An AI Safety Threat from Learned Planning Models

Toryn Q. Klassen[1,2,3] Sheila A. McIlraith[1,2,3] Christian Muise[4]

[1]Department of Computer Science, University of Toronto, Toronto, Canada
[2]Vector Institute for Artificial Intelligence, Toronto, Canada
[3]Schwartz Reisman Institute for Technology and Society, Toronto, Canada
[4]School of Computing, Queen's University, Kingston, Canada

ICAPS'22 Workshop on Reliable Data-Driven Planning and Scheduling (RDDPS)
June 17, 2022

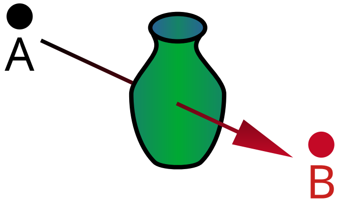
Position

Using **learned planning models** presents both

- a possible **AI safety threat**:
  - people may be more likely to **underspecify** their goals;
- and also a **research opportunity** to make planning more safe.

The threat of side effects from underspecified objectives

- **AI safety** issue: people may create **underspecified objectives**, which can be satisfied in ways that cause negative **side effects**.[1]
- The classic example of a **side effect**: a robot breaks a **vase** because it wasn't told not to.



- This problem has mostly been considered in Markov Decision Processes (MDPs) or similar formalisms, and often with reinforcement learning (RL).

[1] D. Amodei, C. Olah, J. Steinhardt, P. F. Christiano, J. Schulman, and D. Mané. "Concrete Problems in AI Safety". In: *arXiv preprint arXiv:1606.06565* (2016).

Why consider side effects in symbolic planning?

Informal Definition (Side effect)

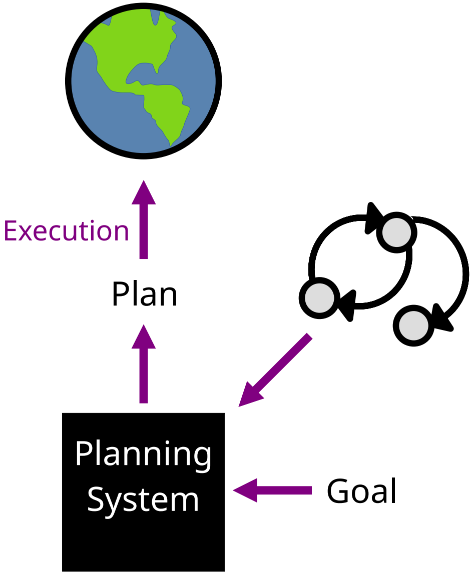
A **side effect** of a plan is any change **in the real world** caused by the execution of the plan, that was not prescribed explicitly as part of the goal.

- With learned models, objective underspecification may become an increasingly important issue for **symbolic planning systems**.
- Investigating side effects in more **restricted settings** (e.g., STRIPS or FOND planning) may
  - allow for finding different, **more efficient algorithms**, and
  - make it easier to **develop concepts** which can later be generalized.

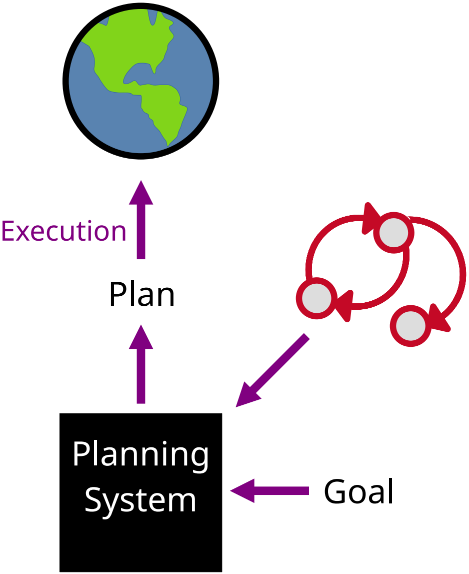
In this talk

- reasons that **planning objectives** may be underspecified and how **learned models** may make that more likely
- algorithmic approaches to **avoiding side effects**
- future directions

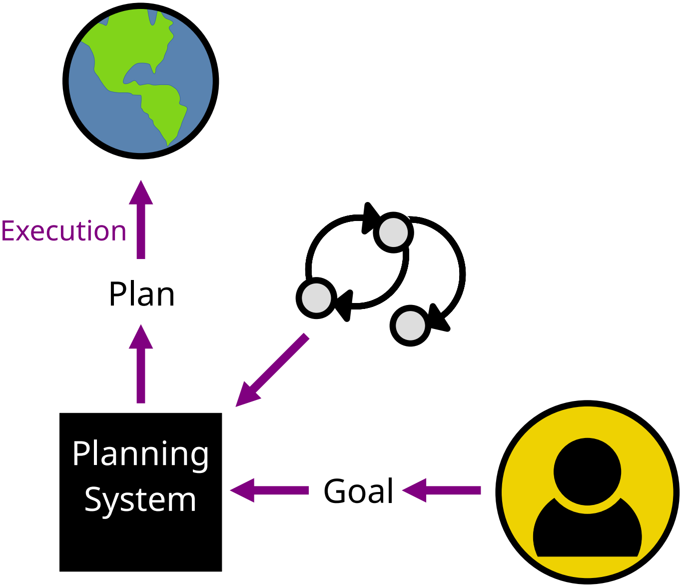
Reasons for objective underspecification to arise



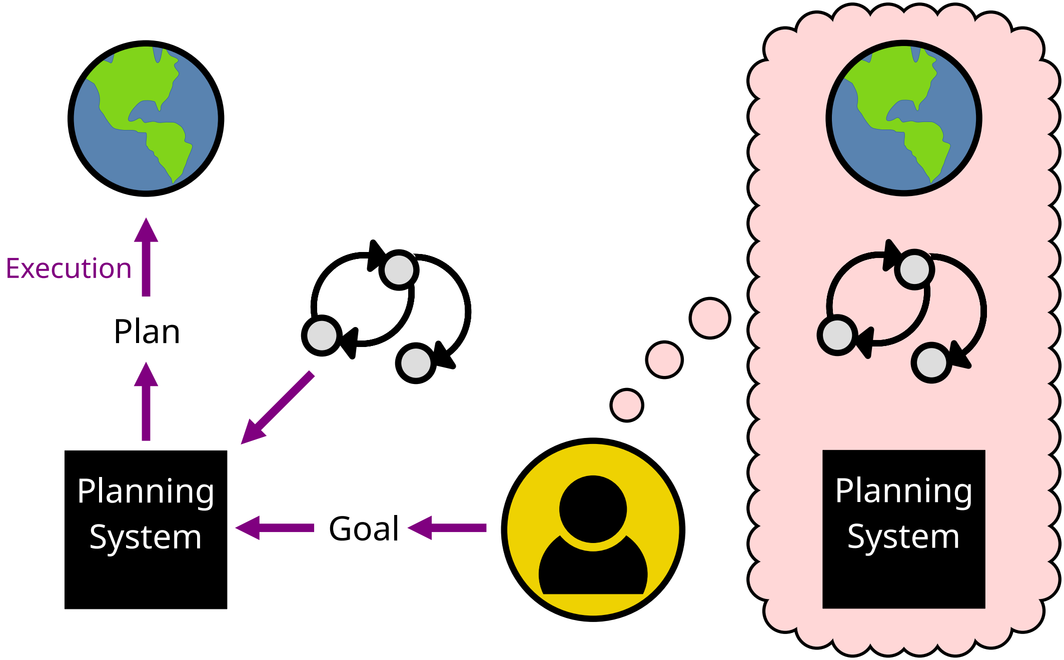
Reasons for objective underspecification to arise



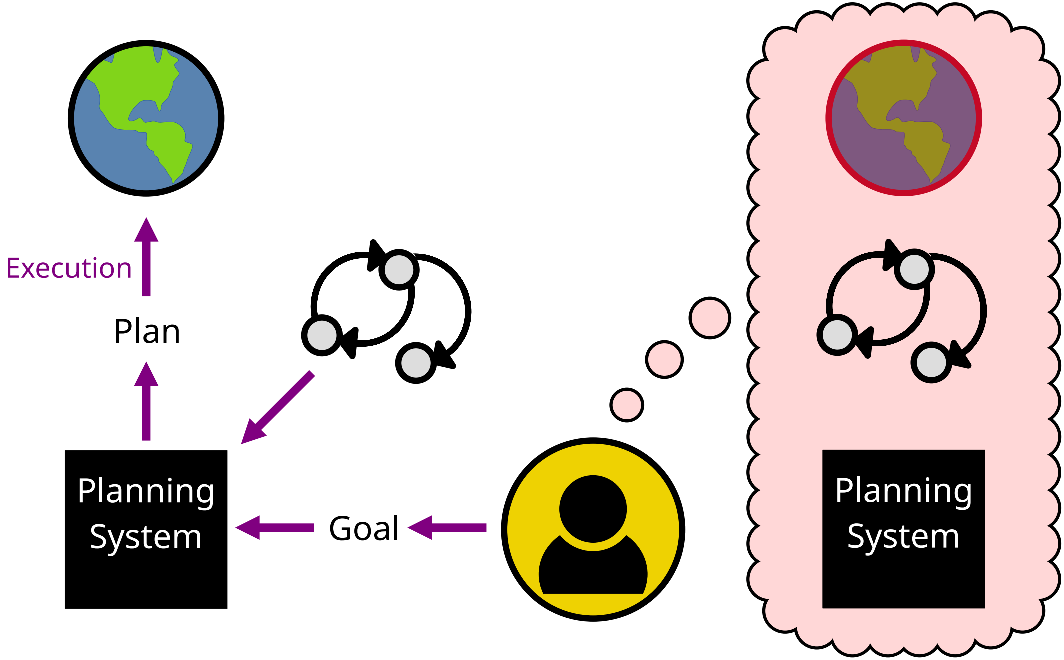
Reasons for objective underspecification to arise



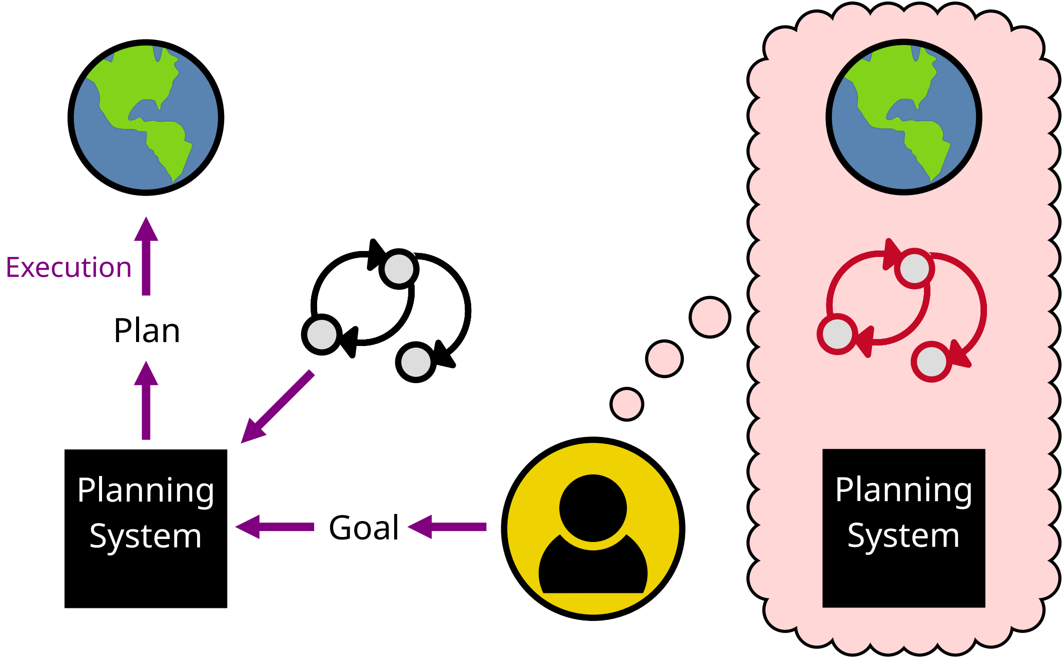
Reasons for objective underspecification to arise



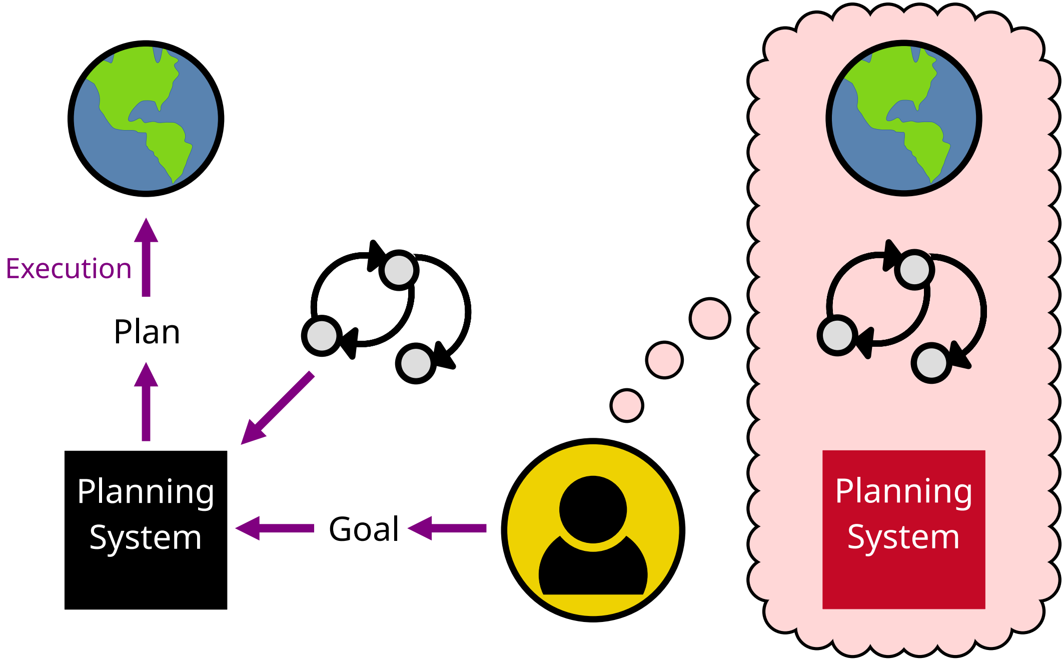
Reasons for objective underspecification to arise



Reasons for objective underspecification to arise



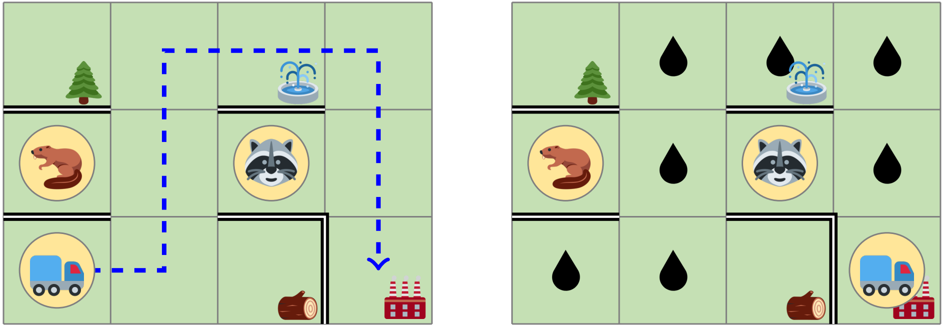
Reasons for objective underspecification to arise



Learned models may provide large vocabularies.

- Large **vocabularies** (of fluents) may allow for **representing side effects**.
- **Automated** methods can **recognize** represented side effects before plan execution and try to deal with them
  - on their own – e.g., by trying to **minimize** how many fluents are changed,
  - or by **consulting a human** to determine which side effects are **negative**.

The Canadian wildlife domain[2]



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

[2] T. Q. Klassen, S. A. McIlraith, C. Muise, and J. Xu. "Planning to Avoid Side Effects". In: *AAAI*. 2022.

Algorithms for avoiding side effects

We considered a number of algorithms, which minimize different things:[3]

1. how many **fluents** are changed
2. how many possible **goals** are made **unreachable** for other agents
   (given a set of possible goal-agent pairs)
3. how many goals are made **unreachable** for agents **using particular policies**
   (given a set of possible goal-policy pairs)

These optimization problems are compiled into planning problems with costs.

[3] T. Q. Klassen, S. A. McIlraith, C. Muise, and J. Xu. "Planning to Avoid Side Effects". In: *AAAI*. 2022.

Avoiding negative side effects interactively

- ask the human **what features** the plan is **allowed** to change[4]
- generate a **diverse set** of plans, and ask the human to **pick** the best one[5]
- learn from other forms of feedback, like human **approval** of actions[6]

[4] S. Zhang, E. H. Durfee, and S. P. Singh. "Minimax-Regret Querying on Side Effects for Safe Optimality in Factored Markov Decision Processes". In: *IJCAI*. 2018, pp. 4867–4873.

[5] T. A. Nguyen, M. B. Do, A. Gerevini, I. Serina, B. Srivastava, and S. Kambhampati. "Generating diverse plans to handle unknown and partially known user preferences". In: *Artificial Intelligence* 190 (2012), pp. 1–31.

[6] S. Saisubramanian, E. Kamar, and S. Zilberstein. "A Multi-Objective Approach to Mitigate Negative Side Effects". In: *IJCAI*. 2020, pp. 354–361.

Summary

Learned planning models

- may raise the risk of **incomplete** goal specifications being used,
  - which may be satisfied by plans that cause **negative side effects**,
- but may have sufficient vocabularies to represent, and allow algorithms to **avoid**, some side effects.

Some possible future directions

- To **minimize human effort**, incorporate additional **(possibly learned) information** into the planning process, e.g.,
  - **possible goals** of other agents that shouldn't be interfered with,
  - or **social norms**.
- **Execution monitoring** that keeps track not just of whether the goal is still achievable but of what side effects might occur or had occurred?
- New **benchmarks** or **competitions** for avoiding (negative) side effects?