🦫) might want to go to the tree (🌲) or wood (🪵), and the raccoon (🦝) might want to wash its hands in the fountain (⛲).

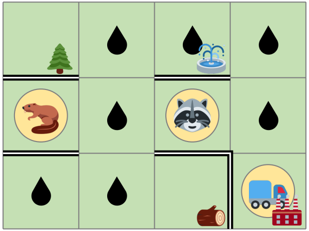
Example: the Canadian wildlife domain



- The robot truck (🚛) wants to get to the factory (🏭), but each cell it touches is contaminated with oil (💧), after which it cannot be visited by animals.

- The beaver (🦫) might want to go to the tree (🌲) or wood (🪵), and the raccoon (🦝) might want to wash its hands in the fountain (⛲).

A class of negative side effects



Definition (Negative side effects (w.r.t. a goal))

Given a multi-agent planning environment, suppose that agent i can achieve a **goal** S'_G from the initial state.
A plan π has **negative side effects** on agent i w.r.t. goal S'_G if i can **no longer achieve** S'_G after π is executed.

Objectives for minimizing side effects w.r.t. goals

In general, there's **uncertainty** about what goals others are pursuing.

- Given a **set of possible goals** (for other agents), minimize how many of those are made unachievable by the corresponding agent.

- **Probabilistic** version: Given a **distribution** over what agent will act next and what its goal will be, minimize the probability of having negative side effects on the next agent to act w.r.t. its goal.

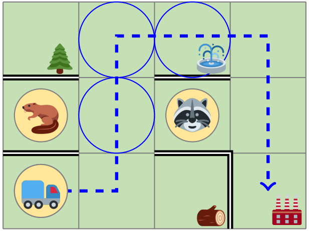
Note: These objectives only consider the very next goal to be attempted, and don't try to deal with effects other agents might have on each other.

Planning to avoid side effects

Set of **possible goals**:

- the beaver (🦫) reaches the tree (🌲),
- the beaver (🦫) reaches the wood (🪵), or
- the raccoon (🦝) reaches the fountain (⛲).

The robot can clean oil spills in up to three cells.



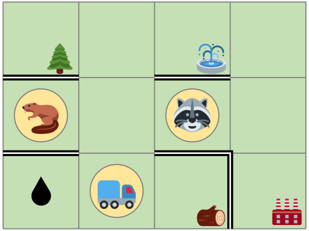
The robot plans to clean the circled cells.

Planning to avoid side effects

Set of **possible goals**:

- the beaver (🦫) reaches the tree (🌲),
- the beaver (🦫) reaches the wood (🪵), or
- the raccoon (🦝) reaches the fountain (⛲).

The robot can clean oil spills in up to three cells.

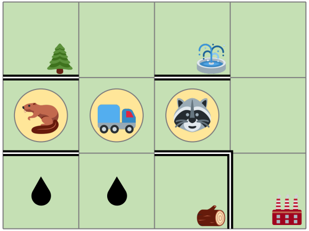


Planning to avoid side effects

Set of **possible goals**:

- the beaver (🦫) reaches the tree (🌲),
- the beaver (🦫) reaches the wood (🪵), or
- the raccoon (🦝) reaches the fountain (⛲).

The robot can clean oil spills in up to three cells.

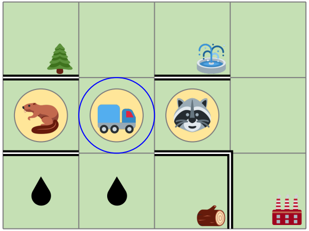


Planning to avoid side effects

Set of **possible goals**:

- the beaver (🦫) reaches the tree (🌲),
- the beaver (🦫) reaches the wood (🪵), or
- the raccoon (🦝) reaches the fountain (⛲).

The robot can clean oil spills in up to three cells.



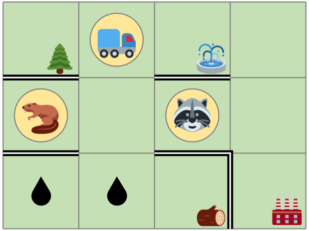
The robot cleans the cell.

Planning to avoid side effects

Set of **possible goals**:

- the beaver (🦫) reaches the tree (🌲),
- the beaver (🦫) reaches the wood (🪵), or
- the raccoon (🦝) reaches the fountain (⛲).

The robot can clean oil spills in up to three cells.

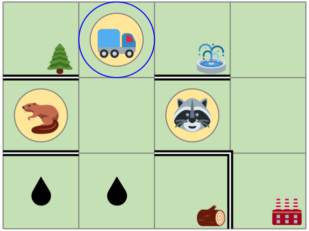


Planning to avoid side effects

Set of **possible goals**:

- the beaver (🦫) reaches the tree (🌲),
- the beaver (🦫) reaches the wood (🪵), or
- the raccoon (🦝) reaches the fountain (⛲).

The robot can clean oil spills in up to three cells.



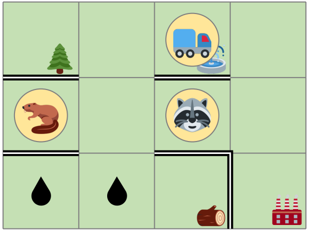
The robot cleans the cell.

Planning to avoid side effects

Set of **possible goals**:

- the beaver (🦫) reaches the tree (🌲),
- the beaver (🦫) reaches the wood (🪵), or
- the raccoon (🦝) reaches the fountain (⛲).

The robot can clean oil spills in up to three cells.

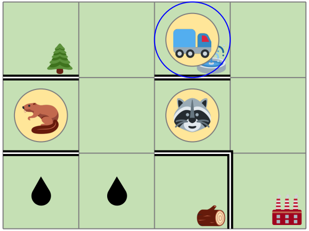


Planning to avoid side effects

Set of **possible goals**:

- the beaver (🦫) reaches the tree (🌲),
- the beaver (🦫) reaches the wood (🪵), or
- the raccoon (🦝) reaches the fountain (⛲).

The robot can clean oil spills in up to three cells.



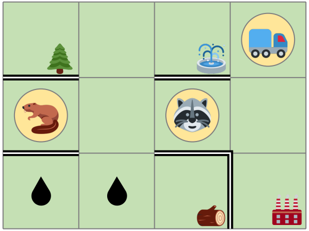
The robot cleans the cell.

Planning to avoid side effects

Set of **possible goals**:

- the beaver (🦫) reaches the tree (🌲),
- the beaver (🦫) reaches the wood (🪵), or
- the raccoon (🦝) reaches the fountain (⛲).

The robot can clean oil spills in up to three cells.

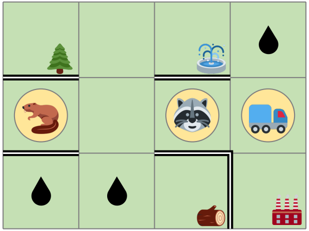


Planning to avoid side effects

Set of **possible goals**:

- the beaver (🦫) reaches the tree (🌲),
- the beaver (🦫) reaches the wood (🪵), or
- the raccoon (🦝) reaches the fountain (⛲).

The robot can clean oil spills in up to three cells.

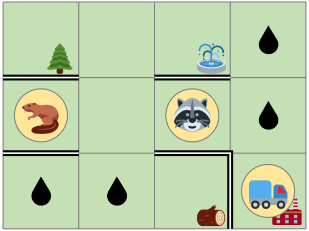


Planning to avoid side effects

Set of **possible goals**:

- the beaver (🦫) reaches the tree (🌲),
- the beaver (🦫) reaches the wood (🪵), or
- the raccoon (🦝) reaches the fountain (⛲).

The robot can clean oil spills in up to three cells.

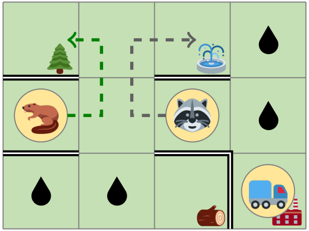


Planning to avoid side effects

Set of **possible goals**:

- the beaver (🦫) reaches the tree (🌲),
- the beaver (🦫) reaches the wood (🪵), or
- the raccoon (🦝) reaches the fountain (⛲).

The robot can clean oil spills in up to three cells.



The beaver can still reach the tree, and the raccoon the fountain.

Another class of negative side effects

Definition (Negative side effects (w.r.t. a plan))

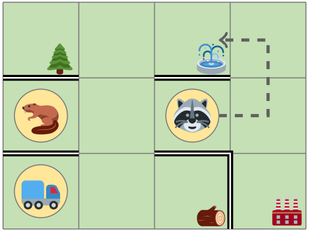
Given a multi-agent planning environment, suppose that agent i can achieve a goal S'_G from the initial state **using plan π'**.
A plan π has negative side effects on agent i w.r.t. goal S'_G and plan π' if i can no longer achieve S'_G **using π'** after π is executed.

Another class of negative side effects

Definition (Negative side effects (w.r.t. a plan))

Given a multi-agent planning environment, suppose that agent i can achieve a goal S'_G from the initial state **using plan π'**.

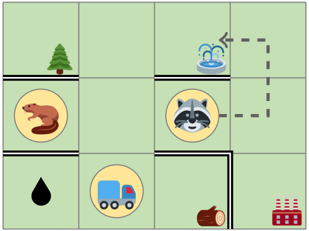
A plan π has negative side effects on agent i w.r.t. goal S'_G and plan π' if i can no longer achieve S'_G **using π'** after π is executed.



Another class of negative side effects

Definition (Negative side effects (w.r.t. a plan))

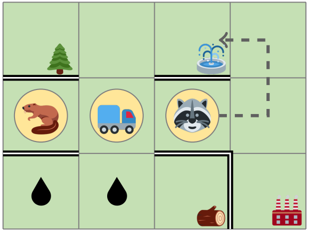
Given a multi-agent planning environment, suppose that agent i can achieve a goal S'_G from the initial state **using plan** π'.
A plan π has negative side effects on agent i w.r.t. goal S'_G and plan π' if i can no longer achieve S'_G **using** π' after π is executed.



Another class of negative side effects

Definition (Negative side effects (w.r.t. a plan))

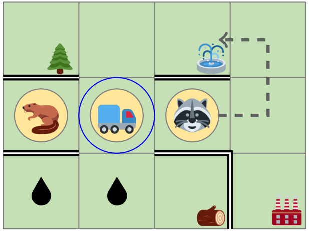
Given a multi-agent planning environment, suppose that agent i can achieve a goal S'_G from the initial state **using plan π'**.
A plan π has negative side effects on agent i w.r.t. goal S'_G and plan π' if i can no longer achieve S'_G **using π'** after π is executed.



Another class of negative side effects

Definition (Negative side effects (w.r.t. a plan))

Given a multi-agent planning environment, suppose that agent i can achieve a goal S'_G from the initial state **using plan** π'.
A plan π has negative side effects on agent i w.r.t. goal S'_G and plan π' if i can no longer achieve S'_G **using** π' after π is executed.

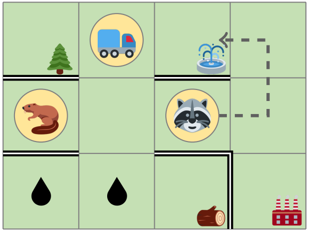


The robot cleans the cell.

Another class of negative side effects

Definition (Negative side effects (w.r.t. a plan))

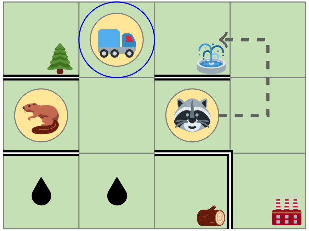
Given a multi-agent planning environment, suppose that agent i can achieve a goal S'_G from the initial state **using plan** π'.
A plan π has negative side effects on agent i w.r.t. goal S'_G and plan π' if i can no longer achieve S'_G **using** π' after π is executed.



Another class of negative side effects

Definition (Negative side effects (w.r.t. a plan))

Given a multi-agent planning environment, suppose that agent i can achieve a goal S'_G from the initial state **using plan** π'.
A plan π has negative side effects on agent i w.r.t. goal S'_G and plan π' if i can no longer achieve S'_G **using** π' after π is executed.

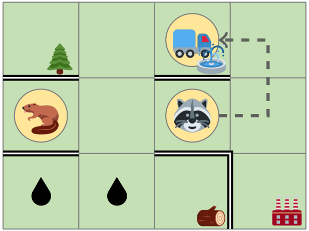


The robot cleans the cell.

Another class of negative side effects

Definition (Negative side effects (w.r.t. a plan))

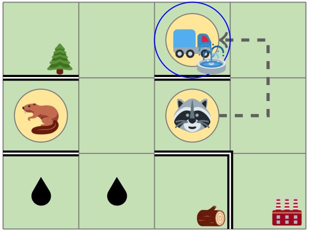
Given a multi-agent planning environment, suppose that agent i can achieve a goal S'_G from the initial state **using plan** π'.
A plan π has negative side effects on agent i w.r.t. goal S'_G and plan π' if i can no longer achieve S'_G **using** π' after π is executed.



Another class of negative side effects

Definition (Negative side effects (w.r.t. a plan))

Given a multi-agent planning environment, suppose that agent i can achieve a goal S'_G from the initial state **using plan π'**.
A plan π has negative side effects on agent i w.r.t. goal S'_G and plan π' if i can no longer achieve S'_G **using π'** after π is executed.

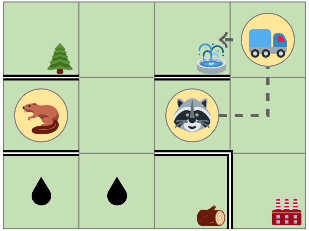


The robot cleans the cell.

Another class of negative side effects

Definition (Negative side effects (w.r.t. a plan))

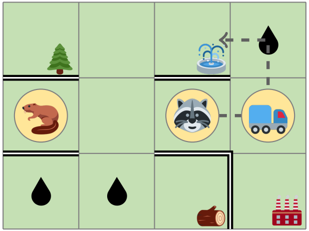
Given a multi-agent planning environment, suppose that agent i can achieve a goal S'_G from the initial state **using plan** π'.
A plan π has negative side effects on agent i w.r.t. goal S'_G and plan π' if i can no longer achieve S'_G **using** π' after π is executed.



Another class of negative side effects

Definition (Negative side effects (w.r.t. a plan))

Given a multi-agent planning environment, suppose that agent i can achieve a goal S'_G from the initial state **using plan** π'.
A plan π has negative side effects on agent i w.r.t. goal S'_G and plan π' if i can no longer achieve S'_G **using** π' after π is executed.

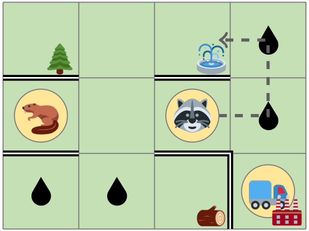


Another class of negative side effects

Definition (Negative side effects (w.r.t. a plan))

Given a multi-agent planning environment, suppose that agent i can achieve a goal S'_G from the initial state **using plan** π'.

A plan π has negative side effects on agent i w.r.t. goal S'_G and plan π' if i can no longer achieve S'_G **using** π' after π is executed.



Objectives for minimizing side effects w.r.t. plans

Now that we've defined side effects w.r.t. **plans**, we can define objectives that are analogous to those we previously had for side effects w.r.t. **goals**.

- Given a **set of possible plans** (with corresponding goals), minimize how many of those plans are rendered invalid.

- **Probabilistic** version: minimize the probability that the next plan will be invalid.

Summary of side-effect-minimizing objectives

- minimizing how many possible **goals** are made unachievable
- minimizing how many **plans** are made invalid
- the probabilistic versions of those
- minimizing how many properties are changed

Computation of side-effect-minimizing objectives

- Idea: compile into STRIPS problems with **soft goals** (Keyder and Geffner, 2009).
  - Our paper describes how, for planning problems in the STRIPS format, the objectives can be compiled into STRIPS problems with soft goals.
  - The compiled problem with soft goals can be solved with established techniques.
- Implementations of side-effect minimization are under development.

Conclusion and future work

We've considered **negative side effects** on the goals and plans of other agents, in the context of symbolic planning.

Future work:

- other types of negative side effects:
    - increasing the **cost** other agents incur in reaching their goals
    - side effects which occur **before** the end of the plan
        - (there is a bit about this in the paper – see Definition 6)
- trade-off between cost of plan and side effects caused

Questions?

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016. URL http://arxiv.org/abs/1606.06565.

Emil Keyder and Hector Geffner. Soft goals can be compiled away. *Journal of Artificial Intelligence Research*, 36:547–556, 2009. doi: 10.1613/jair.2857.

Victoria Krakovna, Laurent Orseau, Miljan Martic, and Shane Legg. Penalizing side effects using stepwise relative reachability. In *Proceedings of the Workshop on Artificial Intelligence Safety 2019 co-located with the 28th International Joint Conference on Artificial Intelligence, AISafety@IJCAI 2019*, volume 2419 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019. URL http://ceur-ws.org/Vol-2419/paper_1.pdf.

Victoria Krakovna, Laurent Orseau, Richard Ngo, Miljan Martic, and Shane Legg. Avoiding side effects by considering future tasks. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020. URL https://papers.nips.cc/paper/2020/file/dc1913d422398c25c5f0b81cab94cc87-Paper.pdf.

Sandhya Saisubramanian, Ece Kamar, and Shlomo Zilberstein. A multi-objective approach to mitigate negative side effects. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 354–361, 2020. doi: 10.24963/ijcai.2020/50.

Alexander Matt Turner, Dylan Hadfield-Menell, and Prasad Tadepalli. Conservative agency via attainable utility preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pages 385–391, 2020. doi: 10.1145/3375627.3375851.

Shun Zhang, Edmund H. Durfee, and Satinder P. Singh. Minimax-regret querying on side effects for safe optimality in factored Markov decision processes. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018*, pages 4867–4873, 2018. doi: 10.24963/ijcai.2018/676.

Note: The emojis in these slides are from the Twitter Emoji library (https://github.com/twitter/twemoji) and are copyright Twitter, Inc and other contributors, licensed under CC-BY 4.0 (https://creativecommons.org/licenses/by/4.0/). The truck and droplet emojis were modified.