A preference-based approach to machine ethics for automated planning

PhD. Thesis Defense - Martin Jedwabny

Rapporteur	Jean-Gabriel GANASCIA, PU, Sorbonne Université, Paris
Rapporteur	Felipe MENEGUZZI, PU, University of Aberdeen, Aberdeen
Examinatrice	Aurélie BEYNIER, MCF HDR, Sorbonne Université, Paris
Examinatrice	Anne LAURENT, PU, Université de Montpellier
Directrice	Madalina CROITORU, PU, Université de Montpellier
Co-encadrant	Pierre BISQUERT, CR, INRAE

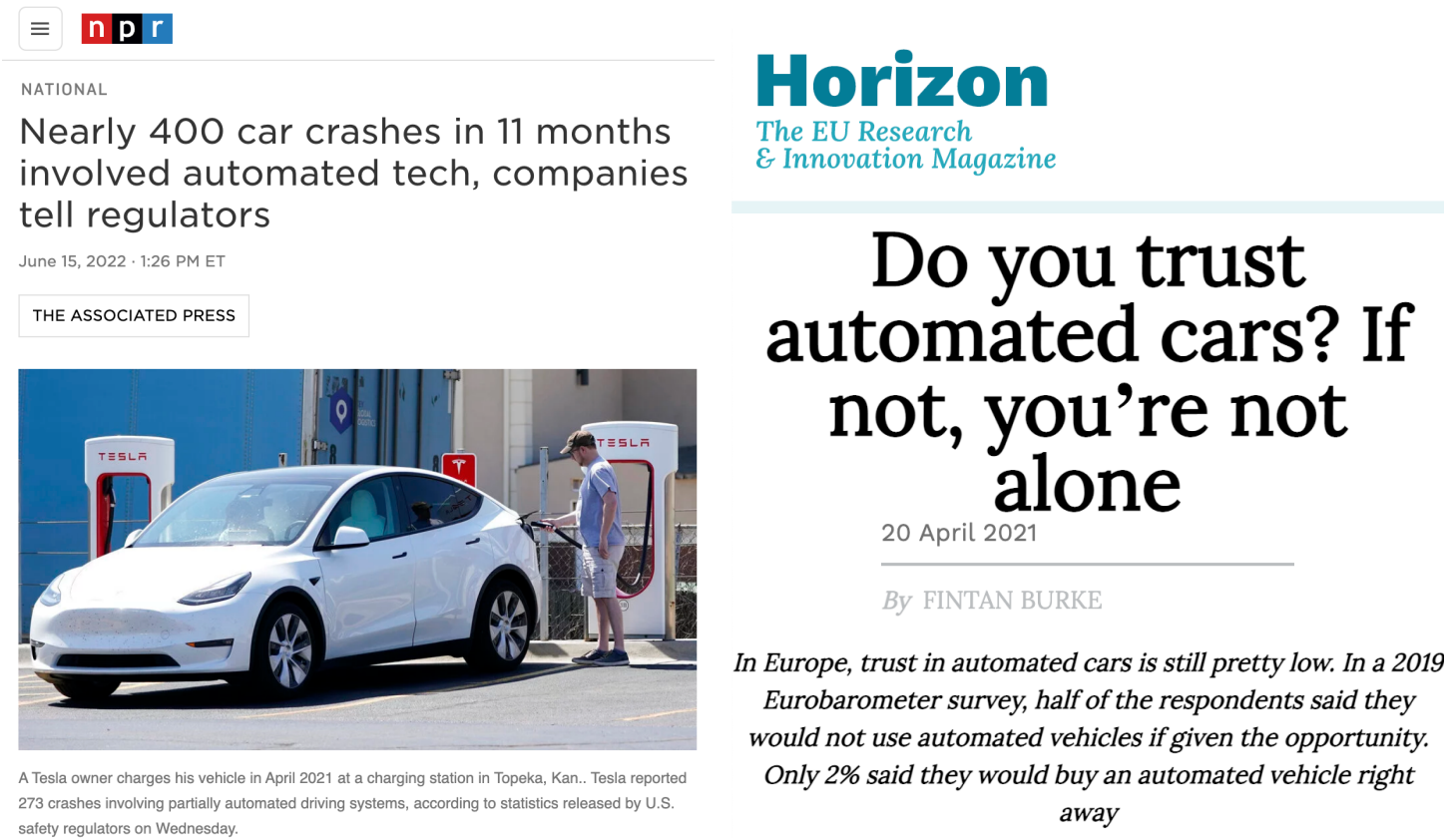
December 2nd 2022

LIRMM, Inria BOREAL, Univ Montpellier, CNRS, Montpellier, France

Introduction

Introduction

Recent years of progress in computer science have resulted in the *widespread use of AI*, **but** the introduction of automated agents in certain domains has stirred much **public concern**:



NATIONAL

Nearly 400 car crashes in 11 months involved automated tech, companies tell regulators

June 15, 2022 · 1:26 PM ET

THE ASSOCIATED PRESS

A Tesla owner charges his vehicle in April 2021 at a charging station in Topeka, Kan.. Tesla reported 273 crashes involving partially automated driving systems, according to statistics released by U.S. safety regulators on Wednesday.

Horizon
The EU Research & Innovation Magazine

Do you trust automated cars? If not, you're not alone

20 April 2021

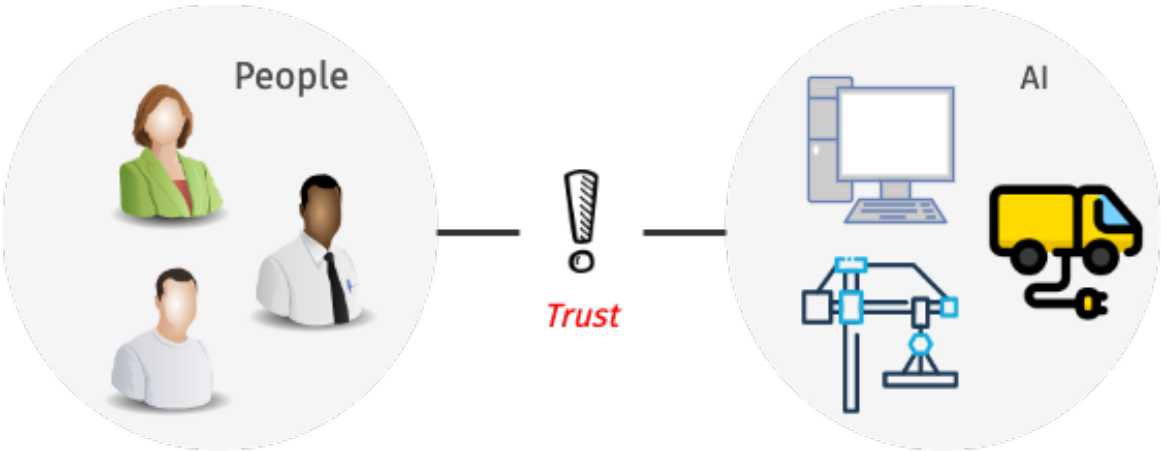
By FINTAN BURKE

In Europe, trust in automated cars is still pretty low. In a 2019 Eurobarometer survey, half of the respondents said they would not use automated vehicles if given the opportunity. Only 2% said they would buy an automated vehicle right away

Figure 1: News articles from [NPR Website, 2022] and [Horizon Magazine Website, 2021]

Introduction

How can we develop systems we can **trust**?



Two ways research has identified for developing trustworthy AI:

- Ethically-aligned agents [Shahriari and Shahriari, 2017], and
- Explainable AI (XAI) [Arrieta et al., 2020].

Introduction

- **Machine ethics** [Anderson and Anderson, 2007] is the subfield of artificial intelligence (AI) that studies the *automation of ethical reasoning*.

- **However**, determining which decision is the *most ethical* can be a very hard problem:

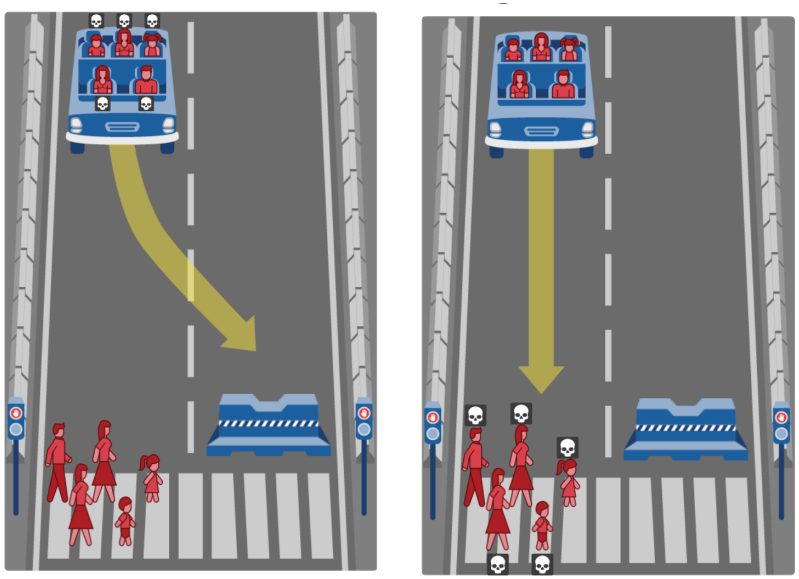


Figure 2: Moral Machine [Awad et al., 2018] example case.

Introduction

- **However**, determining which decision is the *most ethical* can be a very hard problem:

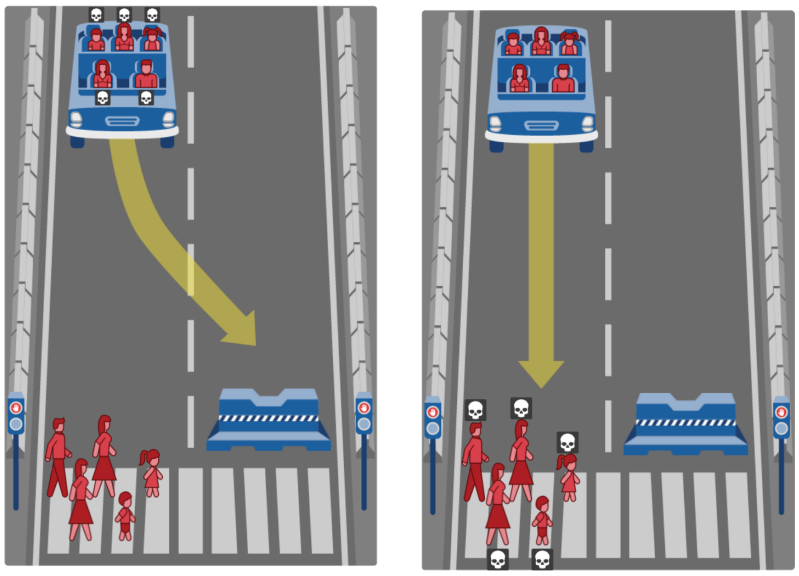


Figure 3: Moral Machine [Awad et al., 2018] example case.

- Permissibility?
- Utilities?
- Causality?

Introduction

Machine ethics *implementations* can be **classified** [Moor, 2006] into:

1. Top-down: **explicitly model** the rightness of action,

2. Bottom-up: **learn** what is ethical from human behavior,

3. Hybrid: a **mix**, i.e: learning following a high-level symbolic theory.

Introduction

⚠ **Problem 1 (top-down)**: too **restrictive** and adaptations require the opinion of **experts**.

⚠ **Problem 2 (bottom-up)**: lack of **transparency** and **explainability**.

⚠ **Problem 3 (hybrid)**: most do not handle either reasoning **many steps** in advance, or **noisy datasets**.

Introduction

★ **Research question**: how can an automated agent

1. **Assess** actions according to many **ethical theories** under the same framework,

2. In a context that allows reasoning **many steps** in advance,

3. And **align** its behavior to societal opinions of ethics, i.e: right and wrong?

Introduction

★ **Research question**: how can an automated agent

1. **Assess** actions according to many **ethical theories** under the same framework,

   ⇒ **Symbolic model of ethics**.

2. In a context that allows reasoning **many steps** in advance,

   ⇒ **Planning**.

3. And **align** its behavior to societal opinions of ethics, i.e: right and wrong?

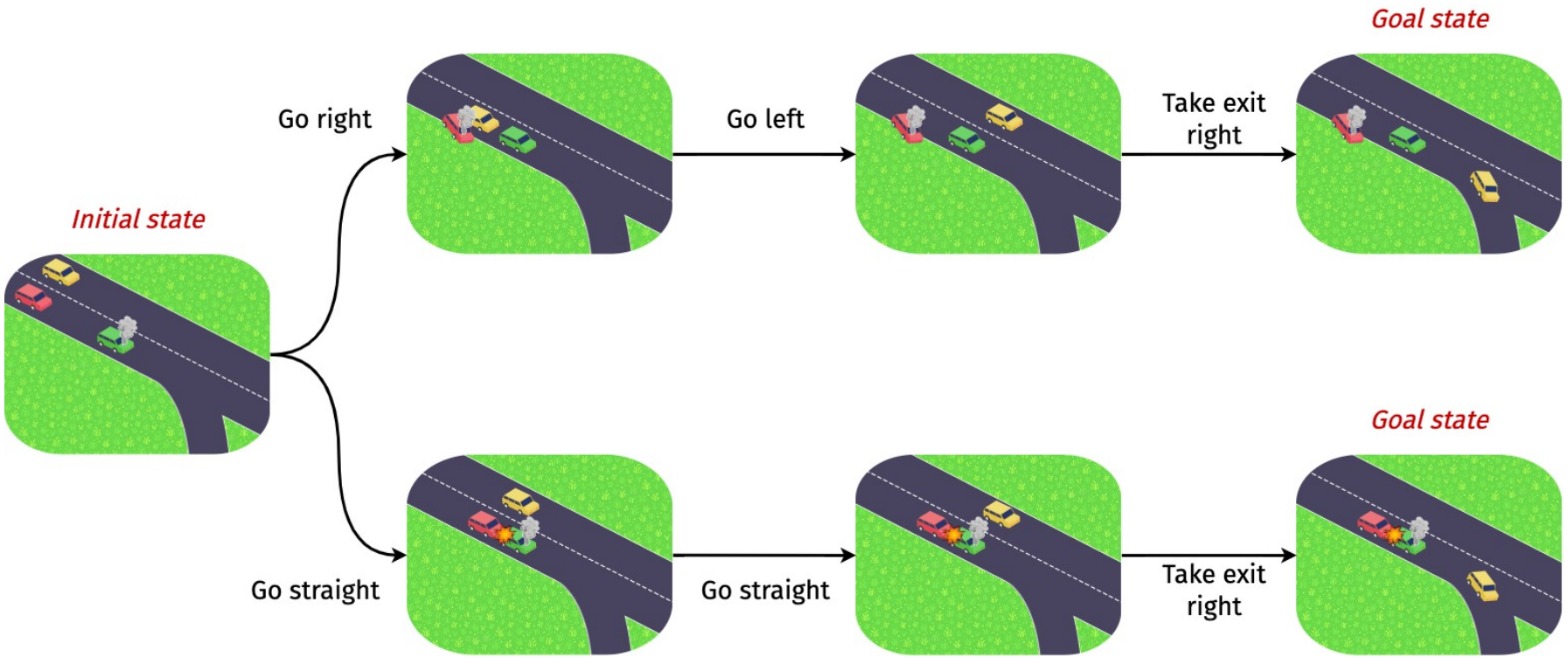
   ⇒ **Learning preferences**.

Outline

Background notions: classical planning

Background notions: classical planning

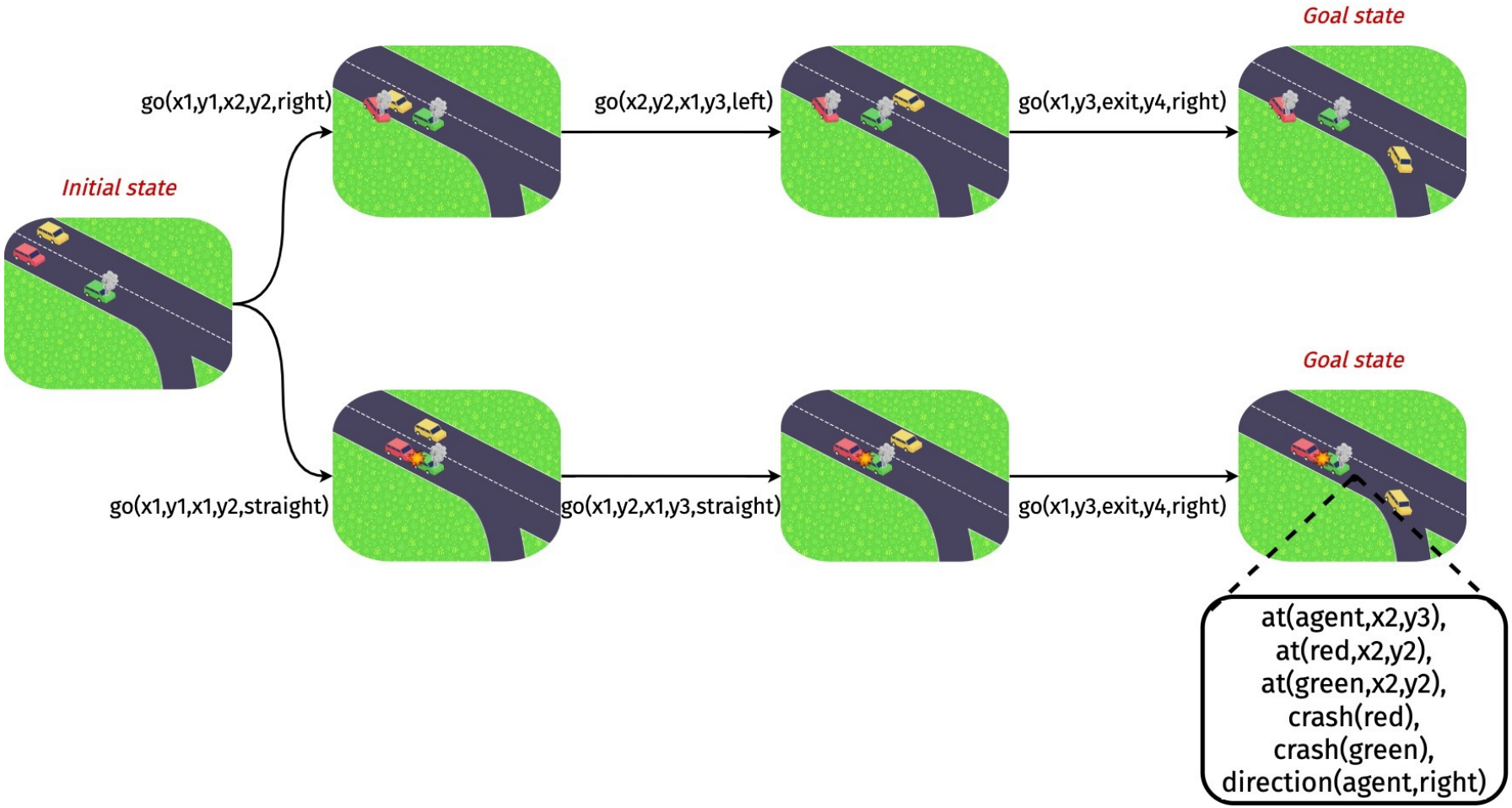
Classical planning, at a glance:



Background notions: classical planning

A **classical planning problem** is a tuple $T = \langle F, O, s_0, g \rangle$, where:

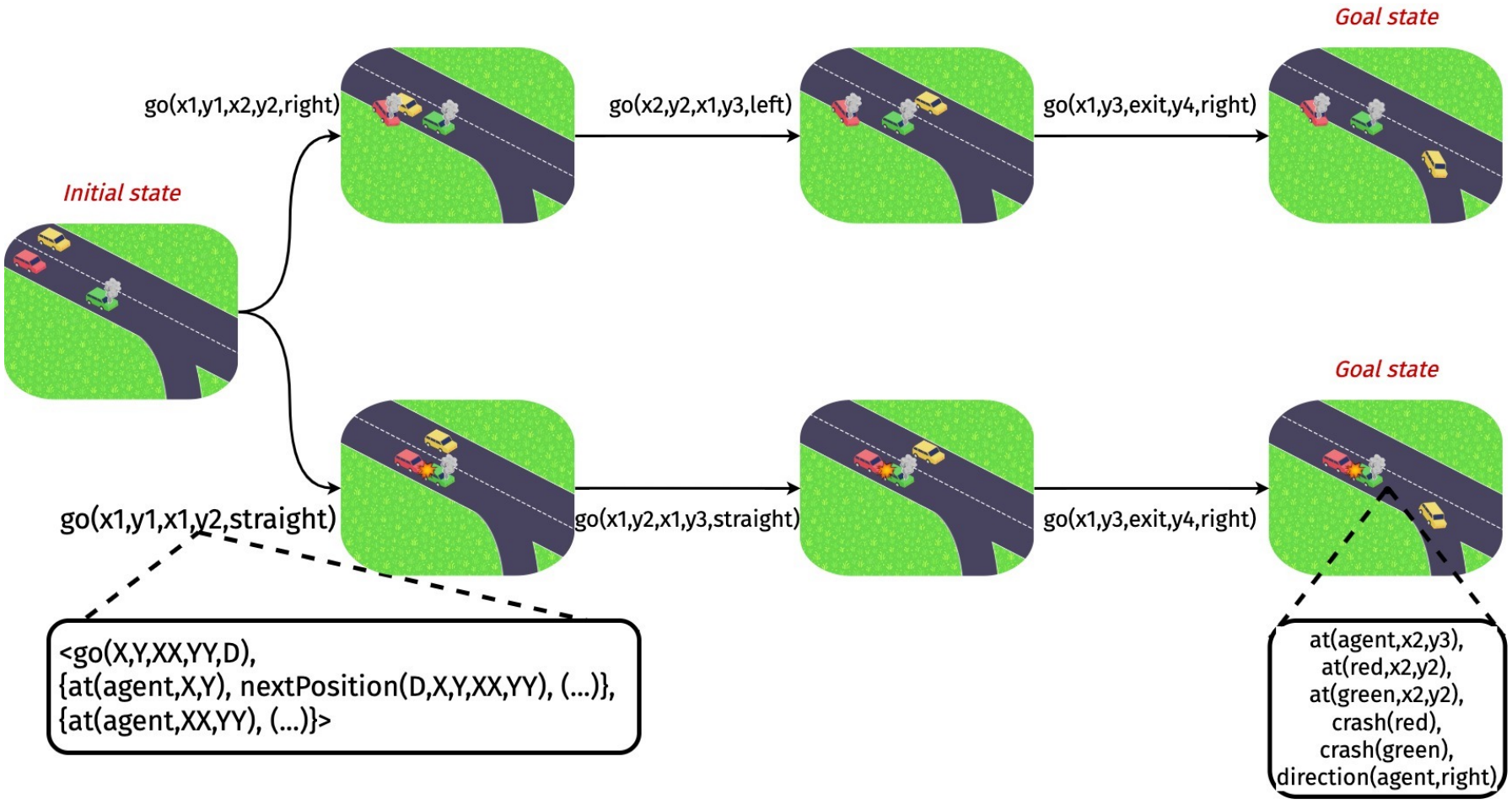
- F is the set of all state properties (atoms over $\mathcal{L}$), called **fluents**, e.g: $at(agent, x_1, y_1), direction(agent, left)$, etc.



Background notions: classical planning

A **classical planning problem** is a tuple $T = \langle F, O, s_0, g \rangle$, where:

- O are the **operators** (over F), and an **action** is a grounded operator.

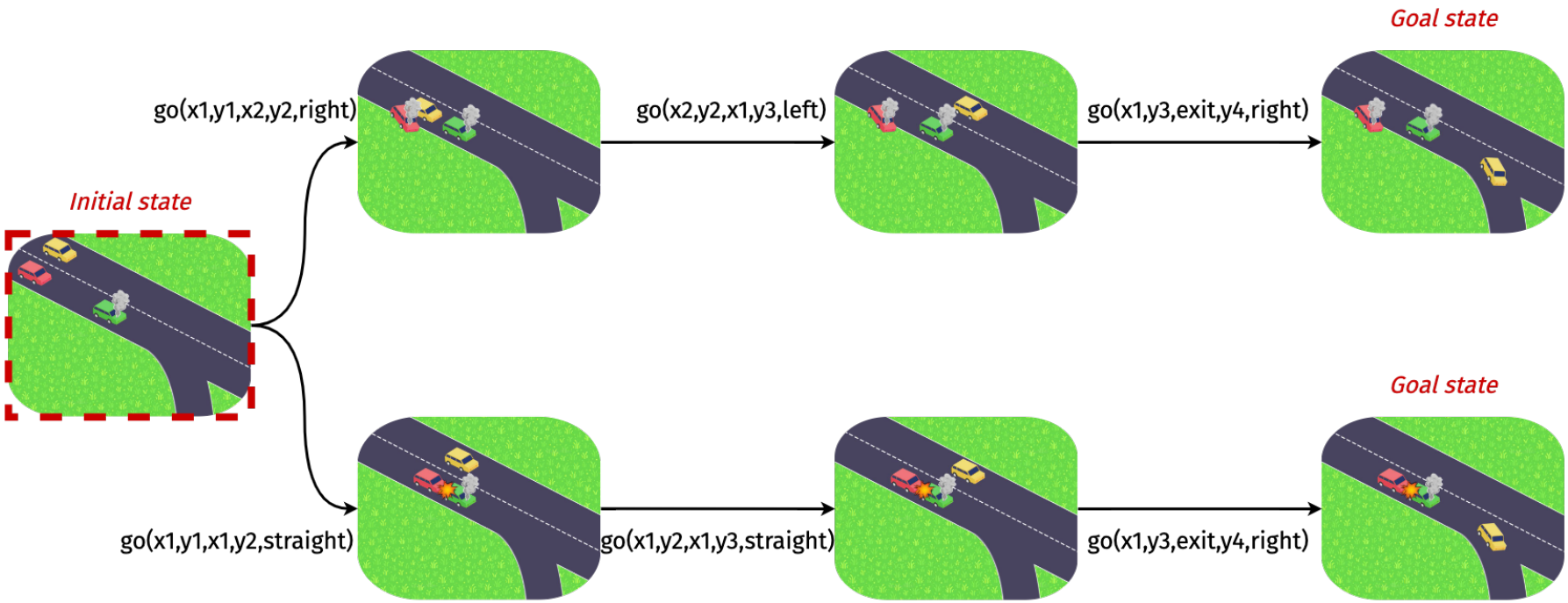


Background notions: classical planning

A **classical planning problem** is a tuple $T = \langle F, O, s_0, g \rangle$, where:

- $s_0 \subseteq F$ is the **initial state**, e.g:

$$s_0 = \{at(agent, x_1, y_1), direction(agent, straight),$$
$$at(red, x_2, y_1), at(green, x_2, y_3), \ldots\}$$

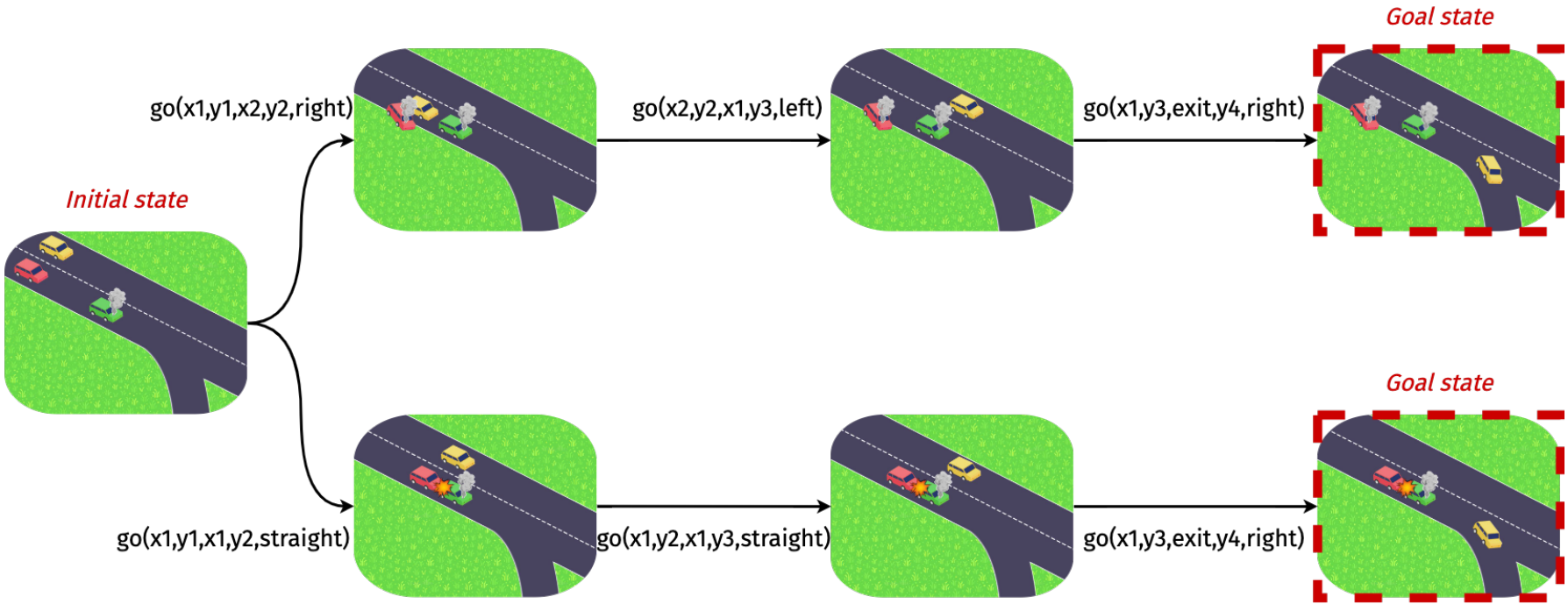


Background notions: classical planning

A **classical planning problem** is a tuple $T = \langle F, O, s_0, g \rangle$, where:

- g is the **goal** (a logic formula over F), denoting success, e.g:

$$g = at(agent, exit, y_4) \wedge direction(agent, right)$$

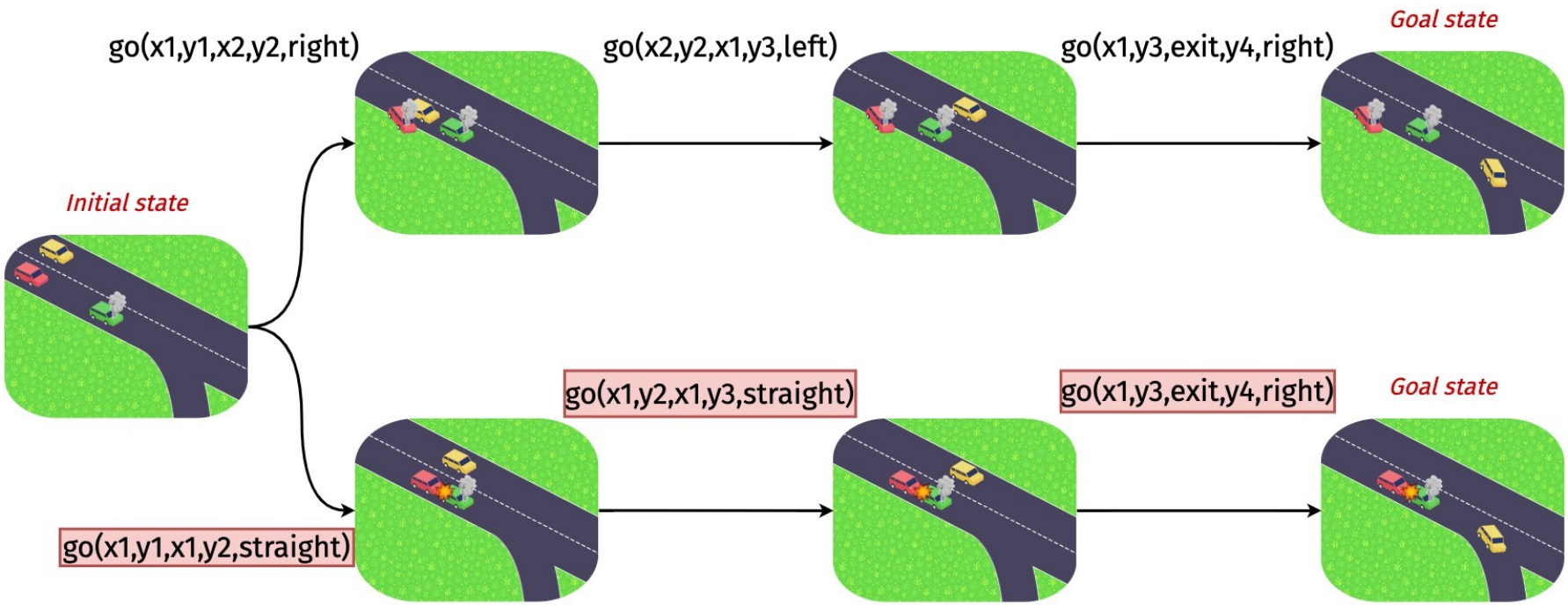


Background notions: classical planning

A **classical planning problem** is a tuple $T = \langle F, O, s_0, g \rangle$, where:

- A **plan** is a sequence of actions $\pi = [a_0, a_1, \ldots, a_n]$ such that: $Succ(a_n, \ldots, Succ(a_1, Succ(a_0, s_0))) \models g$

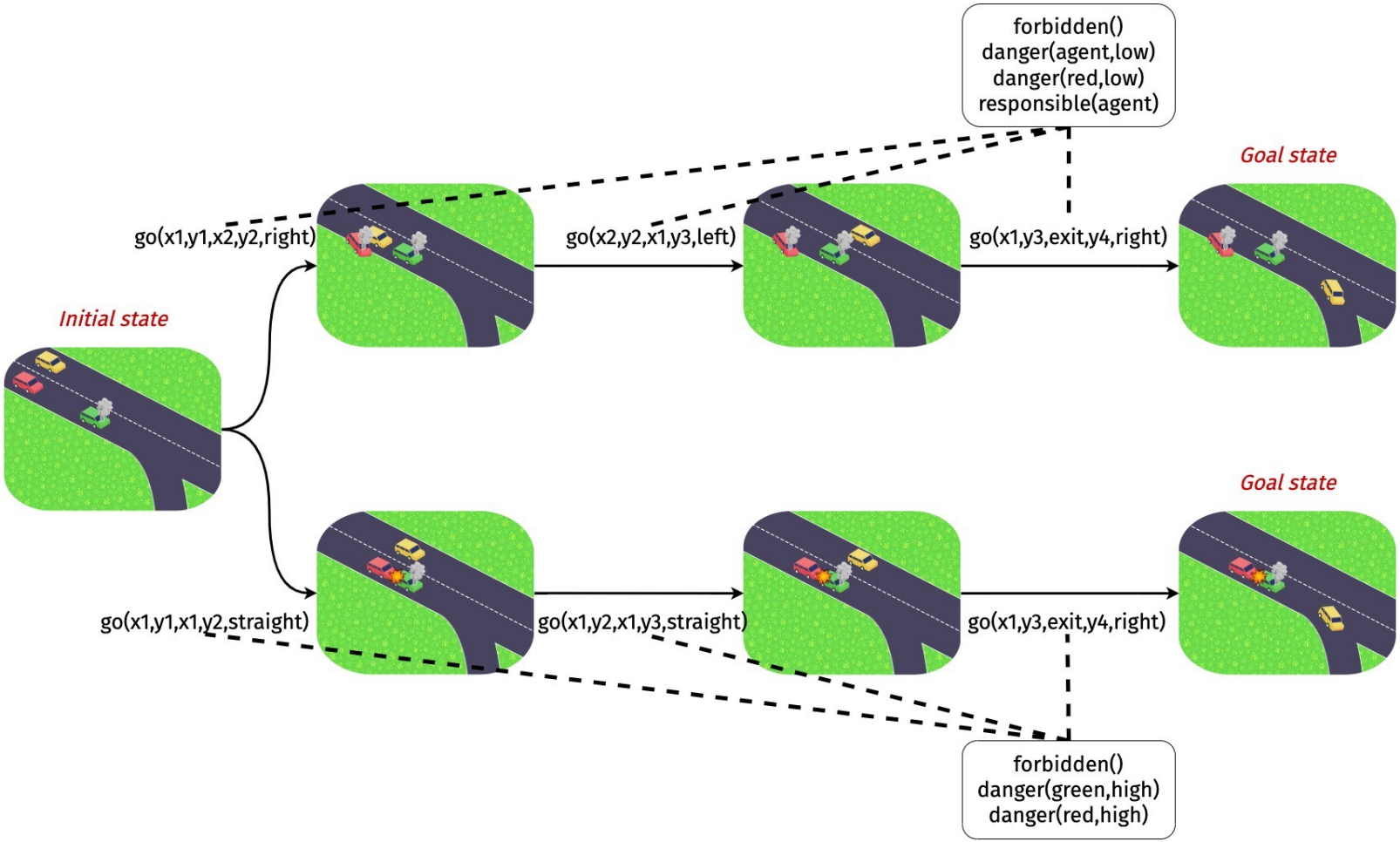
  E.g: $\pi = [go(x_1, y_1, x_1, straight), go(x_1, y_2, x_1, y_3, straight),$
  $go(x_1, y_3, exit, y_4, right)]$



Representing ethics in classical planning

Representing ethics in classical planning

Previous research [Ganascia, 2007, Tolmeijer et al., 2020] has shown how to represent certain ethical theories for decision-making and planning problems.



Question: how can we compare plans ethically?

Representing ethics in classical planning: ethical features

★ **Contribution 1**: an extension [Jedwabny et al., 2021] of classical planning for ethical representation and alignment, we called **ethical planning problem**:

$$T = \langle \overbrace{F, O, s_0, g}^{\text{classical}}, \overbrace{E, R, b}^{\text{ethical}} \rangle$$

- E is the set of **ethical features**,

  E.g: $\mathit{forbidden}()$, $\mathit{responsible}(\mathit{agent})$, $\mathit{danger}(\mathit{agent}, \mathit{low})$, $\mathit{danger}(\mathit{agent}, \mathit{high})$, $\mathit{danger}(\mathit{red}, \mathit{low}), \ldots$

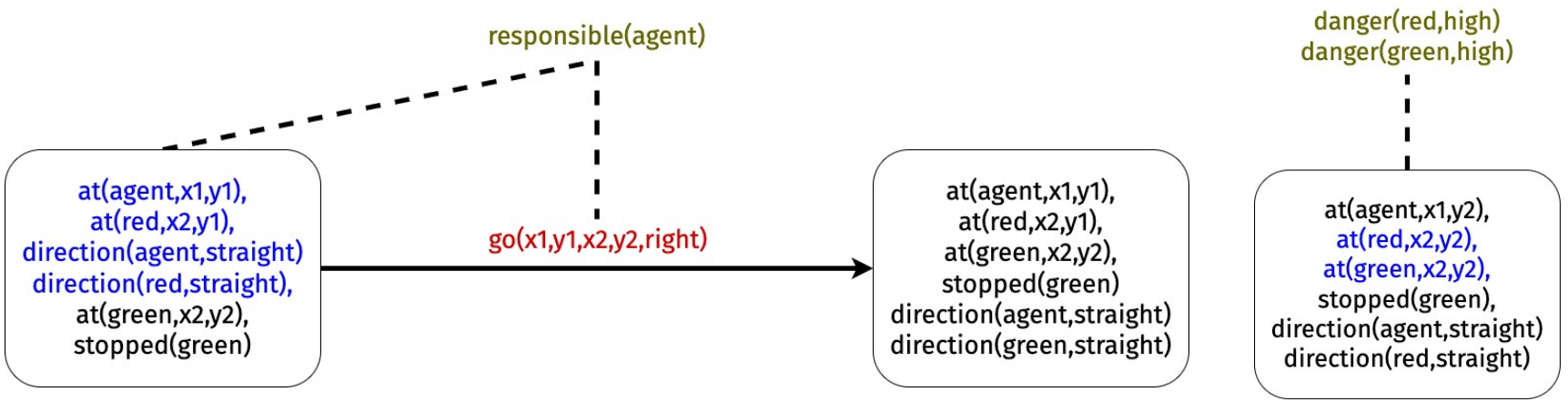
Representing ethics in classical planning: ethical rules

An extension of classical planning for ethical representation and alignment, we called **ethical planning problem**:

$$T = \langle \overbrace{F, O, s_0, g}^{\text{classical}}, \overbrace{E, R, b}^{\text{ethical}} \rangle$$

- R is the set of **ethical rules**, which *assign* ethical features to plans:

$$r = \langle ruleName(X, Y, \ldots), Pre(r), Act(r), E(r) \rangle$$



(a) $Act(r) = go$

(b) $Act(r) = null$

Representing ethics in classical planning: preferences

An extension of classical planning for ethical representation and alignment, we called **ethical planning problem**:

$$T = \langle \overbrace{F, O, s_0, g}^{\text{classical}}, \overbrace{E, R, b}^{\text{ethical}} \rangle$$

- b is the **ethical ranked base**, which compares plans π and π' on ethical terms:

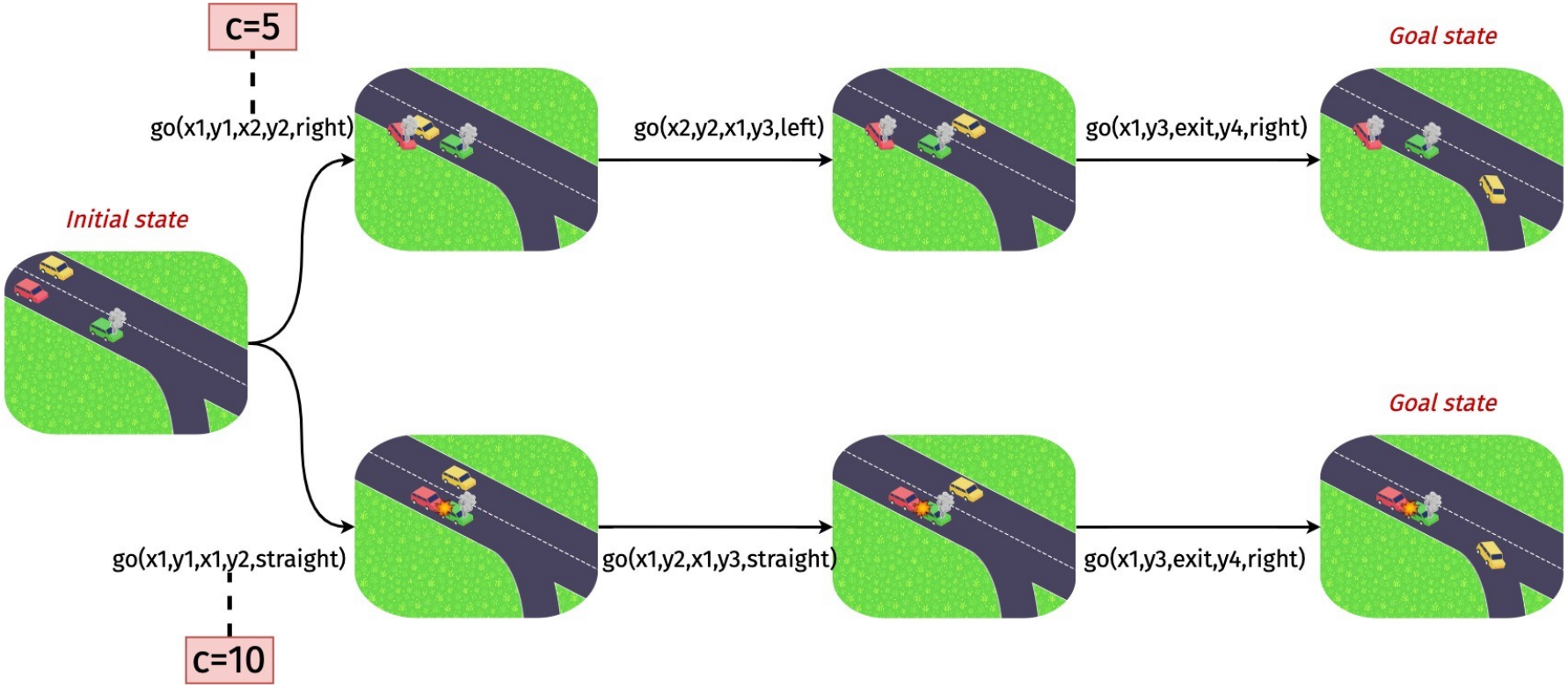
$$\pi \succeq_b \pi' \iff E(\pi) \succeq_b E(\pi')$$

Question: how to compare sets of ethical features?

Representing ethics in classical planning: utilities

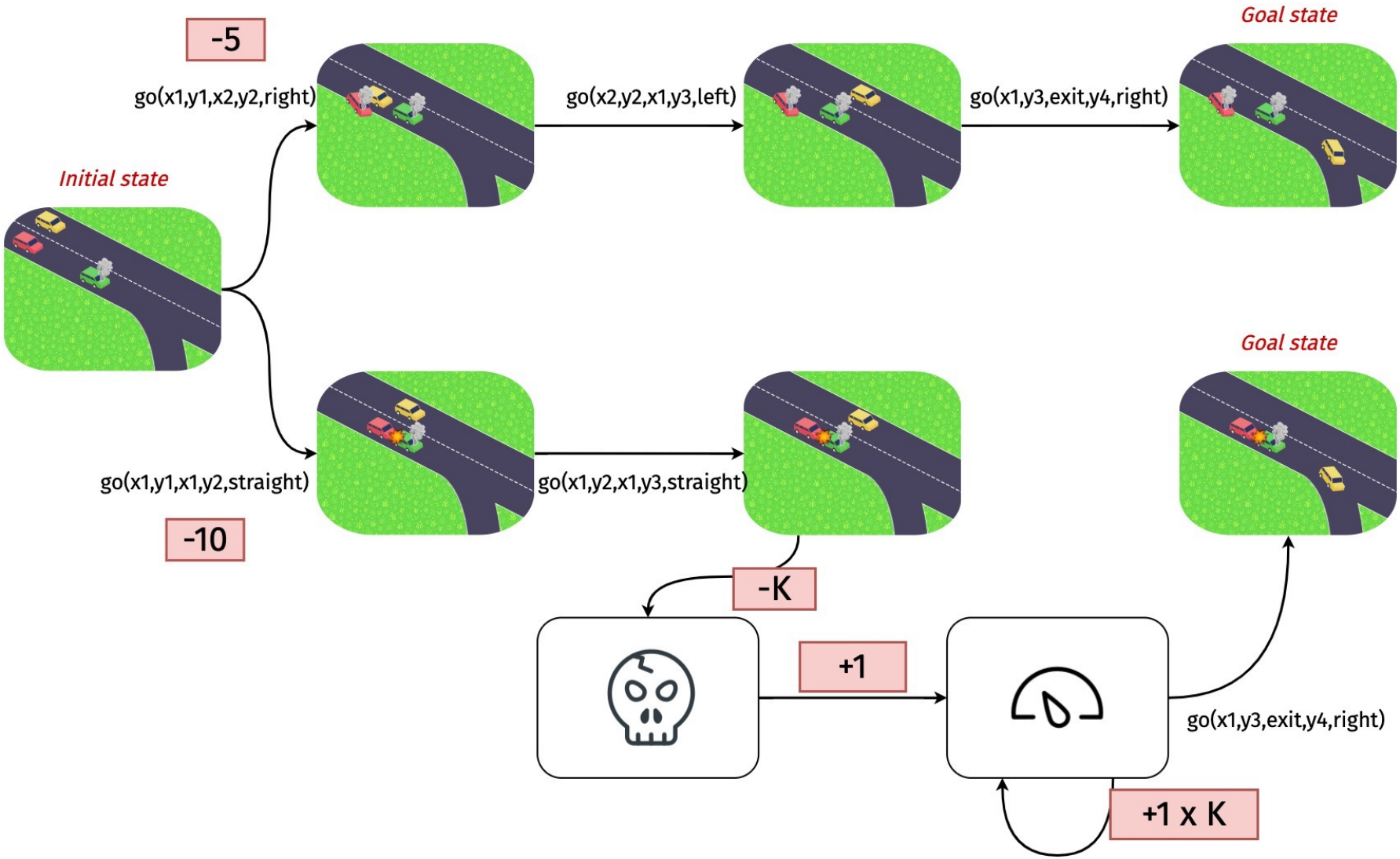
Planning problems can be extended with **action costs** and **utilities** (soft goals) [Gerevini and Long, 2005] as $T = \langle F, O, s_0, g, c, u \rangle$, where:

- $c : A \mapsto \mathbb{R}_{\geq 0}$ is the action cost function,
- $u : S \mapsto \mathbb{R}_{\geq 0}$ is the utility/soft goal function,
- A plan π is **optimal** if $u(\pi) \geq u(\pi')$ for every other plan π'



Representing ethics in classical planning: utilities

Problem: cannot prevent this unless with a **qualitative model**.



Representing ethics in classical planning: ethical ranked bases

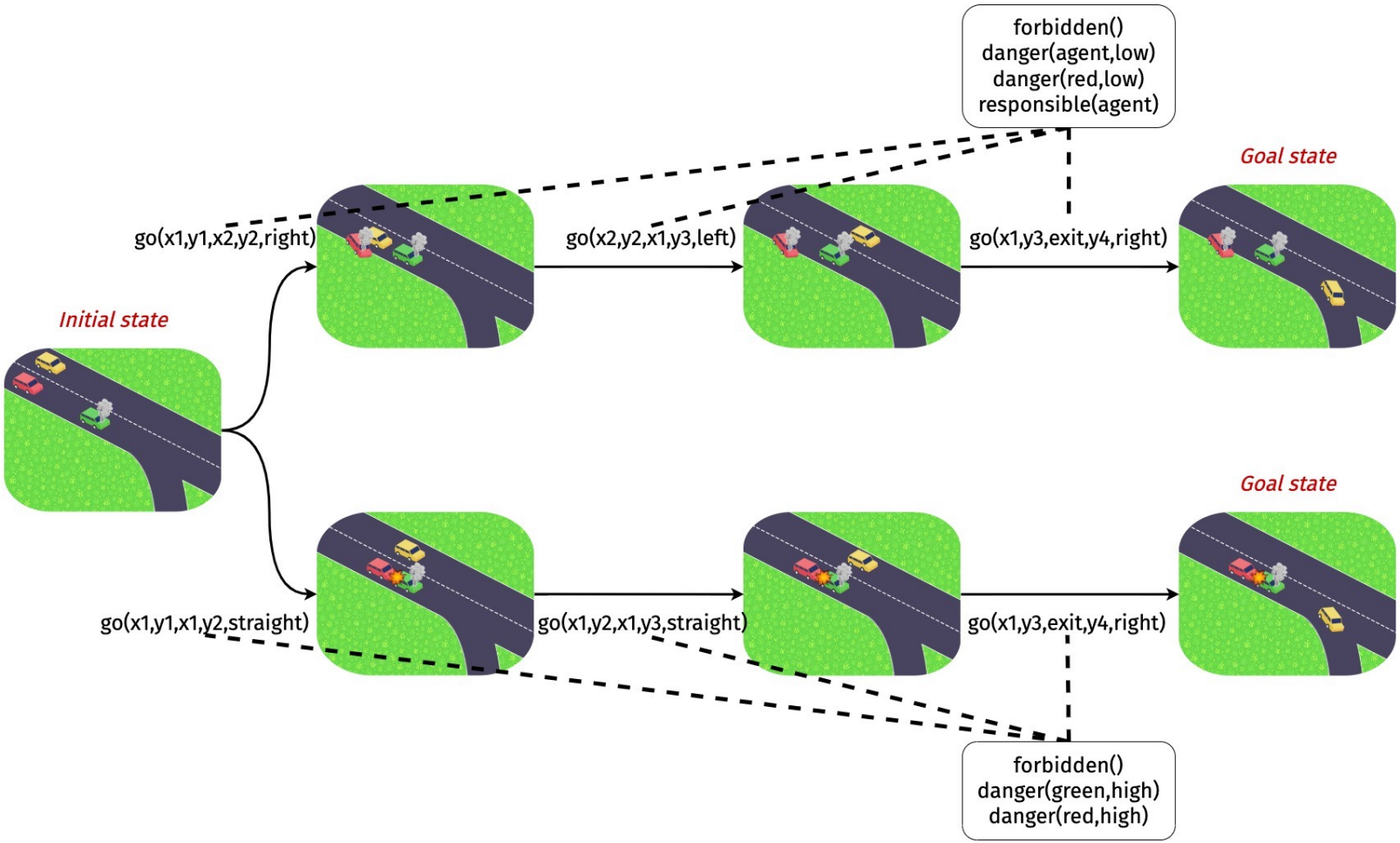
Given an ethical planning problem $T = \langle \overbrace{F, O, s_0, g}^{\text{classical}}, \overbrace{E, R, b}^{\text{ethical}} \rangle$ and $e \in E$:

- $b(e) = \langle Type(e), Rank(e) \rangle$ where:
  - $Type(e) \in \{+, -\}$, and
  - $Rank(e) \in \mathbb{N}_0$.
- For example:
  $b(forbidden()) = \langle -, 4 \rangle$,
  $b(danger(X, high)) = \langle -, 3 \rangle$
  $b(danger(X, low)) = \langle -, 2 \rangle$
  $b(responsible(agent)) = \langle -, 1 \rangle$.

Representing ethics in classical planning: ethical ranked bases

Given an ethical planning problem $T = \langle \overbrace{F, O, s_0, g}^{\text{classical}}, \overbrace{E, R, b}^{\text{ethical}} \rangle$ and $e \in E$:

- b compares plans π and π' on ethical terms: $\pi \succeq_b \pi' \iff E(\pi) \succeq_b E(\pi')$



Representing ethics in classical planning: ethical ranked bases

Given an ethical planning problem $T = \langle \overbrace{F, O, s_0, g}^{\text{classical}}, \overbrace{E, R, b}^{\text{ethical}} \rangle$:

- For example:

$$\pi \succeq_b \pi' \iff$$

$$E(\pi) \succeq_b E(\pi') \iff$$

$$\{forbidden(),danger(agent,low),(...)\} \succeq_b \{forbidden(),danger(green,high),(...)\}$$

Rank 4: $forbidden() \in E(\pi)$ and $forbidden() \in E(\pi')$
$\Rightarrow E(\pi) \succeq_b^4 E(\pi')$

Rank 3: $danger(green, high) \notin E(\pi)$ and $danger(green, high) \in E(\pi')$
$\Rightarrow E(\pi) \succ_b^3 E(\pi')$

Therefore, $E(\pi) \succ_b E(\pi')$

Representing ethics in classical planning: ethical ranked bases

Given an ethical planning problem $T = \langle \overbrace{F, O, s_0, g}^{\text{classical}}, \overbrace{E, R, b}^{\text{ethical}} \rangle$:

- b compares plans π and π' on ethical terms:

$$\pi \succeq_b \pi' \iff \boldsymbol{E(\pi) \succeq_b E(\pi')}$$

- Let $A, B \subseteq E$, then $\boldsymbol{A \succeq_b B}$ if and only if:

$\forall i \in \mathbb{N}$, it holds that $b_i^+(A) = b_i^+(B)$ and $b_i^-(A) = b_i^-(B)$, or

$\exists i \in \mathbb{N}$, such that $\big(b_i^+(B) \subset b_i^+(A) \wedge b_i^-(A) \subseteq b_i^-(B)$, or

$b_i^+(B) \subseteq b_i^+(A) \wedge b_i^-(A) \subset b_i^-(B)\big)$, and

$\forall j > i : b_j^+(A) = b_j^+(B)$ and $b_j^-(A) = b_j^-(B)$.

Where given $i \in \mathbb{N}$ and a set of ethical features $C \subseteq E$:

$$b_i^+(C) = \{e \in C : \ Type(e) = +, \ Rank(e) = i\}$$
$$b_i^-(C) = \{e \in C : \ Type(e) = -, \ Rank(e) = i\}$$

Modeling ethical theories

Additionally, in the thesis:

★ **Contribution 2**: we show how to model adaptations of:

- Consequentialism,
- Deontological permissibility,
- Virtue ethics,
- Prima facie duties,
- Doctrine of double effect, and
- Do-no-harm principle.

Modeling ethical theories: example

Representing ethical theories, example:

Consequentialist ethics:

$$
\begin{aligned}
e_1 &= danger(agent, high) \\
r_1 &= \langle ruleConq_1(), \\
&\quad \{hasCrashed(agent)\}, \\
&\quad null, \\
&\quad \{danger(agent, high)\}\rangle \\
b(e_1) &= \langle -, 3\rangle
\end{aligned}
$$

$$
\begin{aligned}
e_2 &= danger(red, high) \\
r_2 &= \langle ruleConq_2(), \\
&\quad \{hasCrashed(red)\}, \\
&\quad null, \\
&\quad \{danger(red, high)\}\rangle \\
b(e_2) &= \langle -, 3\rangle
\end{aligned}
$$

Deontological ethics:

$$
\begin{aligned}
e_3 &= forbidden() \\
r_3 &= \langle ruleDento(C1, X1, X2, X3, \\
&\quad Y1, D1, D2), \\
&\quad \{at(agent, X1, Y1), \\
&\quad at(C1, X2, Y1), \\
&\quad direction(agent, D1), \\
&\quad direction(C1, D2), \\
&\quad nextX(D1, X1, X3), \\
&\quad nextX(D2, X2, X3), \\
&\quad \neg equal(C1, agent)\}, \\
&\quad go(), \\
&\quad \{forbidden()\}\rangle \\
b(e_3) &= \langle -, 4\rangle
\end{aligned}
$$

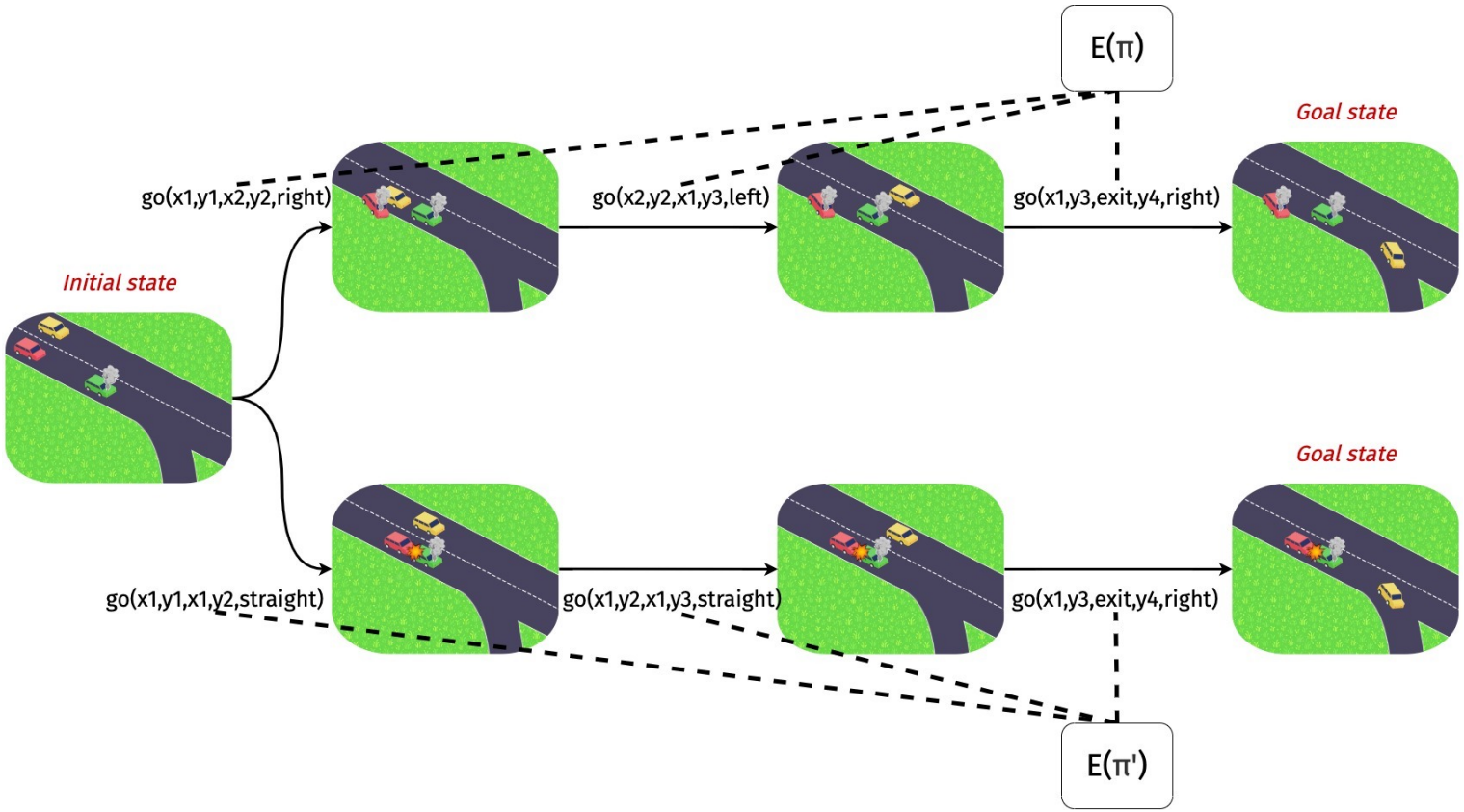
Planning with ethical preferences

Planning with ethical preferences

- We can use ethical features and ranks to align an agent's behavior:

$$T = \langle \overbrace{F, O, s_0, g}^{\text{classical}}, \overbrace{E, R, b}^{\text{ethical}} \rangle$$

- π **optimal** iff $\forall \pi'$ plan, $E(\pi) \succeq_b E(\pi')$



- **Question**: how can we find ethically optimal plans?

Planning complexity and heuristics

Finding plans is highly difficult:

- Determining if a plan exists is PSPACE-complete
  [Bäckström and Nebel, 1995] for propositional classical planning,
- Finding an optimal plan is **very hard in practice** and requires heuristics.
- International Planning Competitions (IPCs):
  - Planning Domain Definition Language (PDDL) [Ghallab et al., 1998].
  - PDDL planners:
    - IPC5 [Gerevini et al., 2009]: **utilities** (soft goals) + **action costs**.
    - More recent IPCs [Torralba and Pommerening, 2018]: only **action costs**.

Planning with preferences: research results

Research results:

PDDL + rank-based preferences

$\Downarrow$ [Feldmann et al., 2006]

PDDL + utilities (soft goals)

$\Downarrow$ [Keyder and Geffner, 2009]

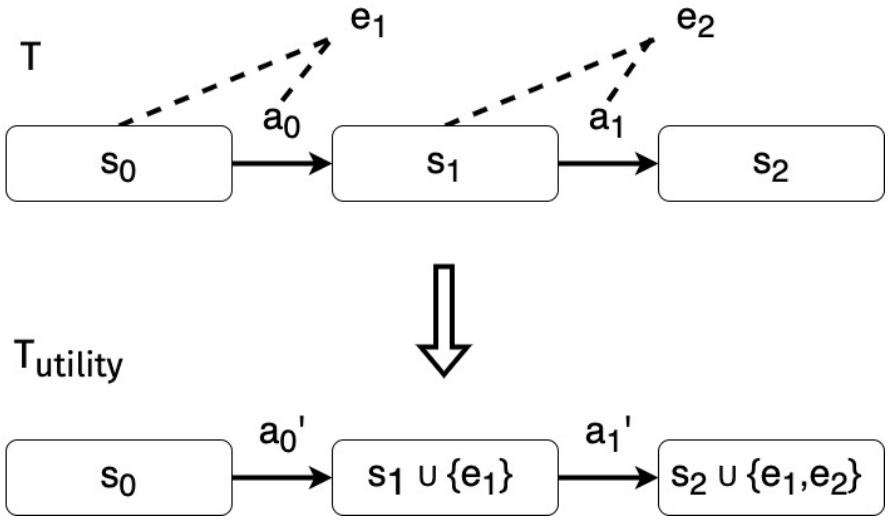
PDDL + action costs

Translation into utilities and action costs

★ **Contribution 3**: demonstrated that ethical planning problems can be translated into utility planning [Jedwabny et al., 2021]:

$$T = \langle F, O, s_0, g, E, R, b \rangle \Rightarrow T_{utility} = \langle F', O', s_0, g', \varnothing, u \rangle$$

Proof intuition: plan **validity** and **optimality**:



- $u(s) = \sum_{i \in \mathbb{N}} |b_i^+(E \cap s) \cup (b_i^-(E) - b_i^-(E \cap s))| \times val_i$
- $u(\pi_1) \geq u(\pi_2) \Rightarrow E(\pi_1) \succeq_b E(\pi_2)$

Implementation of ethical planning problems

★ **Contribution 4**: we implement our model as an extension [Jedwabny et al., 2021] of PDDL, that encodes $T = \langle F, O, s_0, g, E, R, b \rangle$ as:

Math:

$e_1 = danger(agent, high)$

$e_2 = danger(red, high)$

$e_3 = danger(green, high)$

$r_1 = \langle ruleConq(C),$
$\{hasCrashed(C)\},$
$null,$
$\{danger(C, high)\}\rangle$

$b(e_1) = \langle -, 3\rangle$

$b(e_2) = \langle -, 3\rangle$

$b(e_3) = \langle -, 3\rangle$

Code:

```
(:ethical-features
    (danger ?C ?G))
```

```
(:ethical-rule ruleConseq1
    :parameters (?C)
    :precondition (hasCrashed ?C)
    :activation null
    :features (danger ?C high))
```

```
(:ethical-rank
    :feature (danger ?C high)
    :type -
    :rank 3)
```

Translation into utilities and action costs

★ **Contribution 5**: implementation of translation routines:

- Routine 1:

$$\text{PDDL} + \text{ethical}$$

$$\Downarrow \text{(parsing} + \text{translation)}$$

$$\text{PDDL} + \text{utilities (soft goals)}$$

- Routine 2:

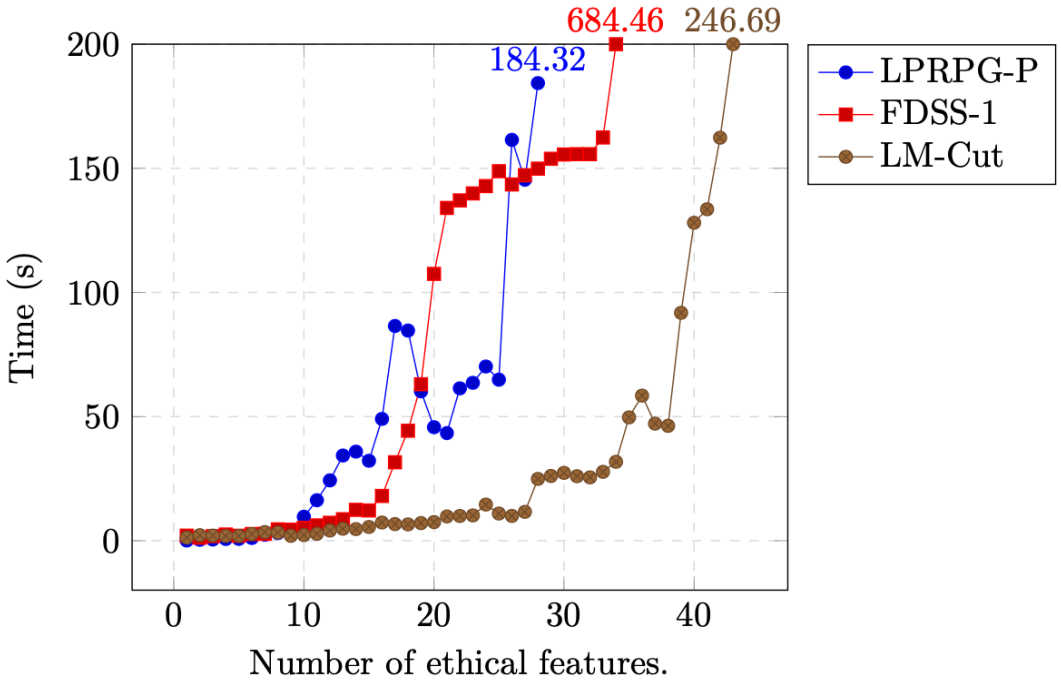
$$\text{PDDL} + \text{ethical}$$

$$\Downarrow \text{(parsing} + \text{translation)}$$

$$\text{PDDL} + \text{action costs}$$

Experimentation: ethical features

Experimentation results (planning time vs. ethical features):

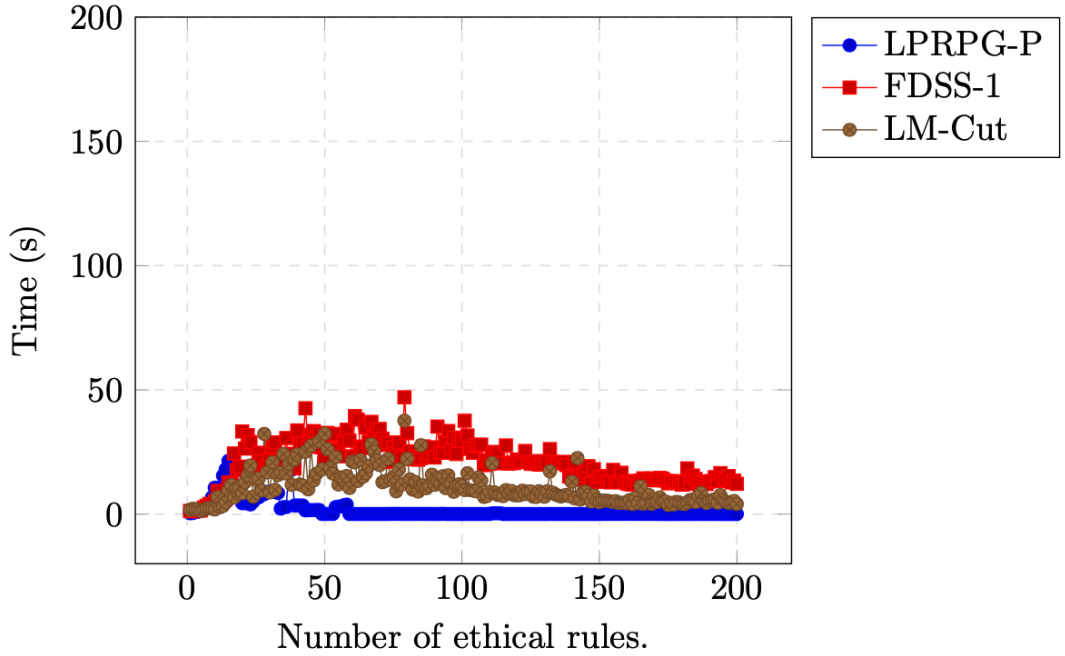


Openstacks planning runtime by ethical features.

- Planning time degradation is above linear w.r.t. ethical features.

Experimentation: ethical rules

Experimentation results (planning time vs. ethical rules):



Openstacks planning runtime parametrized by rules.

- Fixed amount of ethical features ⇒ more ethical rules do not necessarily make planning harder.

Learning ethical preferences

Learning ethical preferences

We have so far shown:

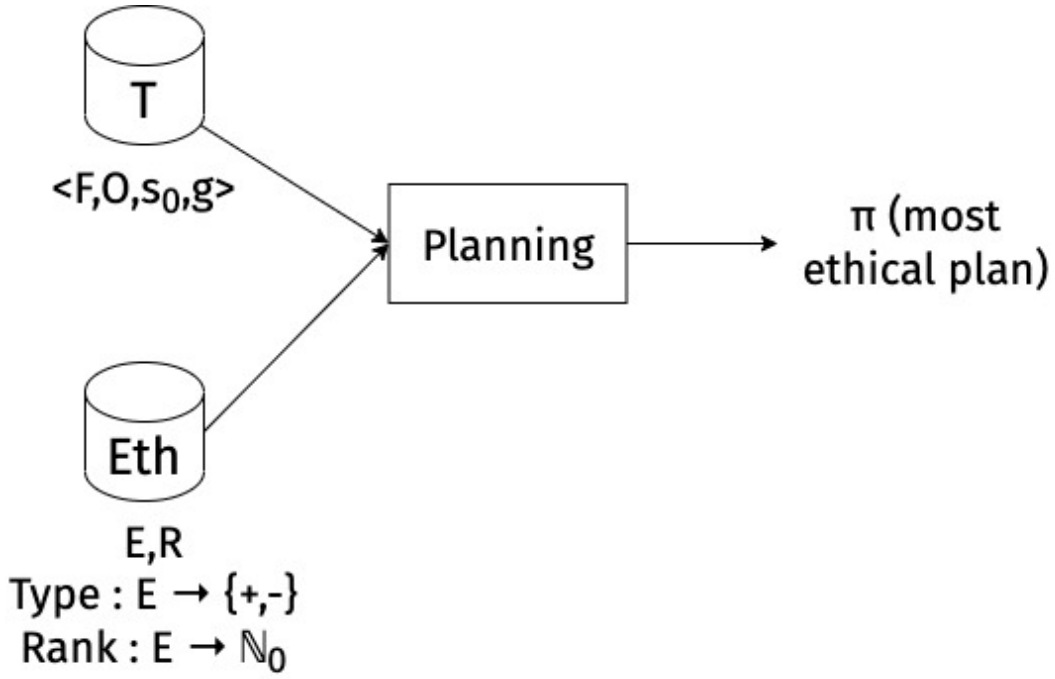
- How to represent ethical features and compare plans on ethical terms to align an agent.
- How to obtain an optimal plan via a translation routine and state-of-the-art planners.

However: how do we align the preferences to societal values?

Learning ethical preferences

So far:

- An ethical planning problem is $T = \langle \overbrace{F, O, s_0, g}^{\text{classical}}, \overbrace{E, R, b}^{\text{ethical}} \rangle$.

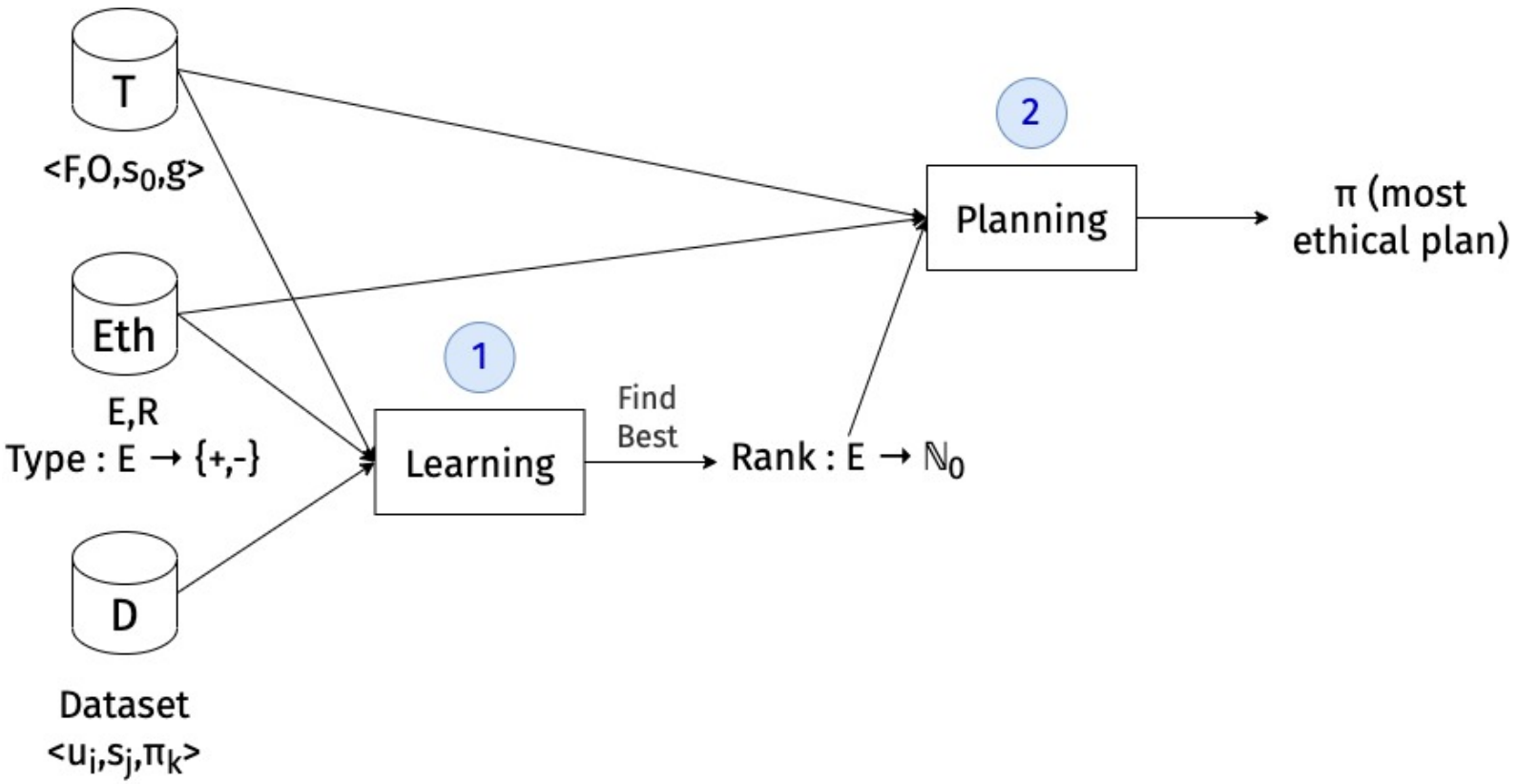


- How do we align the preferences to societal values?
  $\Rightarrow$ **Rank learning**.

Rank learning: overview

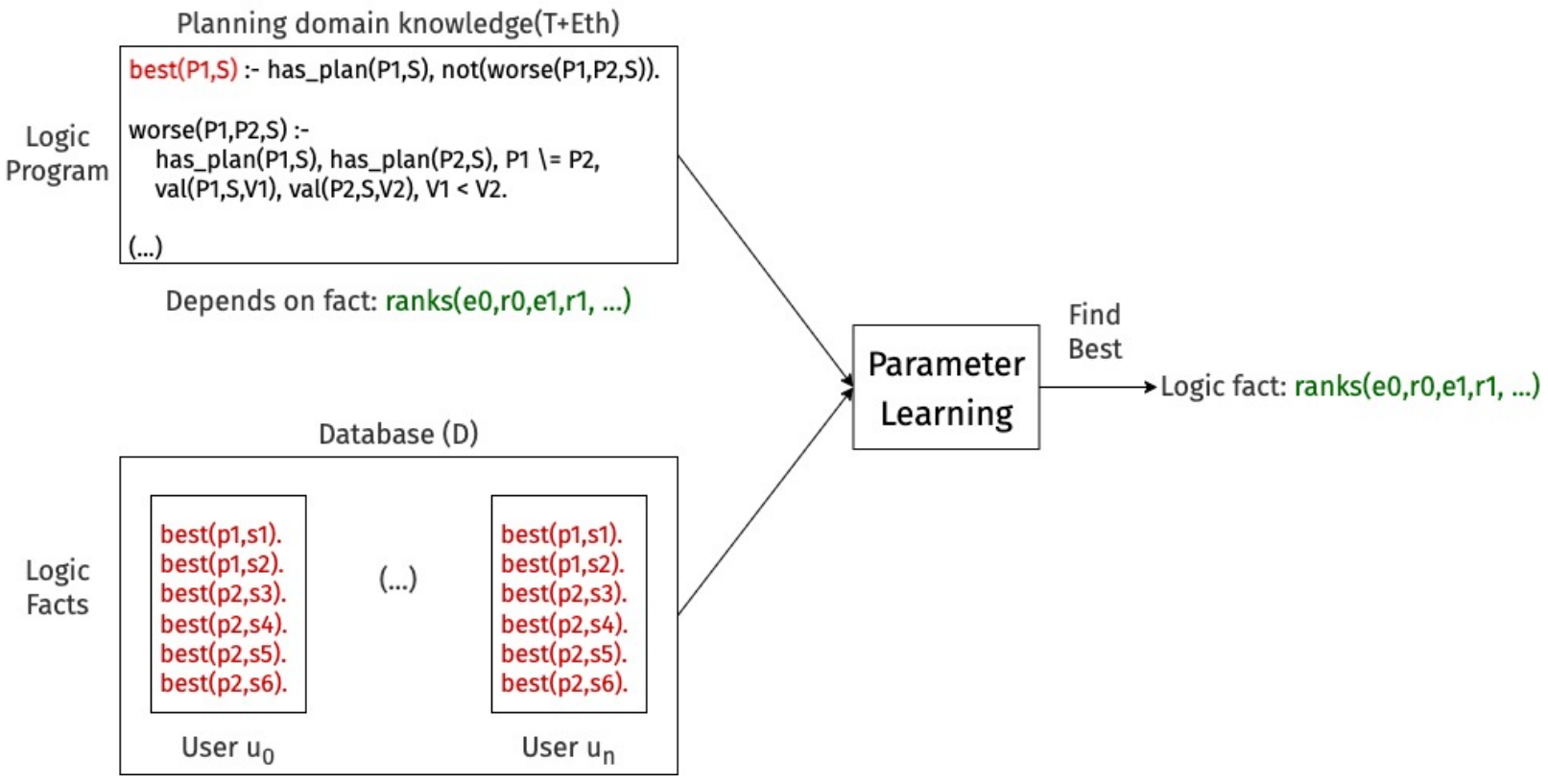
★ **Contribution 6**: an approach to learn ethical ranks from datasets.

- An ethical planning problem is $T = \langle \overbrace{F, O, s_0, g}^{\text{classical}}, \overbrace{E, R, b}^{\text{ethical}} \rangle$.
  Obs: $b(e) = \langle Type(e), Rank(e) \rangle$



Rank learning: implementation

Finding optimal ranks via **parameter learning** [Gutmann et al., 2011] and Problog [De Raedt et al., 2007] (probabilistic logic programming) :



Rank learning: example result

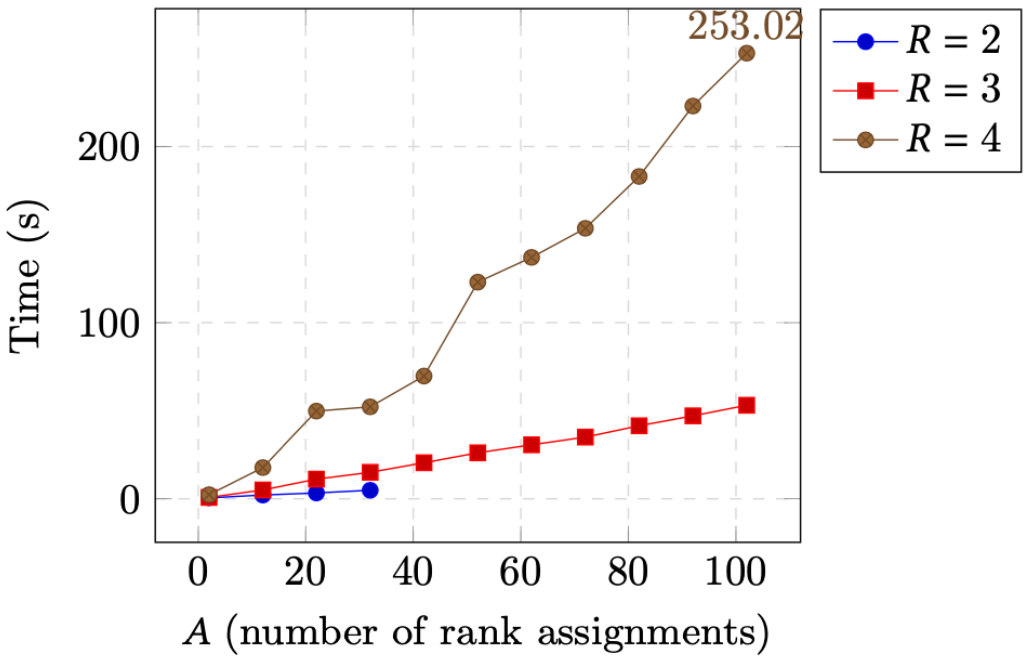
For example (result):

```
0.281427::ranks([danger(agent),1,
                 danger(c1),1,
                 danger(c2),1,
                 responsible(agent),1]).
 ⋮
0.323127::ranks([danger(agent),2,
                 danger(c1),1,
                 danger(c2),1,
                 responsible(agent),1]).
 ⋮
```

- **Observation**: we select the ranks with the maximal probabilistic annotation.

Experimentation: rank assignments

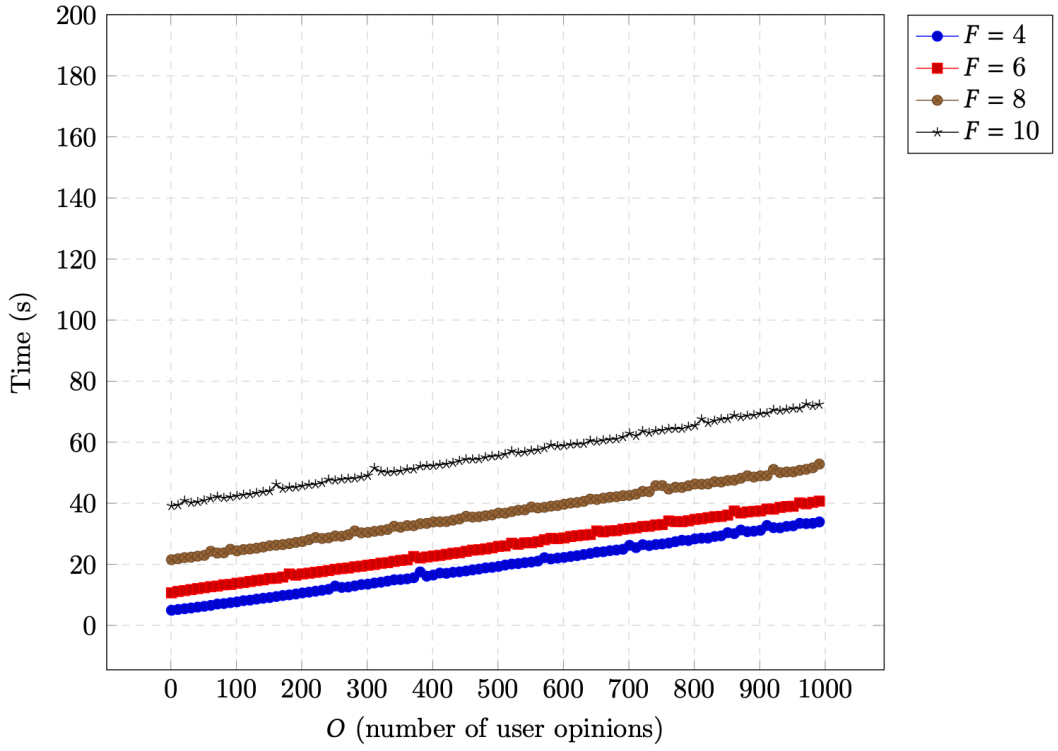
Experimental results (rank assignments):



- A = rank assignments = ethical feature × rank combinations.
- The number of rank assignments impacts the parameter learning time greatly.

Experimentation: dataset size

Experimental results (dataset):



- O = dataset size = number of user/expert opinions.
- The size of the dataset did not seem to affect the performance as much.

Conclusion

Conclusion

Summary:

- **Contribution 1**: we developed an extension of classical planning for ethical representation and alignment.
- **Contribution 2**: we show how to model adaptations of many well-known ethical theories with our framework.
- **Contribution 3**: we implement our model as an extension of PDDL.
- **Contribution 4**: we showed how to translate PDDL+ethical into PDDL with utilities or PDDL with action costs.
- **Contribution 5**: implementation of both translation routines.
- **Contribution 6**: we show how to learn preferences for PDDL+ethical with parameter learning.

Conclusion

List of publications:

- (2020) Explaining ethical planning using ASP - XLoKr (KR workshop).
- (2021) Generating preferred plans with ethical features - FLAIRS conference full paper.
- (2021) Probabilistic rule induction for transparent CBR under uncertainty - SGAI conference full paper.
- (2022) Scrutable Robot Actions Using a Hierarchical Ontological Model - ICCS conference full paper.

Conclusion

Future work:

- Real-life practical validation: test if people agree.
- Explainability testing: whether people can understand the reasoning process of the agent.
- Develop a planner specific to our model of ethics.
- Analyse the complexity of ethical planning.
- Generalize to other planning frameworks (e.g: non-deterministic, probabilistic).
- Improve implementations to handle planning and learning with more ethical features.

Thank you

Thank you for listening!

References i


Anderson, M. and Anderson, S. L. (2007).
The status of machine ethics: a report from the aaai symposium.
Minds and Machines, 17(1):1–10.


Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020).
Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai.
Information Fusion, 58:82–115.


Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., and Rahwan, I. (2018).
The moral machine experiment.
Nature, 563(7729):59–64.

References ii


Bäckström, C. and Nebel, B. (1995).
Complexity results for sas+ planning.
Computational Intelligence, 11(4):625–655.


De Raedt, L., Kimmig, A., and Toivonen, H. (2007).
Problog: A probabilistic prolog and its application in link discovery.

In *IJCAI*, volume 7, pages 2462–2467. Hyderabad.


Feldmann, R., Brewka, G., and Wenzel, S. (2006).
Planning with prioritized goals.
In *KR*, pages 503–514.


Ganascia, J.-G. (2007).
Ethical system formalization using non-monotonic logics.
In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29.

References iii


Gerevini, A. and Long, D. (2005).
Plan constraints and preferences in pddl3.
Technical report, Technical Report 2005-08-07, Department of Electronics for Automation ....


Gerevini, A. E., Haslum, P., Long, D., Saetti, A., and Dimopoulos, Y. (2009).
Deterministic planning in the fifth international planning competition: Pddl3 and experimental evaluation of the planners.
AIJ, 173(5-6):619–668.


Ghallab, M., Knoblock, C., Wilkins, D., Barrett, A., Christianson, D., Friedman, M., Kwok, C., Golden, K., Penberthy, S., Smith, D., Sun, Y., and Weld, D. (1998).
Pddl - the planning domain definition language.

References iv

Gutmann, B., Thon, I., and Raedt, L. D. (2011).
Learning the parameters of probabilistic logic programs from interpretations.
In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 581–596. Springer.

Horizon Magazine Website, F. B. (2021).
Do you trust automated cars? if not, you're not alone.
`https://ec.europa.eu/research-and-innovation/en/horizon-magazine/do-you-trust-automated-cars-if-not-youre-not-alone`.
[Online; accessed 14-November-2022].

References v


Jedwabny, M., Bisquert, P., and Croitoru, M. (2021).
Generating preferred plans with ethical features.
In Bell, E. and Keshtkar, F., editors, *Proceedings of the Thirty-Fourth International Florida Artificial Intelligence Research Society Conference, North Miami Beach, Florida, USA, May 17-19, 2021*.


Keyder, E. and Geffner, H. (2009).
Soft goals can be compiled away.
Journal of Artificial Intelligence Research, 36:547–556.


Moor, J. H. (2006).
The nature, importance, and difficulty of machine ethics.
IEEE intelligent systems, 21(4):18–21.

References vi


NPR Website, A. P. (2022).
Nearly 400 car crashes in 11 months involved automated tech, companies tell regulators.
https://www.npr.org/2022/06/15/1105252793/nearly-400-car-crashes-in-11-months-involved-automated-tech-companies-tell-regul.
[Online; accessed 14-November-2022].


Shahriari, K. and Shahriari, M. (2017).
Ieee standard review — ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems.
In *IEEE*, pages 197–201.

References vii


Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., and Bernstein, A. (2020).
Implementations in machine ethics: a survey.
ACM Computing Surveys (CSUR), 53(6):1–38.


Torralba, A. and Pommerening, F. (2018).
IPC 2018 Website.
`https://ipc2018-classical.bitbucket.io/`.