Model-Based Reasoning for Explainable AI Planning as a Service

Daniele Magazzeni

KING'S College LONDON

AI Planning

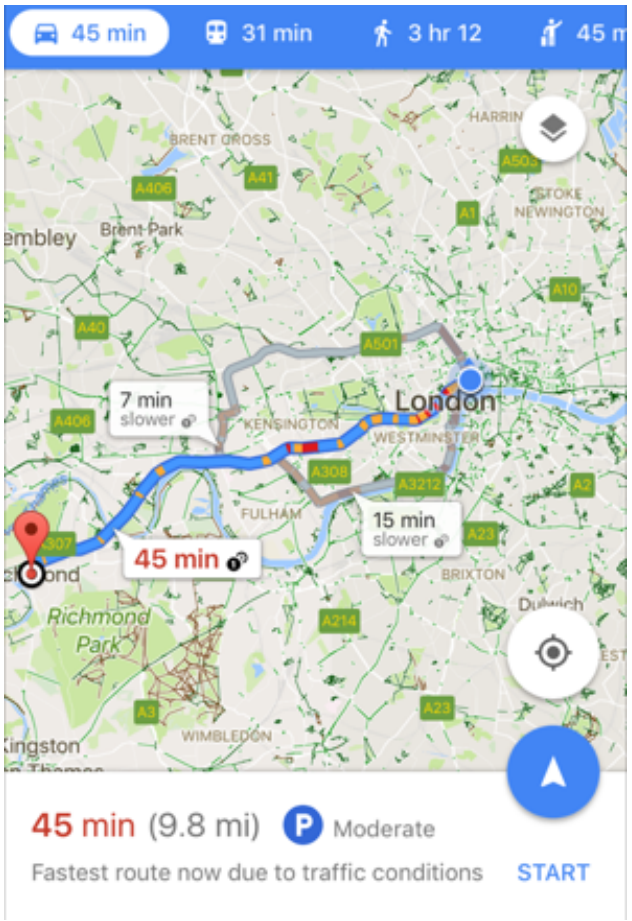
- Planning is about **determining actions *before* doing them**, anticipating the things that will need to be done and preparing for them

- Planning is usually done by (teams of) humans: AI planning is for when this job needs to be done fast, frequently, or is too complicated for humans

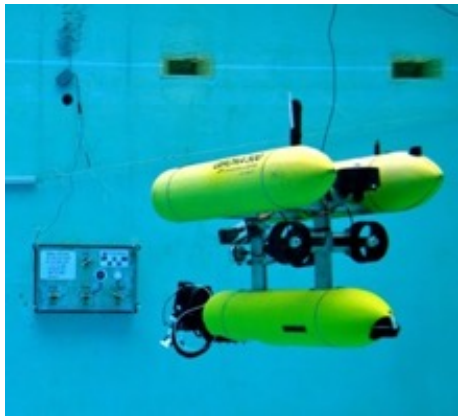
- Where there are resources to be managed, productivity to increase, pollution to reduce, or when a strategy needs to be designed quickly, planning can do it.

- AI Planning is not meant to replace humans, but instead to assist humans in their decision making: **Human-AI Teaming**.


Model
Data
PLANNER
Plan


45 min
31 min
3 hr 12
BRENT CROSS
Brent Park
STOKE NEWINGTON
A406
A41
A1
A10
A40
A501
7 min
slower
London
KENSINGTON
WESTMINSTER
A308
A3212
A2
FULHAM
15 min
slower
A23
45 min
BRIXTON
Richmond Park
A214
WIMBLEDON
A3
45 min (9.8 mi)
Moderate
Fastest route now due to traffic conditions
START


Model
Data
PLANNER
Plan



PDDL: Planning Domain Definition Language

```
(:durative-action do_hover
 :parameters (?v - vehicle ?from ?to - waypoint)
 :duration ( = ?duration (* (distance ?from ?to)
                            (invtime ?v)))
 :condition (and (at start (at ?v ?from))
                 (at start (connected ?from ?to)))
 :effect (and (at start (not (at ?v ?from)))
              (at end (at ?v ?to))))

(:durative-action observe
 :parameters (?v - vehicle ?wp - waypoint
              ?ip - inspectionpoint)
 :duration ( = ?duration (obstime))
 :condition (and (at start (at ?v ?wp))
                 (at start (cansee ?v ?ip ?wp)))
 :effect (and (at start (not (cansee ?v ?ip ?wp)))
              (at end (increase (observed ?ip)
                       (obs ?ip ?wp)))))
```

Temporal planning with time windows

```
(:durative-action do_hover_controlled ...)

(:durative-action do_hover_fast ...)

(:durative-action correct_position ...)

(:durative-action observe_inspection_point ...)

(:durative-action illuminate_pillar ...)

(:durative-action observe_pillar ...)

(:durative-action examine_panel ...)

(:durative-action turn_valve ...)

(:durative-action recalibrate_arm ...)
```

```
;; time window 2 [400--800]
(at 400 (= (valve_goal v2 270))
(at 400 (not (valve_blocked v2)))
(at 400 (valve_free v2))
(at 400 (not (valve_goal_unchecked v2)))

(at 800 (valve_blocked v2))
(at 800 (not (valve_free v2)))

(at 400 (= (valve_goal v3) 10))
(at 400 (not (valve_blocked v3)))
(at 400 (valve_free v3))
(at 400 (not (valve_goal_unchecked v3)))

(at 800 (valve_blocked v3))
(at 800 (not (valve_free v3)))
```

```
0.000: (correct_position auv wp0)  [10.000]
10.001: (do_hover_controlled auv wp0 wp_strat_p0)  [33.532]
43.534: (turn_valve auv wp_strat_p0 p0 v0)  [120.000]
163.535: (correct_position auv wp_strat_p0)  [10.000]
173.536: (turn_valve auv wp_strat_p0 p0 v1)  [120.000]
293.537: (correct_position auv wp_strat_p0)  [10.000]
293.537: (recalibrate_arm auv wp0)  [180.000]
473.538: (turn_valve auv wp_strat_p0 p0 v2)  [120.000]
593.539: (correct_position auv wp_strat_p0)  [10.000]
603.540: (turn_valve auv wp_strat_p0 p0 v3)  [120.000]
```

KCL Planners

Linear dynamics: **POPF/Optic/Colin**

-Forward heuristic search

-Use Linear Programming and Simple Temporal Networks to check temporal constraints

Polynomial Non-Linear dynamics: **SMTPlan**

-Encode the planning problem as SMT formula

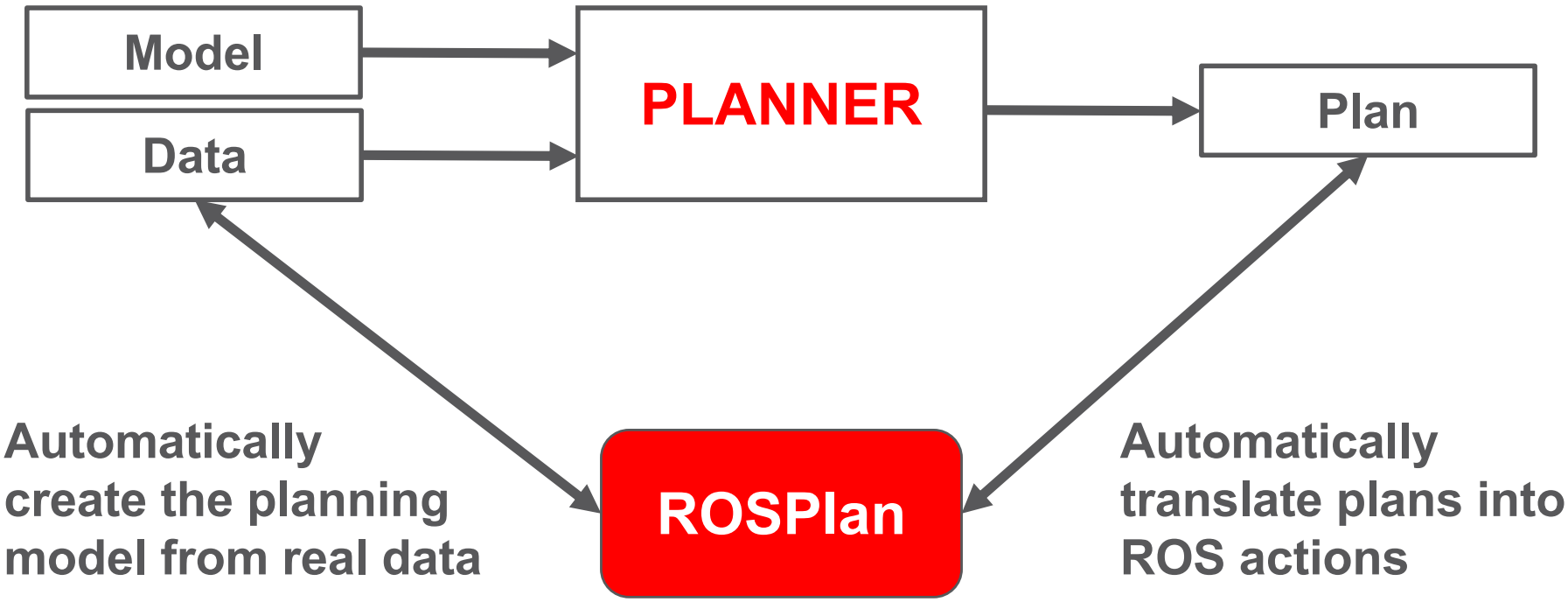
-Use Computer Algebra System to compute indefinite integrals

Non-Linear dynamics: **UPMurphi/DiNO**

-Forward heuristic search

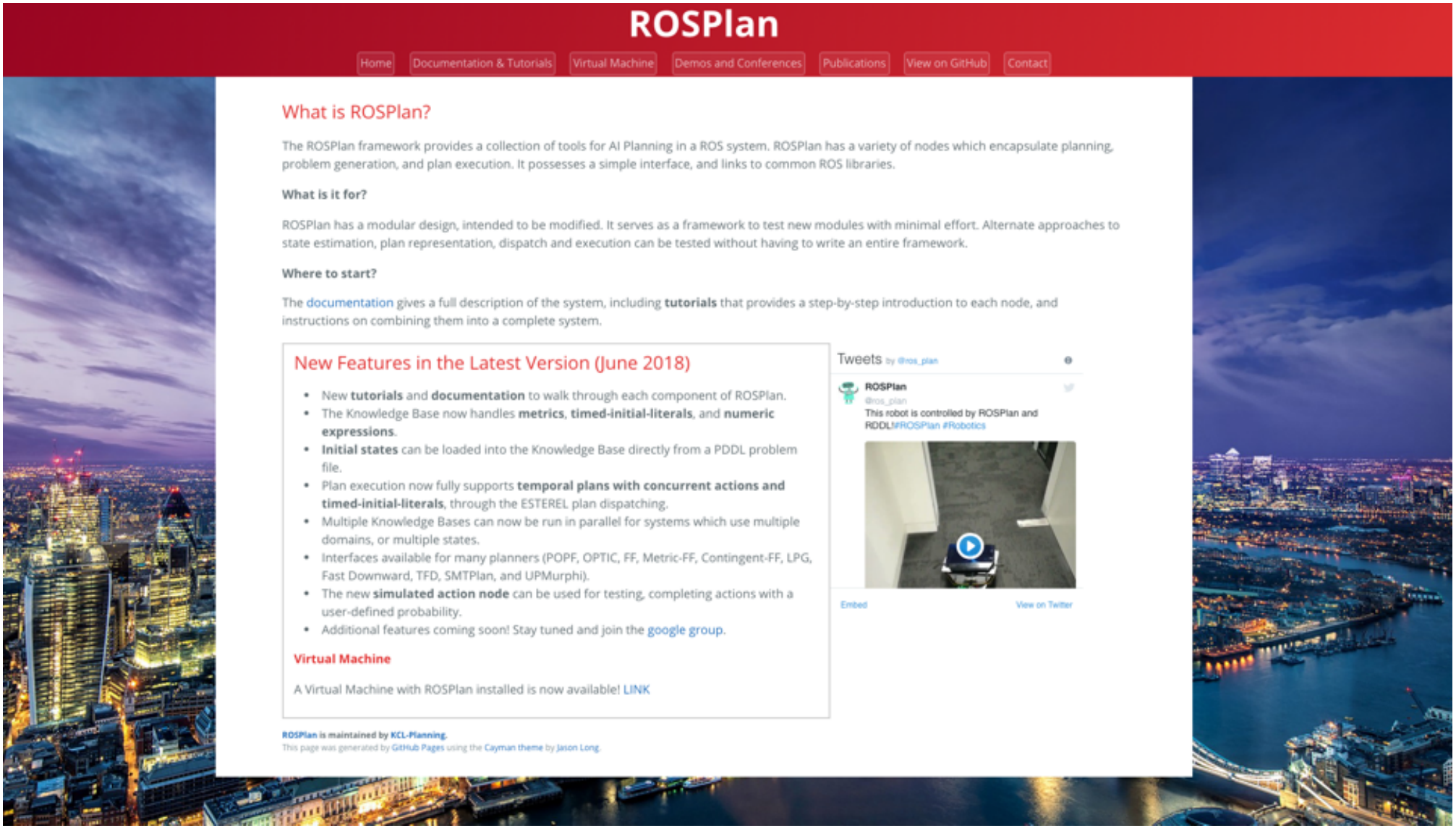
-Use discretisation to handle complex dynamics

All planners are open source



Plan execution
Replanning
Plan failures
Model changes (e.g. equipment failures)
Probabilistic Planning

KING'S College LONDON



ROSPlan is open source: http://kcl-planning.github.io/ROSPlan/

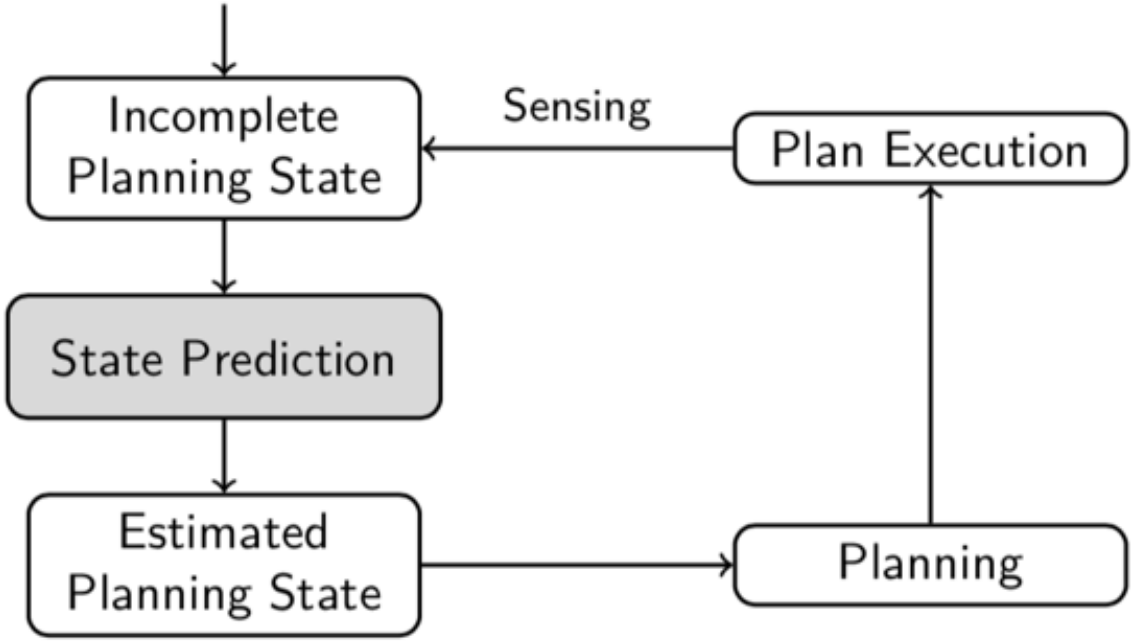
Virtual Machine: kcl-planning.github.io/ROSPlan/vm

Documentation and Tutorials: kcl-planning.github.io/ROSPlan/

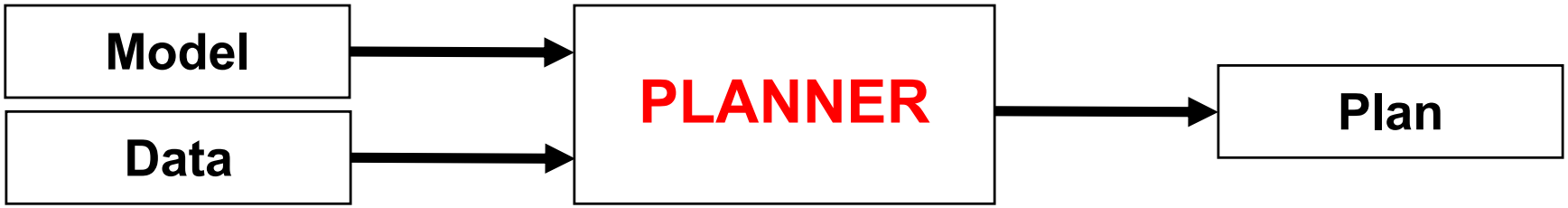
Machine Learning

ROSPlan

Decreasing State Uncertainty



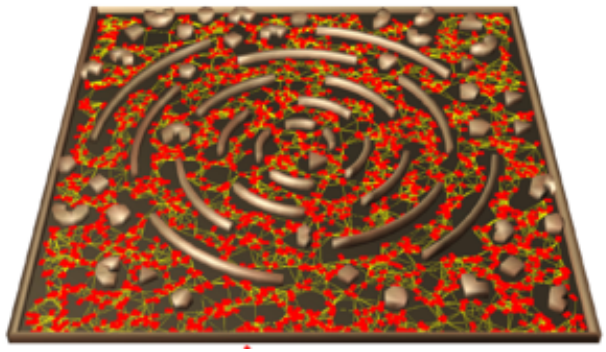
Krivic, Cashmore, Magazzeni, Ridder, Szedmak, Piater. **Decreasing Uncertainty in Planning with State Prediction.** IJCAI 2017.

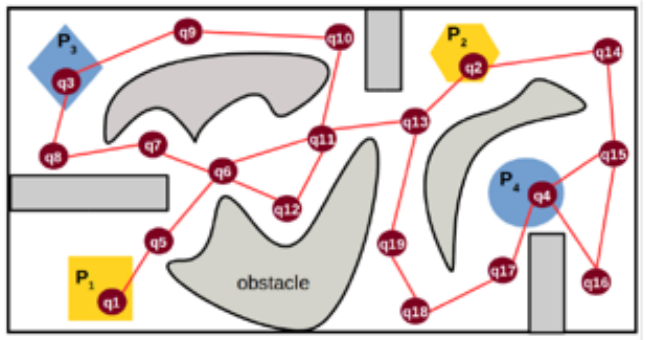


Machine Learning

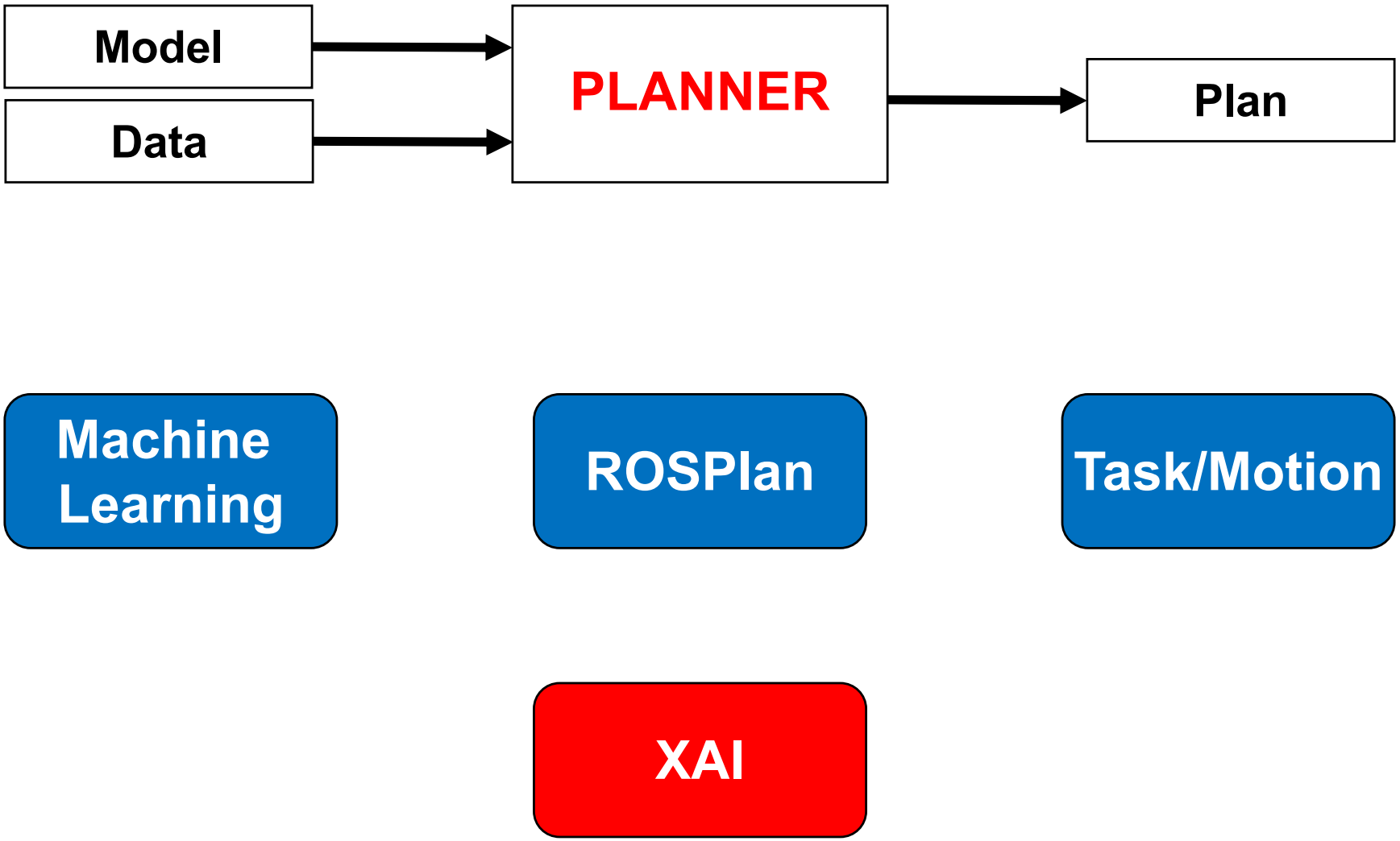
ROSPlan

Task/Motion





Edelkamp, Lahijanian, Magazzeni, Plaku. **Integrating Temporal Reasoning and Sampling-Based Motion Planning for Multi-Goal Problems with Dynamics and Time Windows.** IROS 2018.


Model
Data
PLANNER
Plan
Machine Learning
ROSPlan
Task/Motion
XAI

XAI-Related Meetings

- Explanation-aware Computing (ExaCt) Workshop Series (2005-2012; ECAI(2), IJCAI(2), AAAI, AAAI FSS)
- Human Interpretability in Machine Learning (23 June 2016; ICML-16 WS; New York, NY)
- NIPS-16 Workshops; Barcelona, Spain
  - Future of Interactive Machine Learning (9 December 2016)
  - Interpretable ML for Complex Systems (9 December 2016)
    - Workshop Proceedings
- Interactive Learning (13-17 Feb 2017; Simons Institute; UC Berkeley)
- Designing the User Experience of Machine Learning Systems (27-29 March 2017; AAAI Spring Symposium; Palo Alto, CA)
- Explainability of Learning Machines (17 May 2017; IJCNN-17 Special session; Anchorage, Alaska)
- Explainable Computer Vision Multimedia and Job Candidate Screening Coopetition (26 July 2017; CVPR-17 WS; Honolulu, Hawaii)
- Human in the Loop Machine Learning (11 August 2017; ICML-17 Conference; Sydney, Australia)
- Explainable AI (20 August 2017; IJCAI-17 WS; Melbourne, Australia)
- Explainable Computational Intelligence (XCI) (4 September 2017; INLG-17 WS; Santiago de Compostela, Spain)
- NIPS-17 Events; Long Beach, CA
  - Interpretable Machine Learning (7 December 2017; Symposium)
    - This URL doesn't work (as of 7 February 2018). A brief description of this event can be found here.
  - Transparent and Interpretable Machine Learning in Safety Critical Environments (8 December 2017; Workshop)
  - Interpreting, Explaining and Visualizing Deep Learning...now what? (9 December 2017; Workshop)
- Explainable Robotic Systems (5 March 2018; HRI-18 WS; Chicago, IL)
- Explainable Smart Systems (11 March 2018; IUI-18 Conference; Tokyo, Japan)
- General Data protection regulation: An Opportunity for the HCI community? (21 or 22 April 2018; CHI-18 Conference; Montreal Canada)
- Advances in XAI (Within 11-15 June 2018; IPMU-18 Special Session; Cadiz, Spain)
- Explainable AI Planning (25 or 26 June 2018; ICAPS-18 WS; Delft, The Netherlands)
- FAIM-18 (i.e., Federated AI Meeting of IJCAI/ECAI/ICML/AAMAS/ICCBR/SOCS)
  - ICCBR-18 Workshop on Explainable Case-Based Reasoning (XCBR)
  - Fairness, Interpretability, and Explainability Federation of Workshops
    - Explainable Artificial Intelligence
    - Fairness, Accountability, and Transparency in Machine Learning
    - Human Interpretability in Machine Learning
    - Interpretable & Reasonable Deep Learning and its Applications
- Workshop on XAI (Sometime within 27-30 August 2018; CD-MAKE-18; Hamburg, Germany)
- Luxembourg Logic for AI Summit (17-19 September 2018; Luxembourg)
  - A core objective of LuxLogAI is to present the latest developments and progress made on the crucial question of how to make AI more

Special thanks to

David W. Aha
Head, Adaptive Systems Section
Navy Center for Applied Research in Artificial Intelligence
Naval Research Laboratory (Code 5514); Washington, DC
david.aha@nrl.navy.mil

IJCAI 2019 WORKSHOP ON EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)

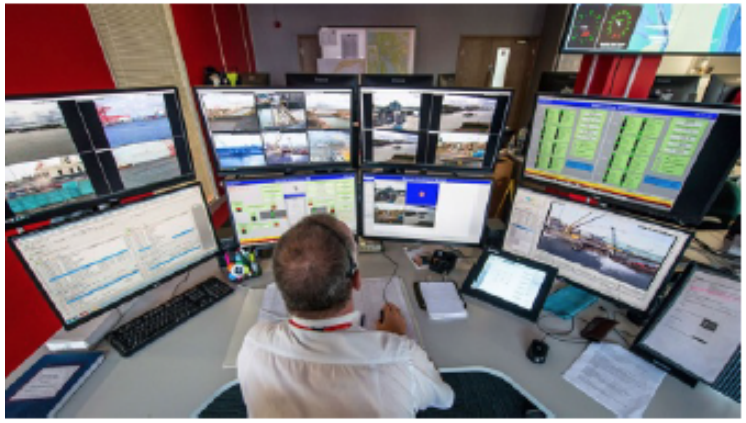
10, 11, or 12 August, 2019. Macau, China

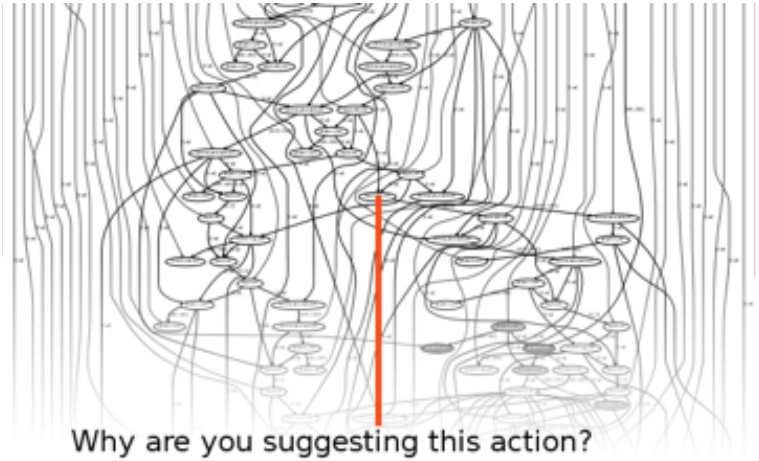
https://www.ijcai19.org/

Co-Chairs:
David Aha (NRL, USA)
Daniele Magazzeni (King's College London)
Tim Miller (University of Melbourne, Australia)
Rosina Weber (Drexel University)

56 submissions !

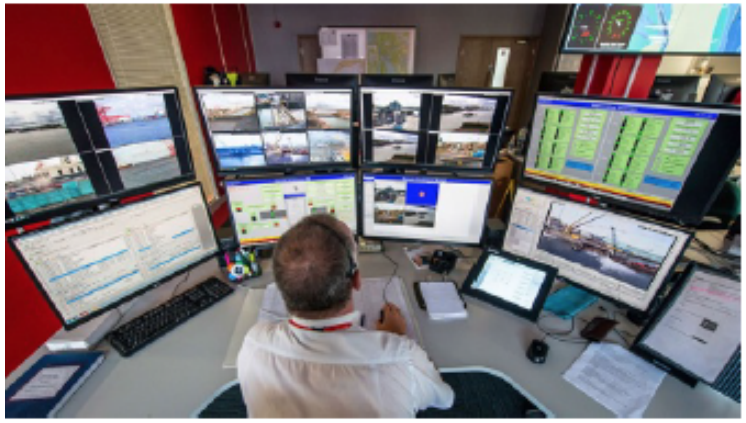
Explainable AI Planning (XAIP)

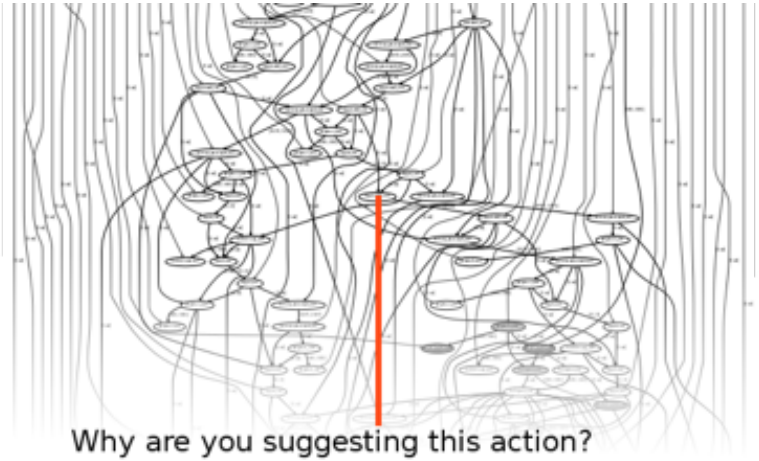




- **Need for Trust, Interaction, and Transparency**
- **Human operators (especially those in charge of /responsible for critical decisions) want to understand why the AI suggests something that they would not do.**
- **Intelligent Situational Awareness.**

Explainable AI Planning (XAIP)





White-Box AI needs to be explained as well !

(*some*) **Things to Be Explained**

- Q1: Why did you do that?
- Q2: Why didn't you do *something else*? (that I would have done)
- Q3: Why is what you propose to do more efficient/safe/cheap than something else? (that I would have done)
- Q4: Why can't you do that ?
- Q5: Why do I need to replan at this point?
- Q6: Why do I not need to replan at this point?

Fox, Long, Magazzeni. **Explainable Planning.** XAI @ IJCAI 2017.

Providing Explanations

- Q2: Why didn't you do *something else*? (that I would have done)

Quick (*and useless*) answer: because the heuristic evaluation was better for the decision the planner made.

One meaningful explanation is to demonstrate that the alternative action would prevent from finding a valid plan or would lead to a plan that is no better than the one found by the planner.

Contrastive Explanations

Contrastive Explanations

- Q2: Why didn't you do *something else*? (that I would have done)

Algorithm:

-re-run the planner up to the decision point questioned by the human

-inject the human choice

-plan from the state obtained after applying the action chosen by the human

(a)

Contrastive Explanations

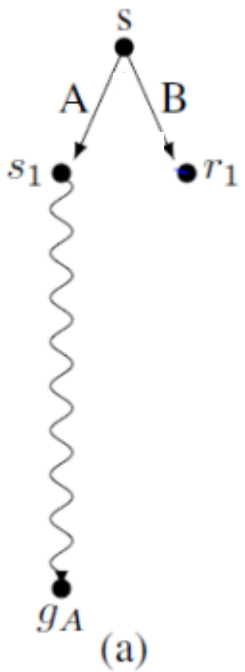
- Q2: Why didn't you do *something else*? (that I would have done)

Algorithm:

-re-run the planner up to the decision point questioned by the human

-inject the human choice

-plan from the state obtained after applying the action chosen by the human



Contrastive Explanations

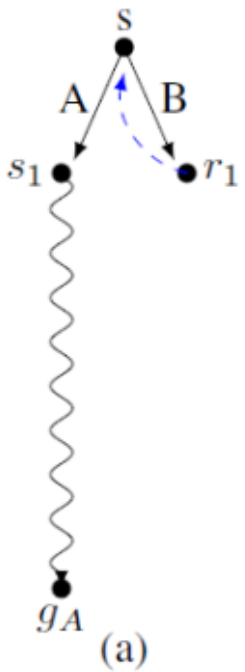
- Q2: Why didn't you do *something else*? (that I would have done)

Algorithm:

-re-run the planner up to the decision point questioned by the human

-inject the human choice

-plan from the state obtained after applying the action chosen by the human



(a)

Contrastive Explanations

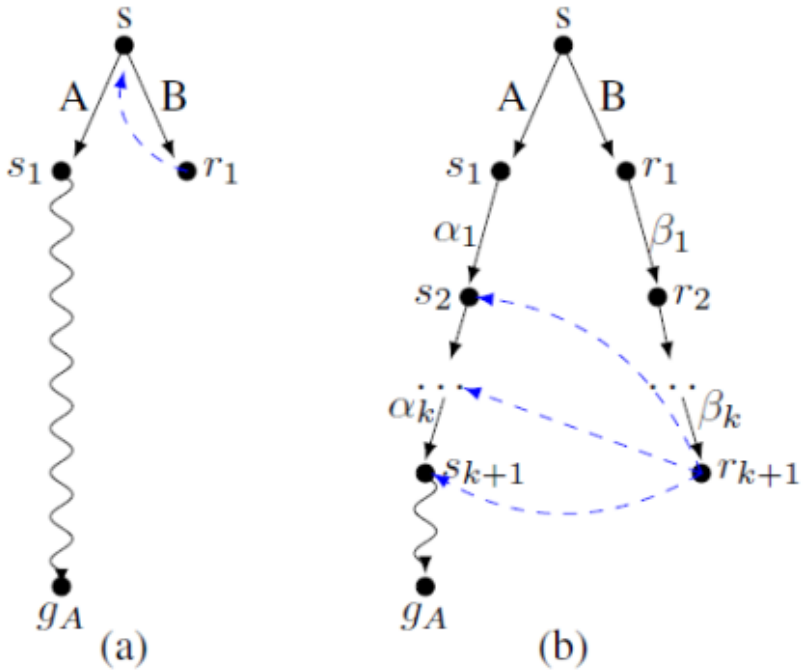
- Q2: Why didn't you do *something else*? (that I would have done)

Algorithm:

-re-run the planner up to the decision point questioned by the human

-inject the human choice

-plan from the state obtained after applying the action chosen by the human



Contrastive Explanations

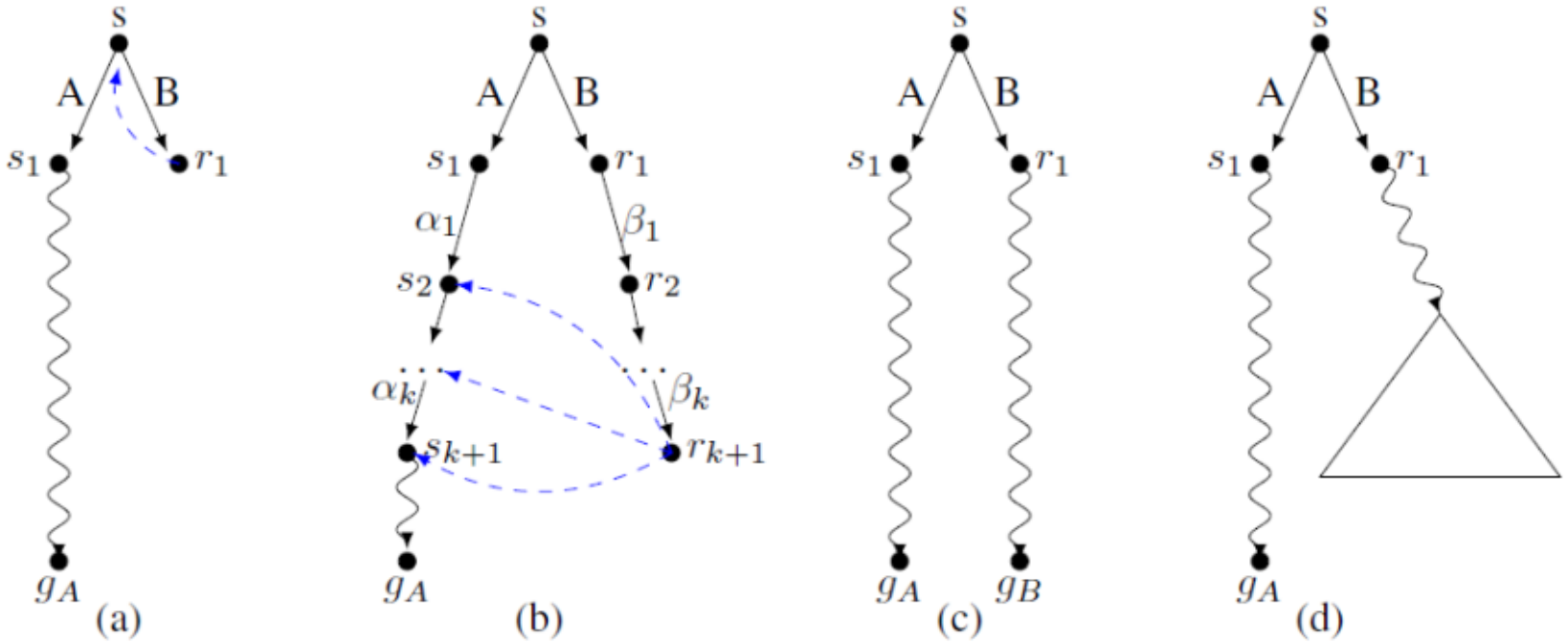
- Q2: Why didn't you do *something else*? (that I would have done)

Algorithm:

-re-run the planner up to the decision point questioned by the human

-inject the human choice

-plan from the state obtained after applying the action chosen by the human



Q&A through Formal Constraints

```
0.00: (goto_waypoint kenny wp0 wp2)   [1.45]
1.45: (goto_waypoint kenny wp2 wp1)   [2.00]
3.45: (goto_waypoint kenny wp1 wp2)   [2.00]
5.45: (goto_waypoint kenny wp2 wp5)   [2.00]
7.45: (goto_waypoint kenny wp5 wp3)   [4.68]
12.13: (goto_waypoint kenny wp3 wp5)   [4.68]
16.81: (goto_waypoint kenny wp5 wp0)   [0.99]
```

```
(:goal (and
        (visited kenny wp1)
        (visited kenny wp2)
        (visited kenny wp3)
        (visited kenny wp5)
)
```

Q: Why didn't kenny recharge the battery at wp4 ?

```
(:durative-action recharge
 :parameters (?r – robot ?w – waypoint)
 :duration (> ?duration 0)
 :condition (at start (robot_at ?r ?w)
                 at start (charge_available ?))
 :effect (increase (charge ?r) (* #t  rate_of_charge ?w))
```

Q&A through Formal Constraints

```
0.00: (goto_waypoint kenny wp0 wp2)   [1.45]
1.45: (goto_waypoint kenny wp2 wp1)   [2.00]
3.45: (goto_waypoint kenny wp1 wp2)   [2.00]
5.45: (goto_waypoint kenny wp2 wp5)   [2.00]
7.45: (goto_waypoint kenny wp5 wp3)   [4.68]
12.13: (goto_waypoint kenny wp3 wp5)   [4.68]
16.81: (goto_waypoint kenny wp5 wp0)   [0.99]
```

```
(:goal (and
        (visited kenny wp1)
        (visited kenny wp2)
        (visited kenny wp3)
        (visited kenny wp5)
)
```

Q: Why didn't kenny recharge the battery at wp4 ?

```
(:durative-action recharge
 :parameters (?r – robot ?w – waypoint)
 :duration (> ?duration 0)
 :condition (at start (robot_at ?r ?w)
                 at start (charge_available ?))
 :effect (increase (charge ?r) (* #t  rate_of_charge ?w))
          (at end (charged_at ?r ?w))
```

```
(:goal (and

        (visited kenny wp1)
        (visited kenny wp2)
        (visited kenny wp3)
        (visited kenny wp5)
(charged_at kenny wp4)
)
```

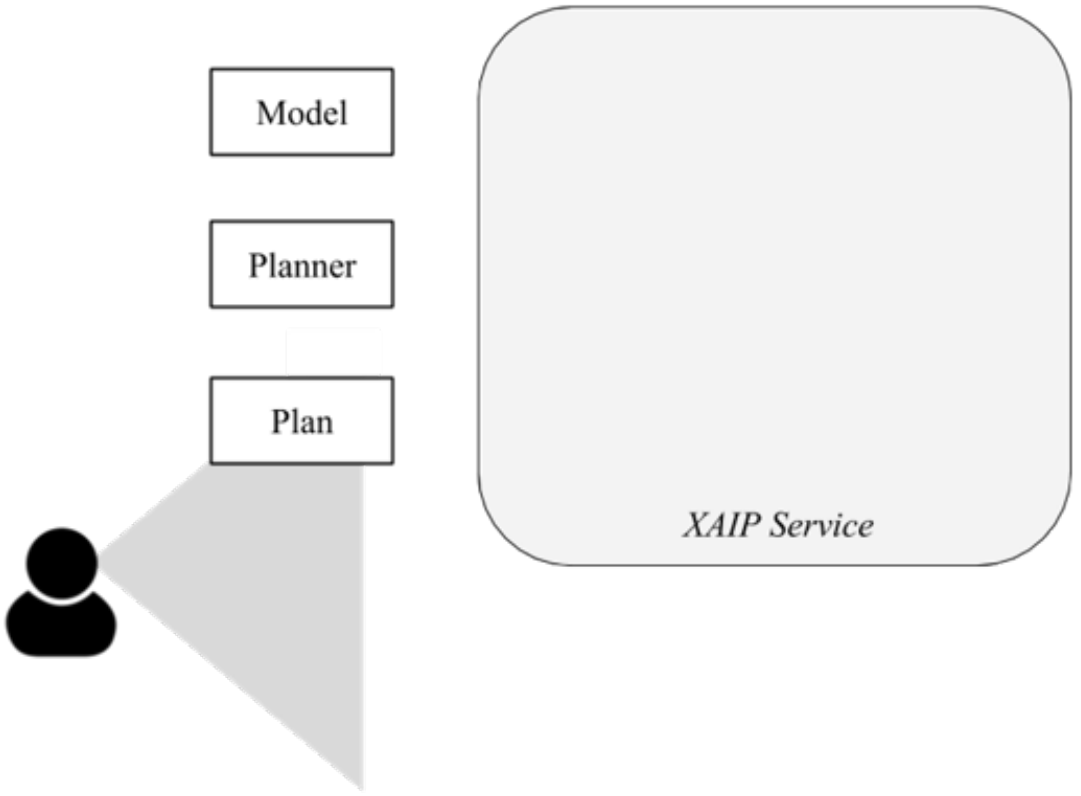
Ben Krarup, Michael Cashmore, Daniele Magazzeni, Tim Miller.
Model-Based Contrastive Explanations for Explainable Planning. XAIP-19.

Explainable Planning as a Service

1. **The supervisor will not accept an explanation generated by a planner different from the one that they use and whose performance they trust.**

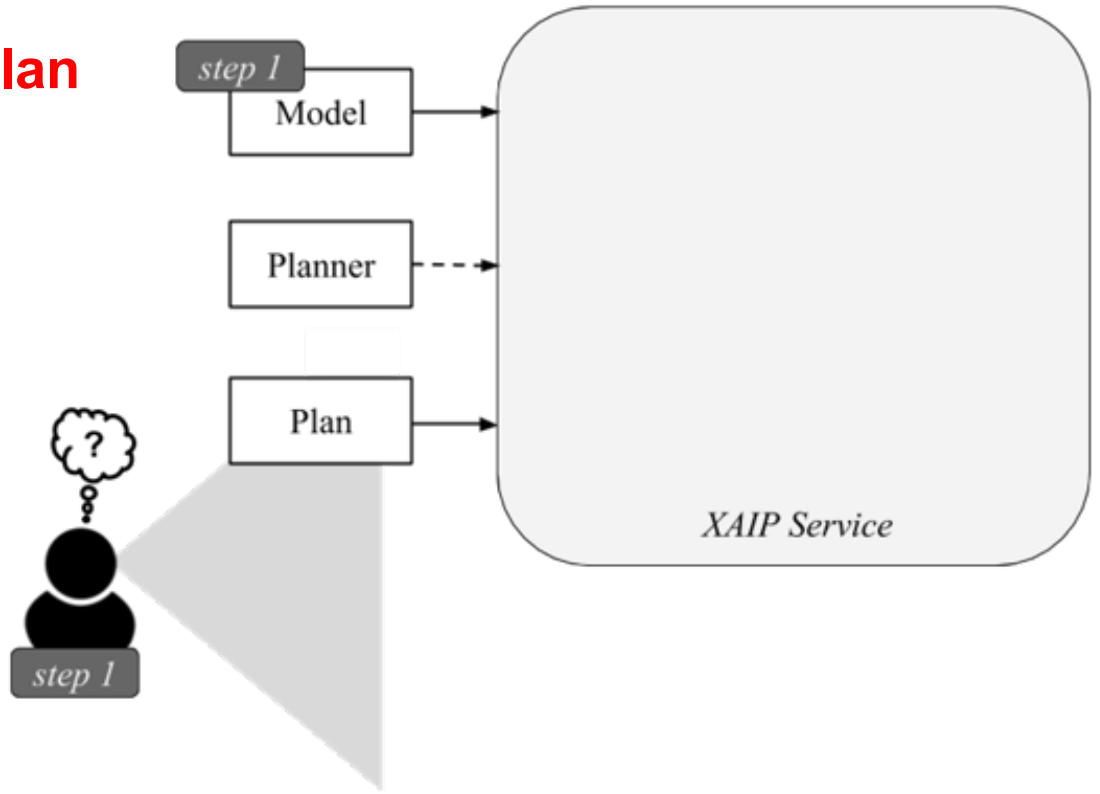
2. **The supervisor will not accept an explanation generated using a model that differs from the one that has been developed by the company's engineers, verified, and is trusted by the supervisor.**

Explainable Planning as a Service



Explainable Planning as a Service

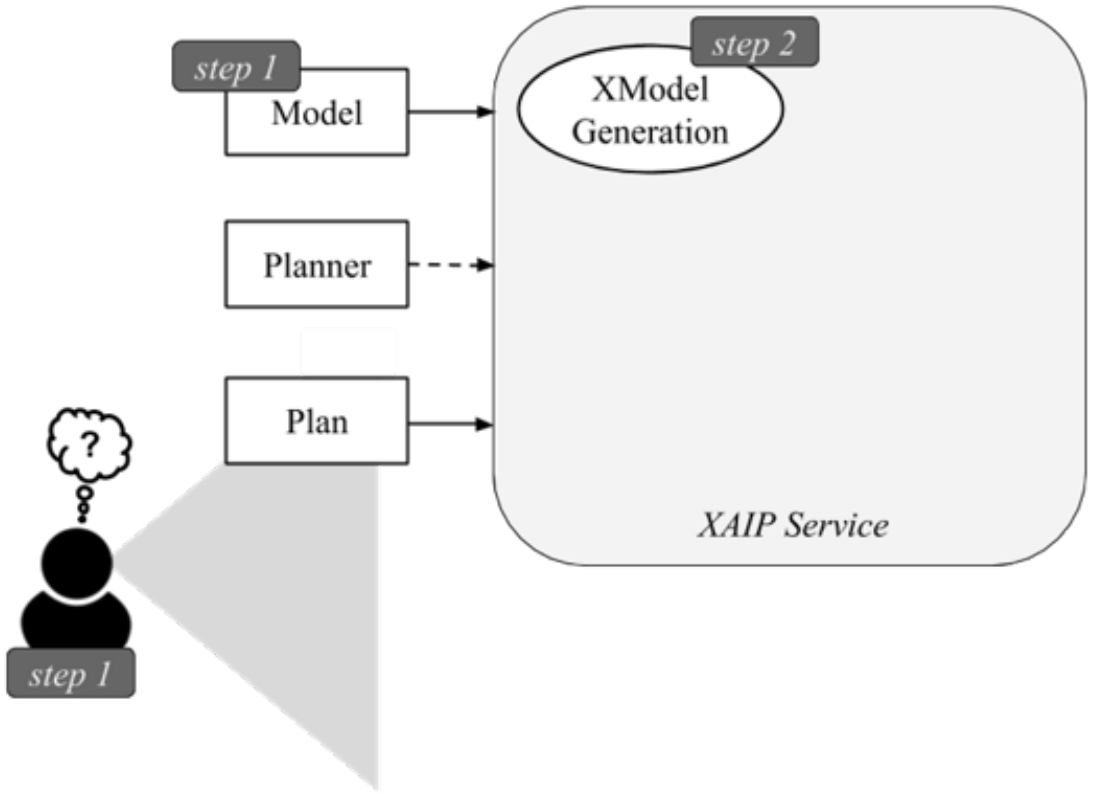
Step 1: Questioning the plan



The XAIP Service takes as input:
the model, the plan, and the question from the user

Explainable Planning as a Service

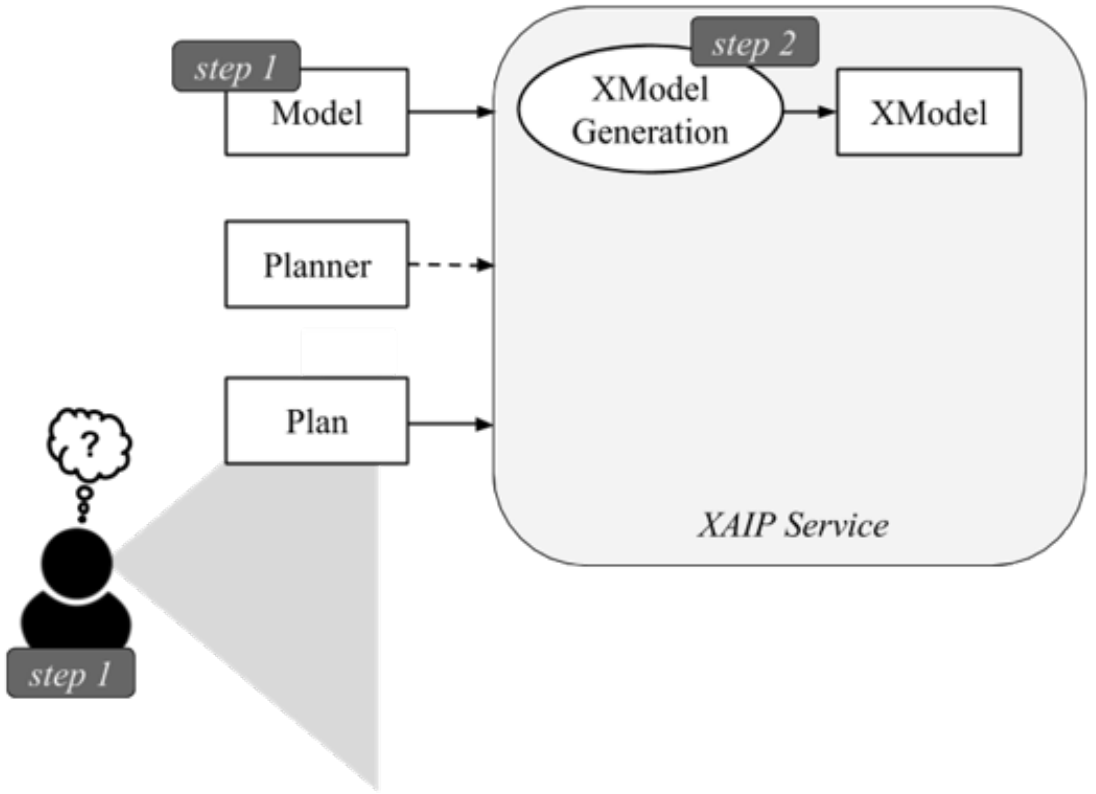
Step 2: Deriving the Hypothetical Model
(using formal constraints)



The query is translated into constraints

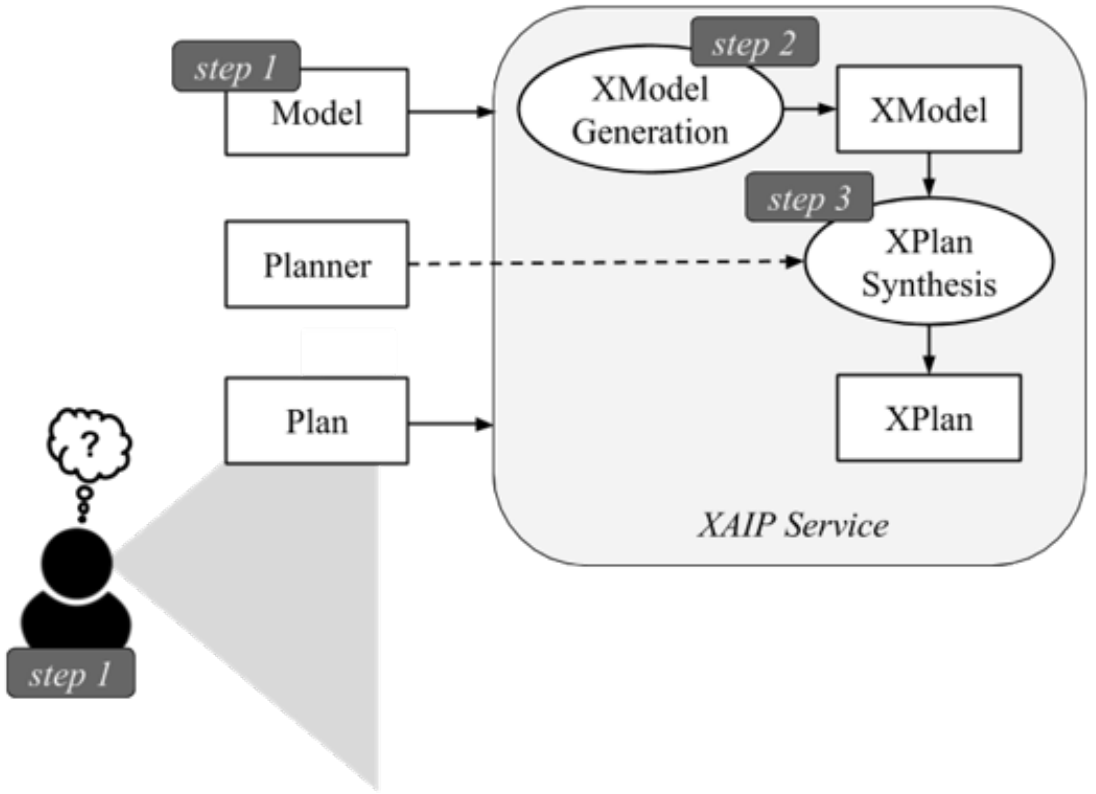
Explainable Planning as a Service

Step 2: Deriving the Hypothetical Model
(using formal constraints)



Explainable Planning as a Service

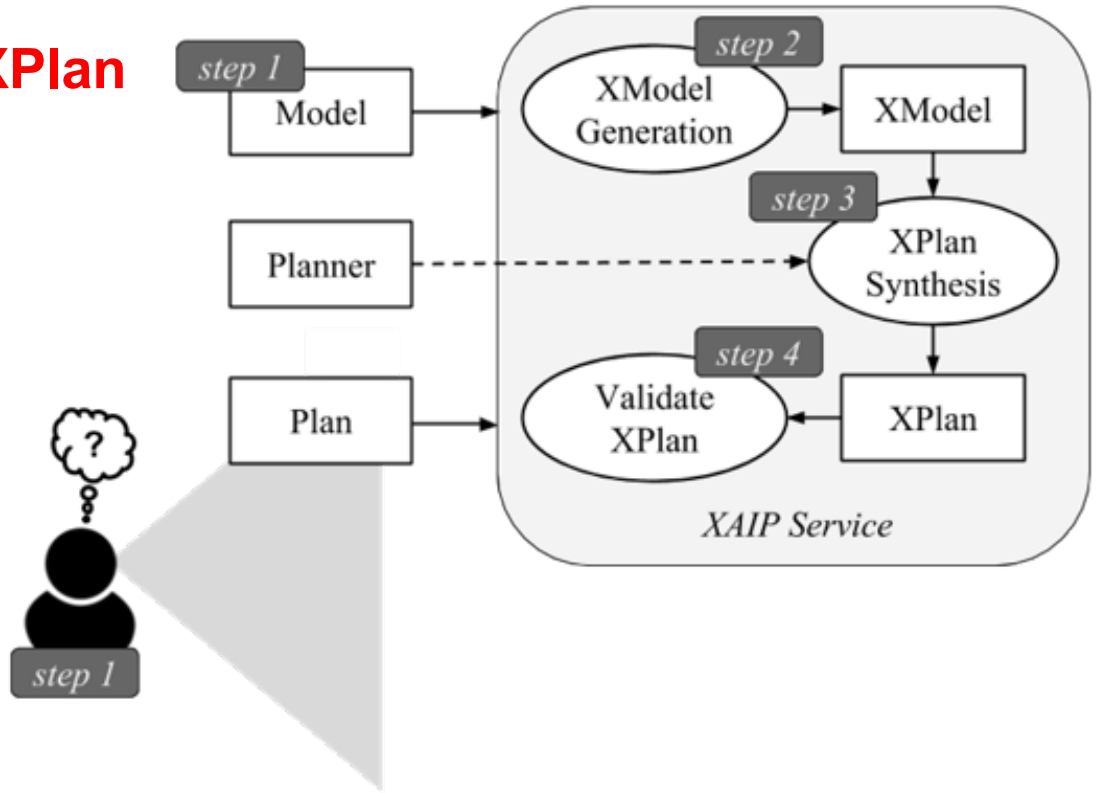
Step 3: Producing the alternative plan (XPlan)



The original planner must be used

Explainable Planning as a Service

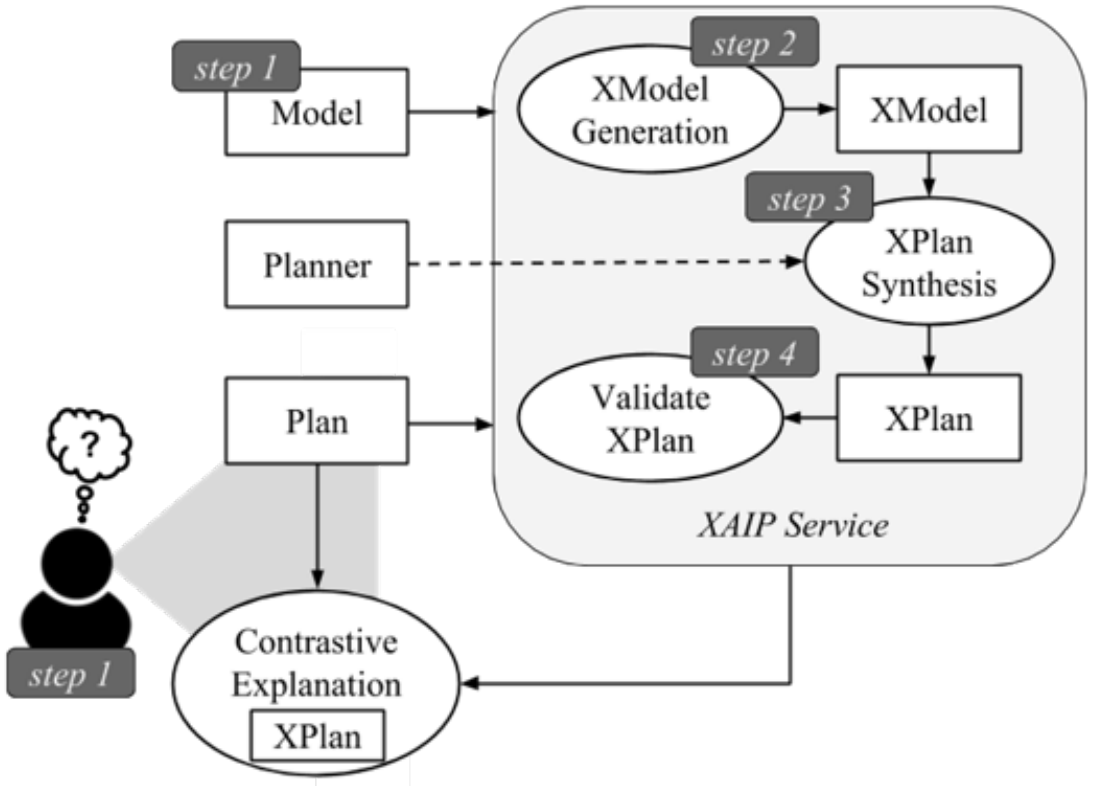
Step 4: Validation of the XPlan



The original planner must be used
The XPlan must be VALid according to the original model

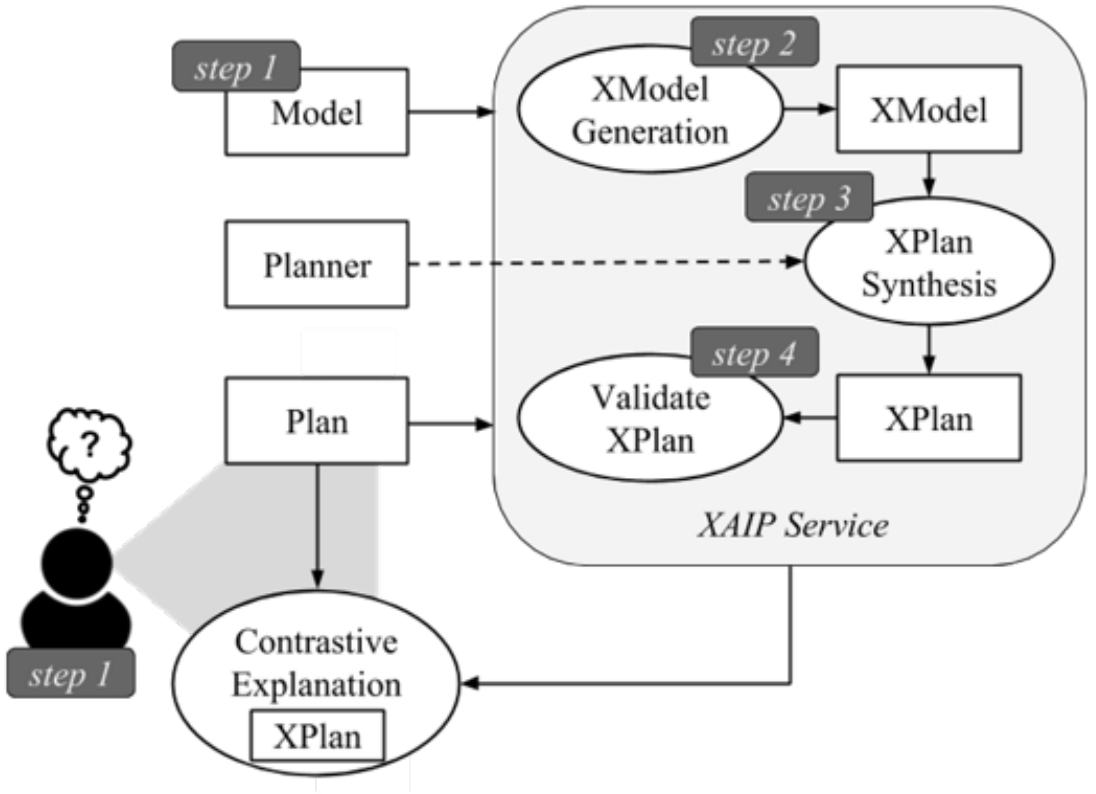
Explainable Planning as a Service

Forming the Contrastive Explanation



Explainable Planning as a Service

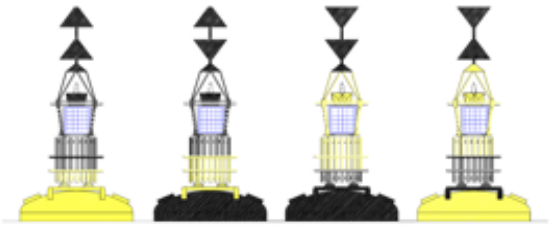
Iterative Process !



Maintenance of Aids to Navigation

Assets and Waypoints

- Harbors
- Lighthouses
- Lightvessels - maintenance time: 12h
- Buoys
  - Type 1 - maintenance time: 6h
  - Type 2 - maintenance time: 3h
- Survey – 3 h



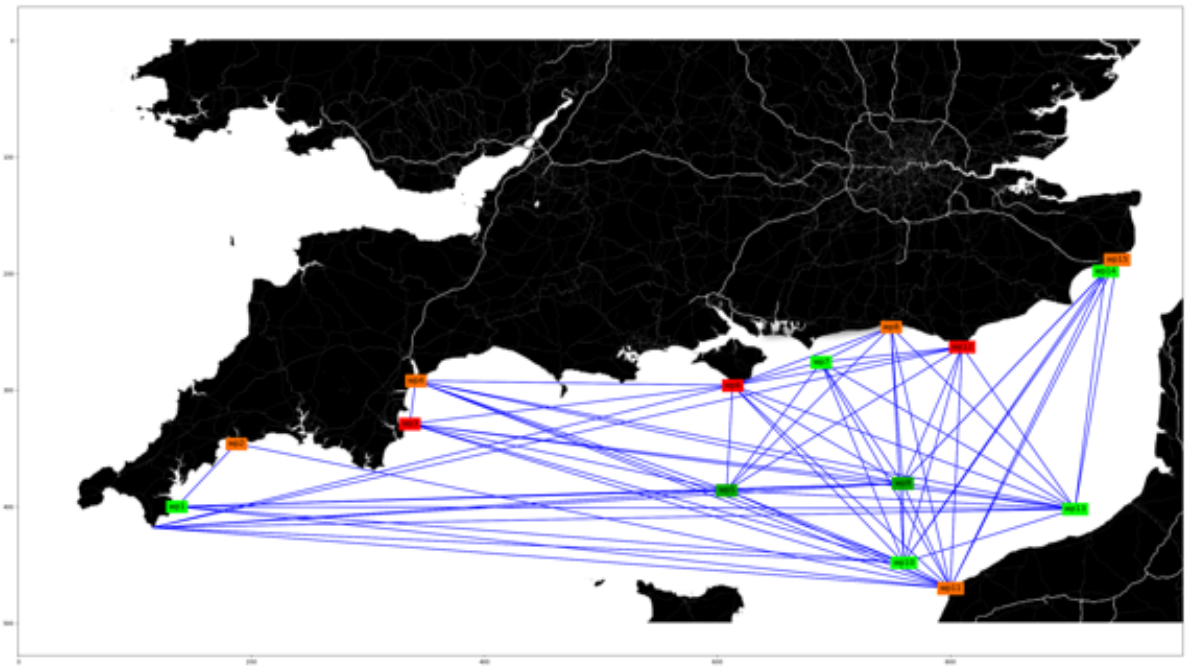
Ships:

- Alert
  - Can survey
- Patricia
  - Can survey, maintain and transport
- Galatia
  - Can survey, maintain and transport

Innovate UK

Operator specifies the mission (Goals)

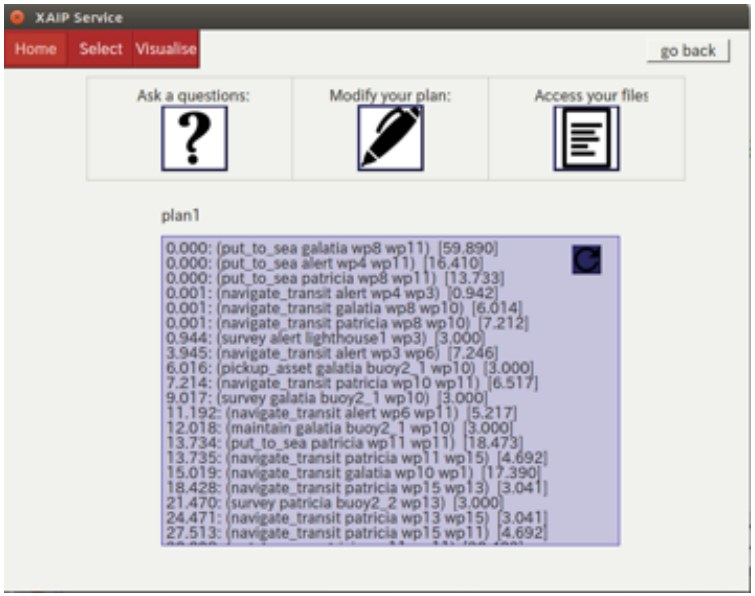
```
(:goal (and
 (survey_done buoy1_1)
(maintenance_done buoy1_1)
(survey_done buoy2_1)
(maintenance_done buoy2_1)
(maintenance_done buoy2_3)
(survey_done buoy2_2)
(at_asset buoy2_1 wp11)
(survey_done lighthouse1)
(> (fuel_level Galatia) 0)
(> (fuel_level Patricia) 0)
(> (fuel_level Alert) 0)
))
```



Innovate UK

AI generates a Plan

XAIP Framework: Original plan visualisation



0.000: (put_to_sea galatia wp8 wp11) [59.890]
0.000: (put_to_sea alert wp4 wp11) [16.410]
0.000: (put_to_sea patricia wp8 wp11) [13.733]
0.001: (navigate_transit alert wp4 wp3) [0.942]
0.001: (navigate_transit galatia wp8 wp10) [6.014]
0.001: (navigate_transit patricia wp8 wp10) [7.212]
0.944: (survey alert lighthouse1 wp3) [3.000]
3.945: (navigate_transit alert wp3 wp6) [7.246]
6.016: (pickup_asset galatia buoy2_1 wp10) [3.000]
7.214: (navigate_transit patricia wp10 wp11) [6.517]
9.017: (survey galatia buoy2_1 wp10) [3.000]
11.192: (navigate_transit alert wp6 wp11) [5.217]
12.018: (maintain galatia buoy2_1 wp10) [3.000]
13.734: (put_to_sea patricia wp11 wp11) [18.473]
13.735: (navigate_transit patricia wp11 wp15) [4.692]
15.019: (navigate_transit galatia wp10 wp1) [17.390]
18.428: (navigate_transit patricia wp15 wp13) [3.041]
21.470: (survey patricia buoy2_2 wp13) [3.000]
24.471: (navigate_transit patricia wp13 wp15) [3.041]
27.513: (navigate_transit patricia wp15 wp11) [4.692]
32.208: (put_to_sea patricia wp11 wp11) [30.423]
32.209: (navigate_transit patricia wp11 wp10) [6.517]
32.409: (survey galatia buoy1_1 wp1) [3.000]
35.410: (maintain galatia buoy1_1 wp1) [6.000]
41.411: (navigate_transit galatia wp1 wp5) [12.318]
41.727: (navigate_transit patricia wp10 wp14) [9.906]
51.633: (maintain patricia buoy2_3 wp14) [3.000]
53.730: (navigate_transit galatia wp5 wp11) [6.159]
54.634: (navigate_transit patricia wp14 wp15) [3.302]
57.937: (navigate_transit patricia wp15 wp11) [4.692]
59.889: (place_asset galatia buoy2_1 wp11) [3.000]

Innovate UK

Operator can then ask questions…

Question

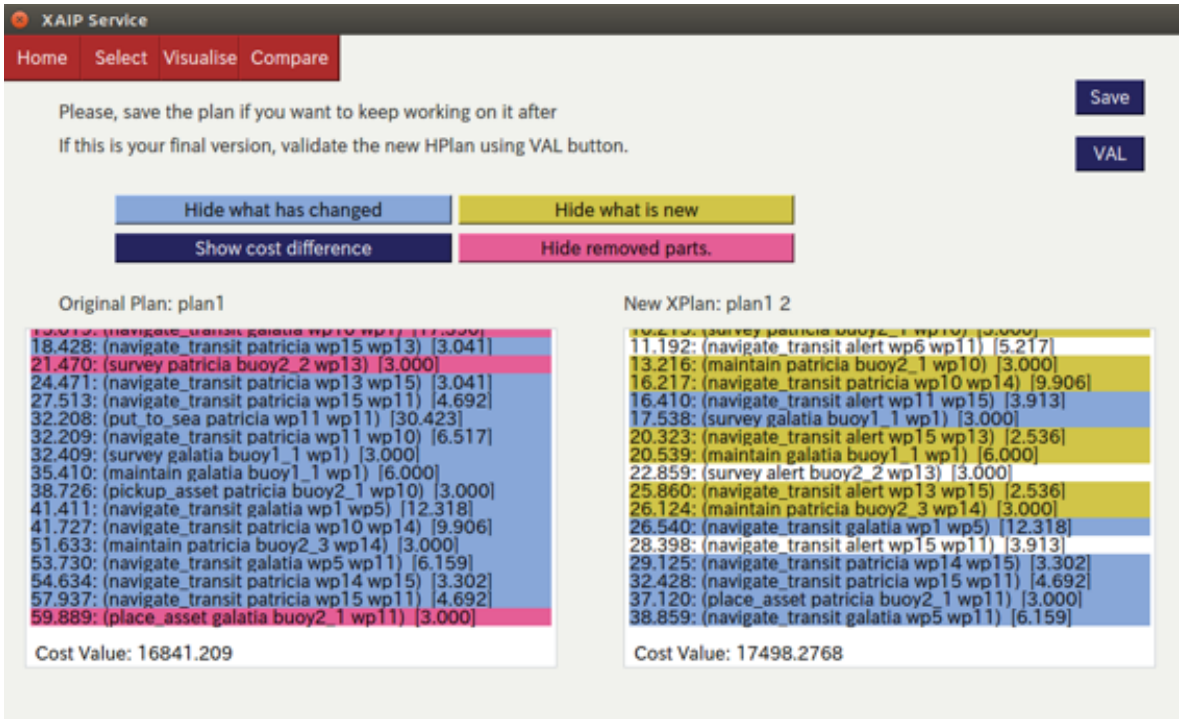
- "why did Patricia *not* pickup buoy2_1?"

0.000: (put_to_sea galatia wp8 wp11) [59.890]
0.000: (put_to_sea alert wp4 wp11) [16.410]
0.000: (put_to_sea patricia wp8 wp11) [13.733]
0.001: (navigate_transit alert wp4 wp3) [0.942]
0.001: (navigate_transit galatia wp8 wp10) [6.014]
0.001: (navigate_transit patricia wp8 wp10) [7.212]
0.944: (survey alert lighthouse1 wp3) [3.000]
3.945: (navigate_transit alert wp3 wp6) [7.246]
6.016: (pickup_asset galatia buoy2_1 wp10) [3.000]
7.214: (navigate_transit patricia wp10 wp11) [6.517]
9.017: (survey galatia buoy2_1 wp10) [3.000]
11.192: (navigate_transit alert wp6 wp11) [5.217]
12.018: (maintain galatia buoy2_1 wp10) [3.000]
13.734: (put_to_sea patricia wp11 wp11) [18.473]
13.735: (navigate_transit patricia wp11 wp15) [4.692]
15.019: (navigate_transit galatia wp10 wp1) [17.390]
18.428: (navigate_transit patricia wp15 wp13) [3.041]
21.470: (survey patricia buoy2_2 wp13) [3.000]
24.471: (navigate_transit patricia wp13 wp15) [3.041]
27.513: (navigate_transit patricia wp15 wp11) [4.692]
32.208: (put_to_sea patricia wp11 wp11) [30.423]
32.209: (navigate_transit patricia wp11 wp10) [6.517]
32.409: (survey galatia buoy1_1 wp1) [3.000]
35.410: (maintain galatia buoy1_1 wp1) [6.000]
41.411: (navigate_transit galatia wp1 wp5) [12.318]
41.727: (navigate_transit patricia wp10 wp14) [9.906]
51.633: (maintain patricia buoy2_3 wp14) [3.000]
53.730: (navigate_transit galatia wp5 wp11) [6.159]
54.634: (navigate_transit patricia wp14 wp15) [3.302]
57.937: (navigate_transit patricia wp15 wp11) [4.692]
59.889: (place_asset galatia buoy2_1 wp11) [3.000]

Innovate UK

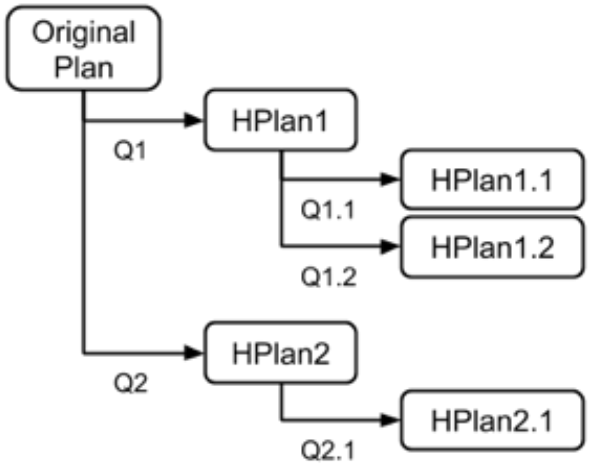
XAIP provides a contrastive explanation

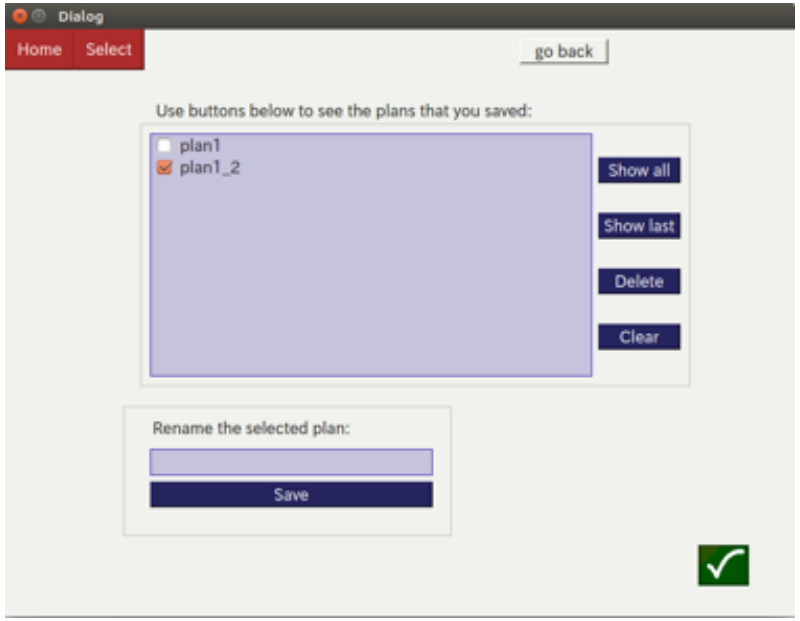
Explanation provided by the XAIP Service



Innovate UK

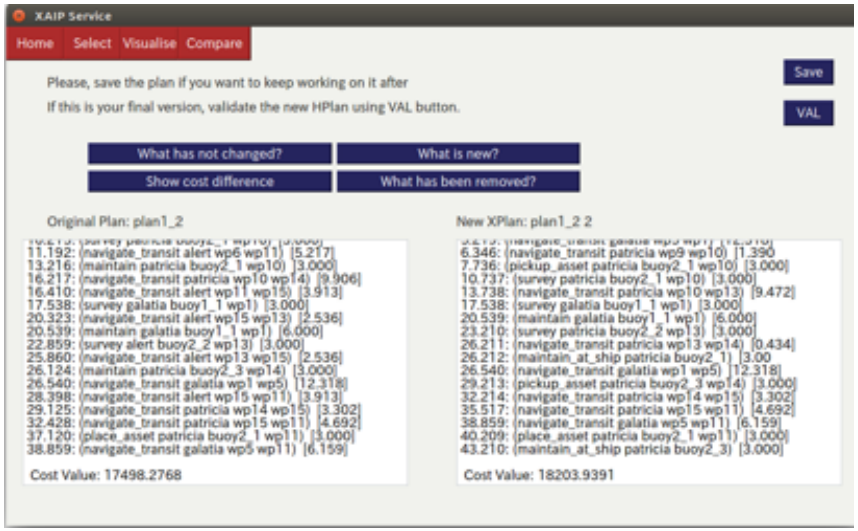
The process can be iterated

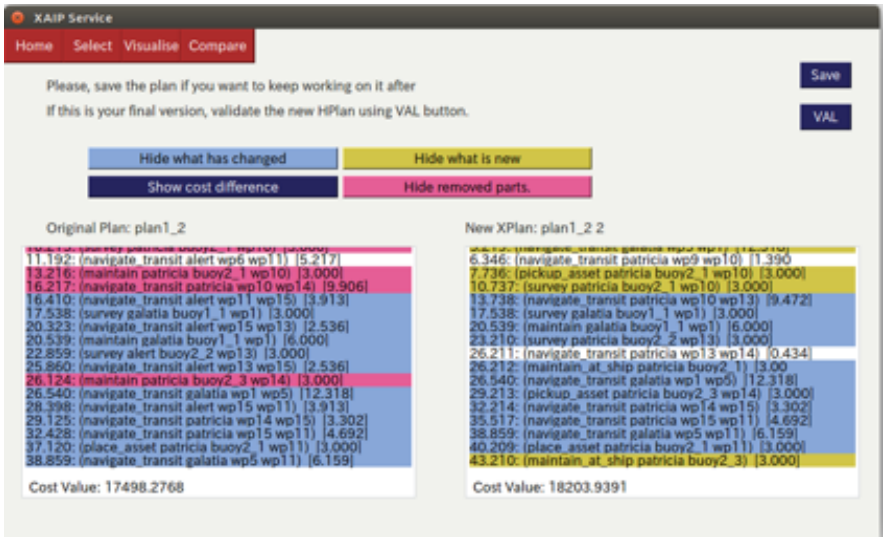




Innovate UK

- Why Patricia maintained buoy2_3 at wp14?





Innovate UK

Explainable Planning as a Service

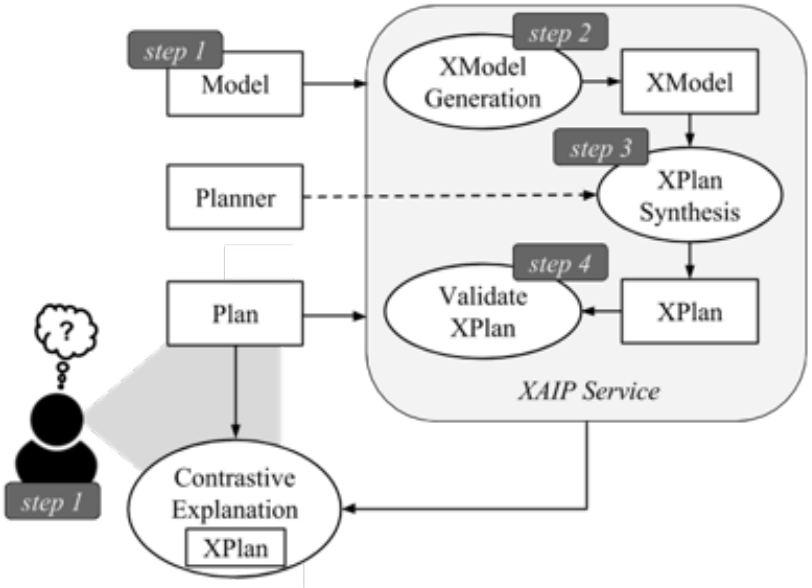
How the user question can be understood, properly taking into account the context in which it was asked?

How to formally characterize the set of questions that can be answered with contrastive explanations?

How constraints can be formally encoded in the XModel?

How to present explanations to the users?

How to assess the effectiveness of the provided explanations?



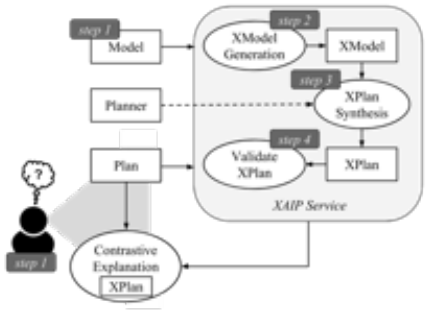
Explainable Planning as a Service

How to formally characterize the set of questions that can be answered with contrastive explanations?

How constraints can be formally encoded in the XModel?

Explaining the Space of Plans. (2018-2023)
Funded by U.S. Air Force Office of Scientific Research ($1m)
We are hiring !!!

Analysis of **plan-property dependencies**, where plan properties are Boolean properties of plans the user is interested in, and dependencies are entailment relations in plan space.
The answer to question "*Why A rather than B*?" consists of those properties C entailed by B.



Cashmore, Eifler, Hoffman, Magazzeni. **Explaining the Space of Plans through Plan-Property Dependencies.** **NOT at** IJCAI 2019.

Trust in Human-Machine Partnership (THuMP)

Funded by UKRI (£1.3m) THuMP involves 6 CoIs and 6 RAs (2019-2022). THuMP focusses on planning and allocation of resources in critical domains, bringing together experts in **AI**, **Provenance**, **Visualisation**, **Law** and **Social Science**.

1. **What are the technical challenges involved in creating Explainable AI Planning systems?**
2. **What are the technical, legal and social challenges involved in instantiating with explanations a planning system for solving resource allocation problems in critical domains?**
3. **What are the legal and social implications of enhancing machines with transparency and the ability to explain?**

Schlumberger

Save the Children

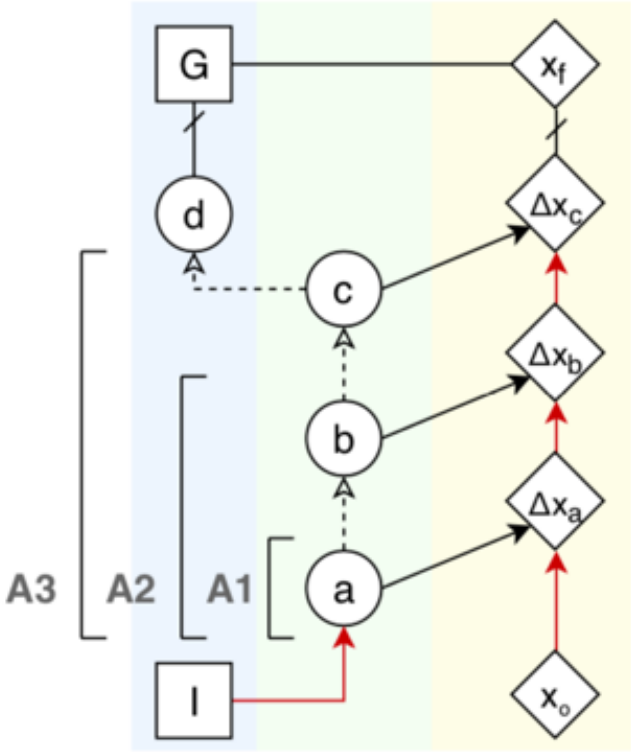
Hogan Lovells

UK Research and Innovation

EPSRC

Argumentation for Explainable Planning

- Abstracting **plan causality** into argumentation frameworks
- **ASPIC+** framework
  - **strict** rules represent action-state causality
  - **defeasible** rules represent action-action causality
- "A1 *causes* Δx_a and *enables* **b**"



Anna Collins, Daniele Magazzeni, Simon Parsons. **Towards an argumentation-based approach to explainable planning .** XAIP-19.

Argumentation for Explainable Planning

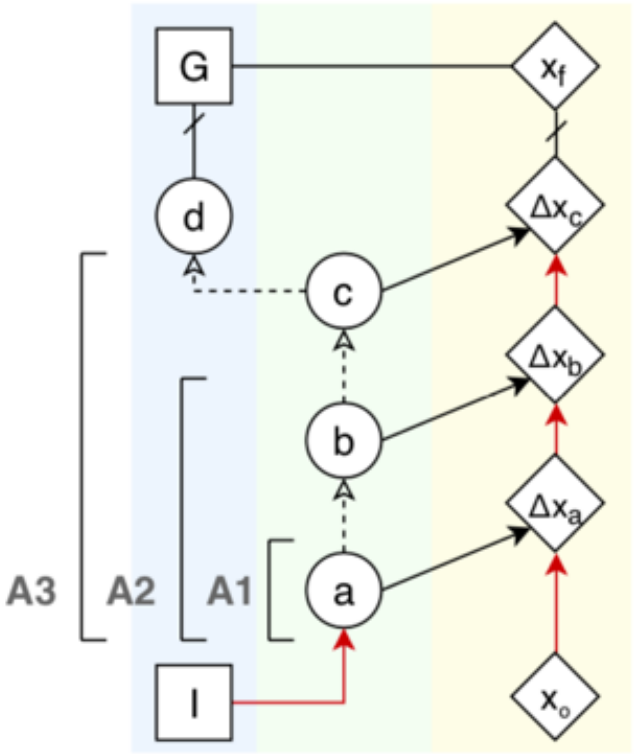
- Use nested arguments as explanations:

'Why b?'

b causes Δx_b and enables c

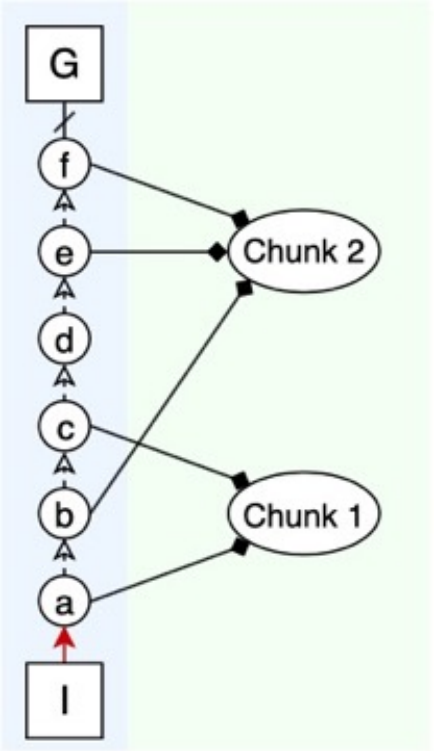
c causes Δx_c and enables d

⋮



Argumentation for Explainable Planning

- Forming 'chunks' comprising related actions
- Helps identify relevant information for a specific user question
- e.g. Grouping actions related to changes of one state variable



Argumentation for Explainable Planning

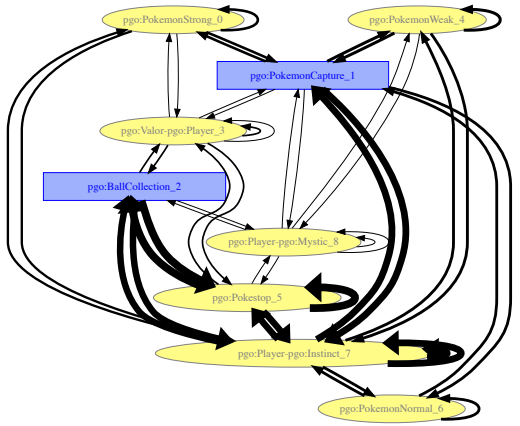
- Use argumentation to **structure the dialogue** between user and the planner.

- Use argumentation to represent **hierarchical explanations**.

Explanations strongly depend on the user !

Provenance and Explainable Planning

Provenance is a record that describes the people, institutions, entities, and activities, involved in producing, influencing, or delivering a piece of data or a thing in the world.



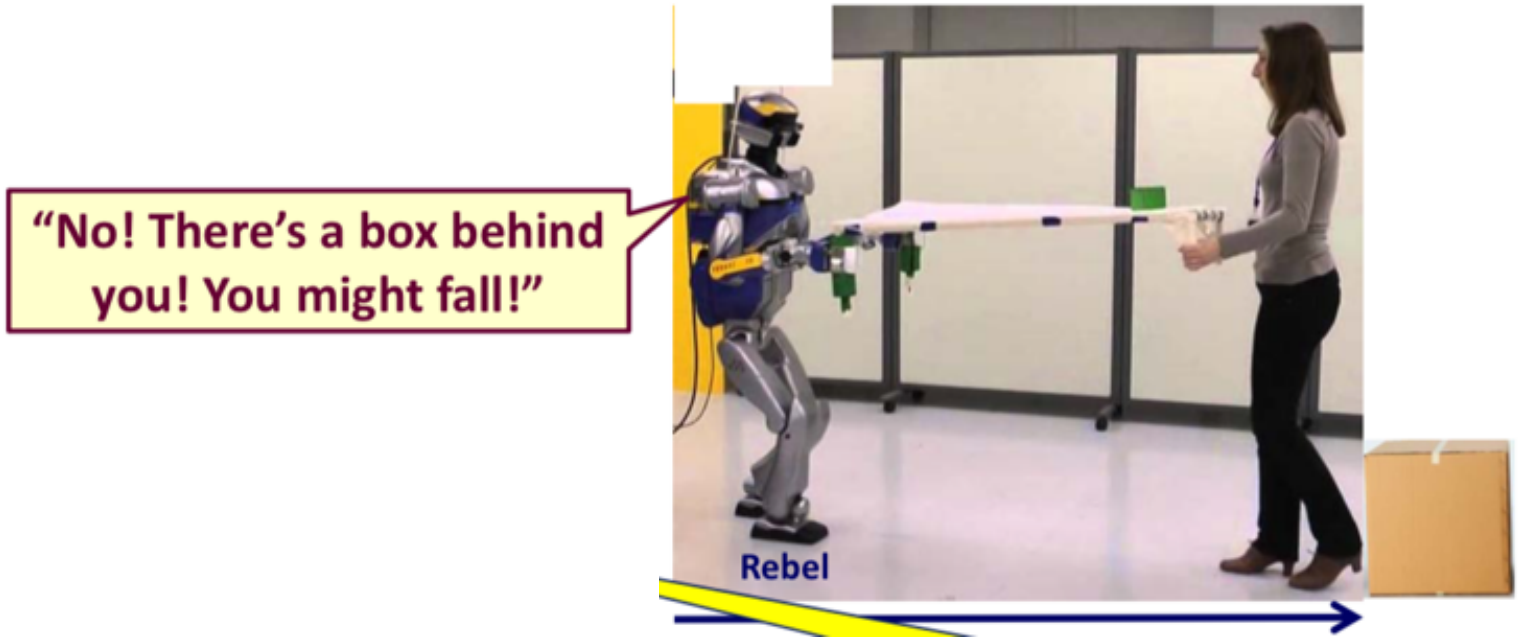
Detailed provenance graph can describe what led to decision/action, i.e. the "provenance of decision/action".

Provenance of information for building your model!

Prof Luc Moreau (KCL)

https://nms.kcl.ac.uk/luc.moreau/

Explaining Rebel Behavior in Goal Reasoning Agents.
D. Dannenhauer, M. Floyd, D. Magazzeni, D. Aha.
Proceedings of ICAPS-18 Workshop on Explainable Planning.



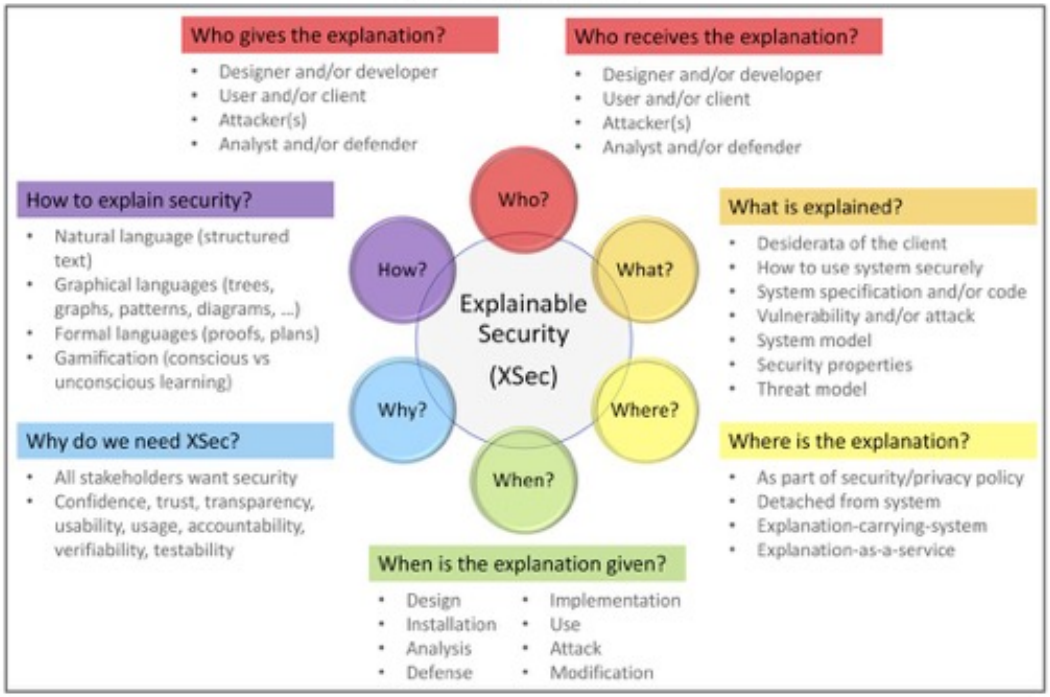
Moral Permissibility of Action Plans.
Felix Lindner, Robert Mattmüller and Bernhard Nebel. AAAI 2019.

XAIP for Training and up-skill.

XAIP for helping understand the models (and why they might be wrong!)

Explainable Security

Luca Viganò and **Daniele Magazzeni**
Department of Informatics
King's College London



Explainable Security has unique and complex characteristics:

- it involves **several different stakeholders** (developers, analysts, users and attackers) and
- is **multi-faceted by nature** (it requires reasoning about system model, threat model, properties of security, privacy and trust, concrete attacks, vulnerabilities, countermeasures).

Who?	What?	Where?
• All stakeholders might need explanations or need to act as explainer	• Explain several "things", at different levels of detail and with different aims	• Explanations can be made available in different places (X-carrying most promising)
When?	**Why?**	**How?**
• Design time • Runtime • Post-hoc	• We make too many mistakes because we don't understand or don't explain	• Explain proof or attack? • Explain explanation process • Trade-off with security threats

If you explain too much, they will attack:

- Explanations might provide information that an attacker can exploit.
- Explanations might need to be "relativized" and made less "powerful" by withholding details.

XAIP Workshops

Home | XAIP 2019 | Archive | Related Work | Contact

2nd ICAPS Workshop on Explainable Planning (XAIP-2019)

Berkeley, CA, 11-12 July 2019.

Co-Chairs:
Tathagata Chakraborti (IBM Research AI, USA)
Dustin Dannenhauer (Naval Research Laboratory, USA)
Joerg Hoffmann (Saarland University, Germany)
Daniele Magazzeni (King's College London, UK)

AIJ Special Issue on XAI

Topics: *Human-centric XAI, Theoretical and Philosophical Foundations, Knowledge Representation and Machine Learning, Interactive Explanation, Historical Perspectives, Case Study Reports*.

Important Dates

Submissions open: 1 August 2019
Submissions close: 1 March 2020
Final decisions: 1 August 2020

Guest Editors

Tim Miller (University of Melbourne, Australia)
Robert Hoffman (Institute for Human and Machine Cognition)
Daniele Magazzeni (King's College London, UK)
Julie Shah (Massachusetts Institute of Technology)
Randy Goebel (University of Alberta)