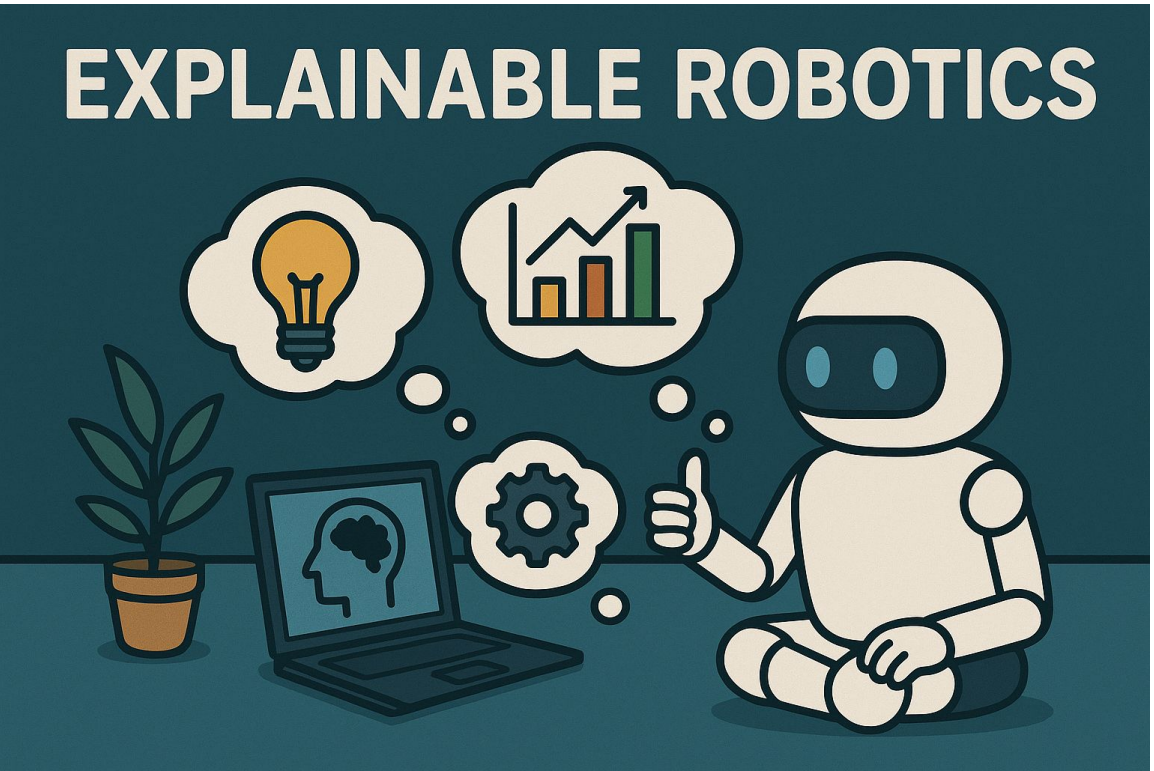


Senka Krivić, PhD

Assistant Professor
Faculty of Electrical Engineering
University of Sarajevo

King's College London

Eastern Europe Machine Learning Summer School, Sarajevo 2025

About me …

Bsc and Msc Control and Electronics
Faculty of Electrical Engineering, University of Sarajevo

2007

2012

Maintenance Engineer, Electrical Team Lead

Volkswagen Sarajevo, d.o.o

2014

PhD in Robot Learning, Dept. of Computer Science,
Universität Innsbruck, Austria
Intelligent and Interactive Systems group

2018

Research Associate

Department of Artificial Intelligence
King's College London
United Kingdom

2021

Lecturer and AI Consultant
Faculty of Electrical Engineering, University of Sarajevo, Burch University

2022

Assistant Professor

Department of Computer Science, Faculty of Electrical Engineering, University of Sarajevo

2025

+ Visiting Lecturer
Department of Artificial Intelligence
King's College London
United Kingdom

Work in XAI and Explainable Robotics

Human-Centred AI Lab in Sarajevo, UNSA

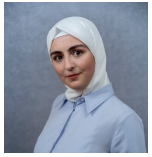
- Explainable Robotics
- Explainable and responsible AI
- AI - assisted learning

Ajla Karajko

Mubina Kamberović

Vahidin Hasić

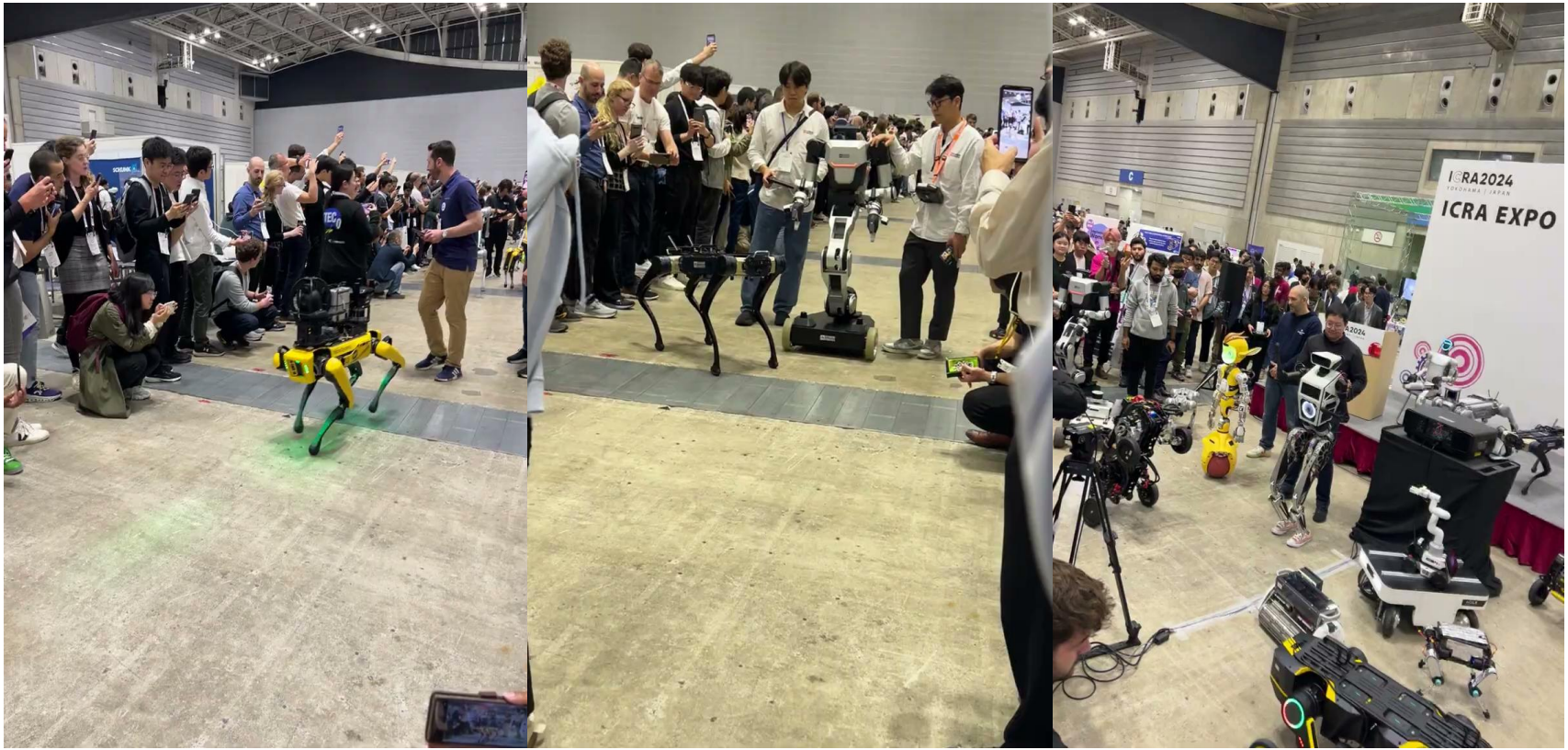
Amar Halilović



Amina Mević

Robotics today

2024 IEEE International Conference on Robotics and Automation



Robots are entering our daily lives

Current trends in robotics

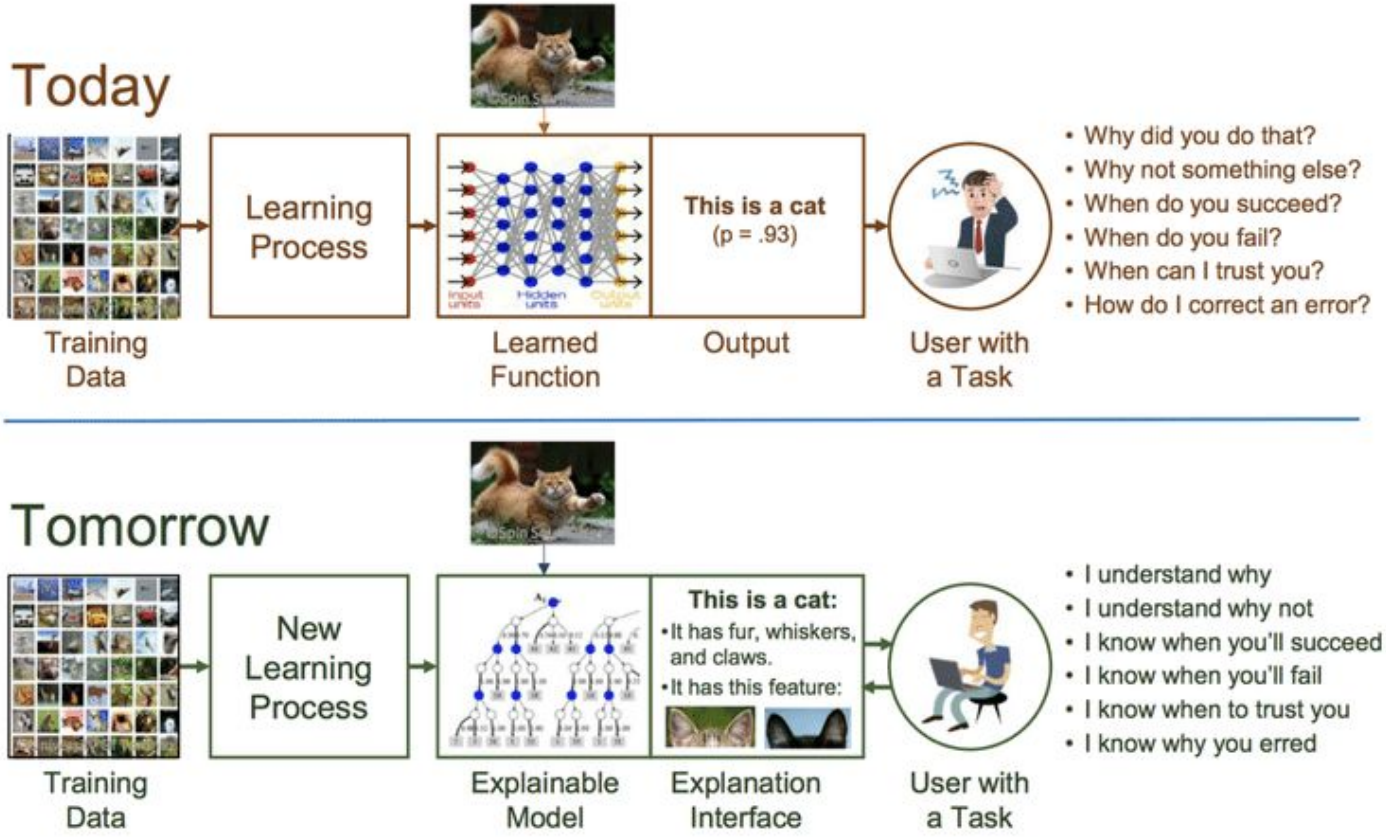
- Widespread adoption in homes, factories, and public spaces
- Growing complexity of robot behavior
- Rising expectations for accountability

Explainable AI

Explainable-AI explores and investigates methods to produce or complement AI models to make accessible and interpretable the internal logic and the outcome of the algorithms, making such process understandable by humans.

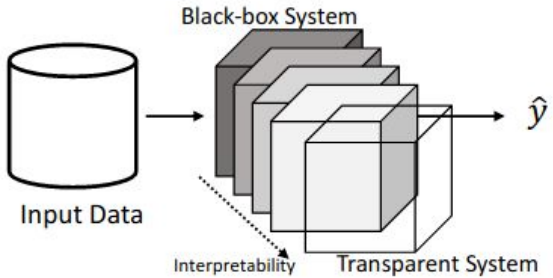
Explainable AI

Turek, M.: DARPA - Explainable Artificial Intelligence (XAI) Program (2017), https://www.darpa.mil/program/explainable-artificial-intelligence.

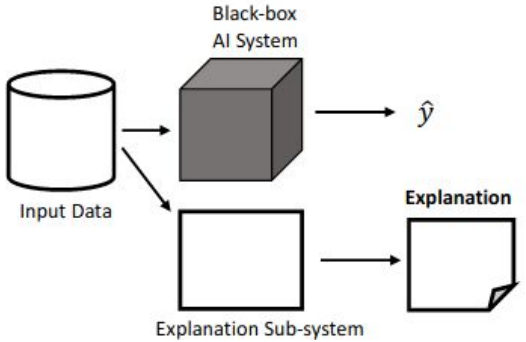


XAI Systems

Transparent-by-design systems



Post-hoc Explanation (black-box explanation) systems



[Mittelstadt et al. 2018]

Explainability vs Interpretability vs Mechanistic Interpretability

Interpretability: How well a human (often a developer/researcher) can make sense of the model's logic or structure.

Mechanistic Interpretability: A subtype of interpretability focused on low-level understanding of model internals.

Explainability: A broader term that includes any method or mechanism that helps understand and communicate a model's behavior, including to itself, other systems, or humans.

AI in media

THE WALL STREET JOURNAL.

Subscribe Now | Sign In

£1 FOR 3 MONTHS

Home World U.S. Politics Economy Business Tech Markets Opinion Life & Arts Real Estate WSJ. Magazine

Search

Qualcomm Won't Revive NXP Deal, After White House Flags China Concession

Trump: China to 'Reduce and Remove' Tariffs on American Cars

GlaxoSmithKline to Acquire Tesaro for $4.16 Billion

Smaller Films Finding Big Problems in China

CIO JOURNAL.

Tech Giants Launch New AI Tools as Worries Mount About Explainability

About 60% of 5,000 executives in IBM survey expressed concern with AI's 'black box'

Analytics & Optimization

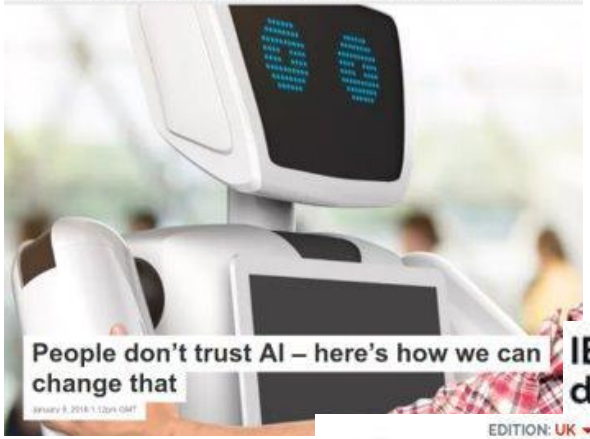
GDPR and Other Regulations Demand Explainable AI

Save

MAY 24 2017

THE CONVERSATION

Arts + Culture Business + Economy Cities Education Environment + Energy Health + Medicine Politics



People don't trust AI – here's how we can change that

COMPUTERWORLD

HEALTH IT, CONVERGENCE

How do you make doctors trust machines in an AI-driven clinical world?

During a panel at the MedCity INVEST Twin Cities conference leaders from the payer, provider and investor spaces spoke about how to actually drive adoption of AI tools in the clinical system.

By KEVIN TRUONG

AI at work: Staff 'hired and fired by algorithm'

25 March 2021 | Comments

IBM Researchers propose transparency docs for AI services

other technologies and industries, artificial intelligence will need to adopt supplier's declaration of conformity uments to build trust. How was that model built exactly?

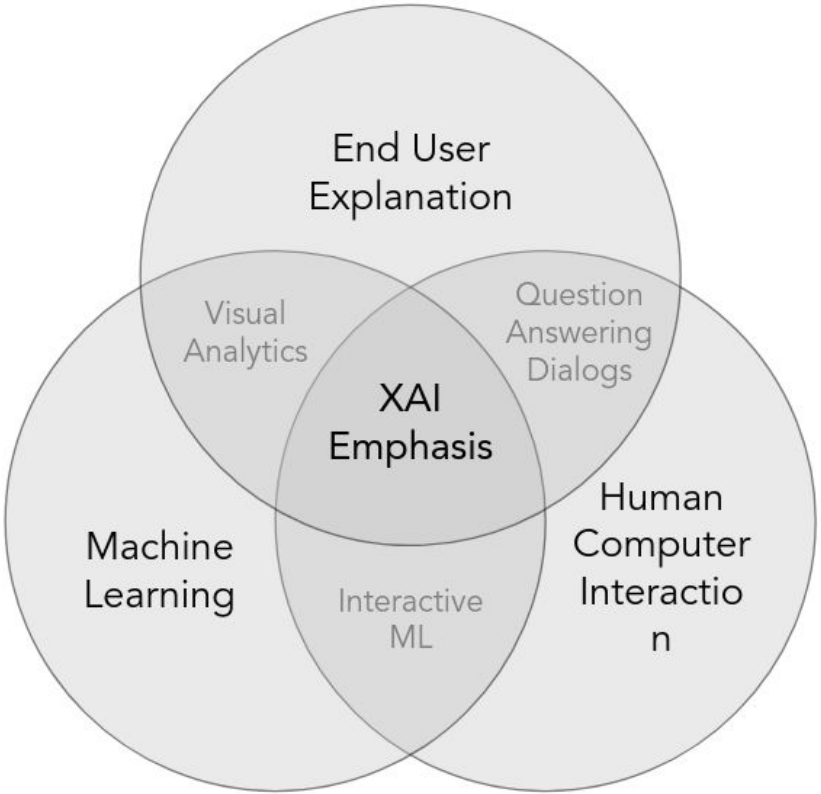
EDITION: UK

ZDNet

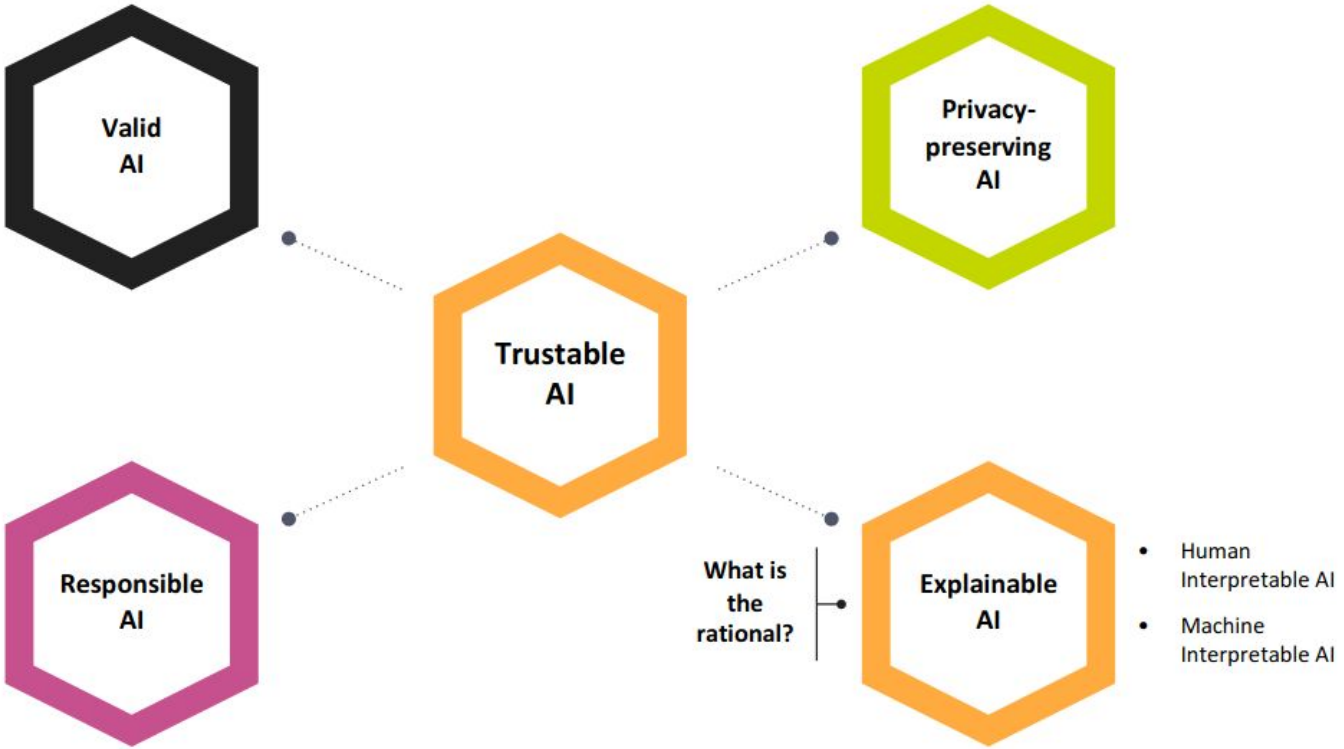
Scope of XAI

For millennia, philosophers have asked the questions about what constitutes an explanation, what is the function of explanations, and what are their structure

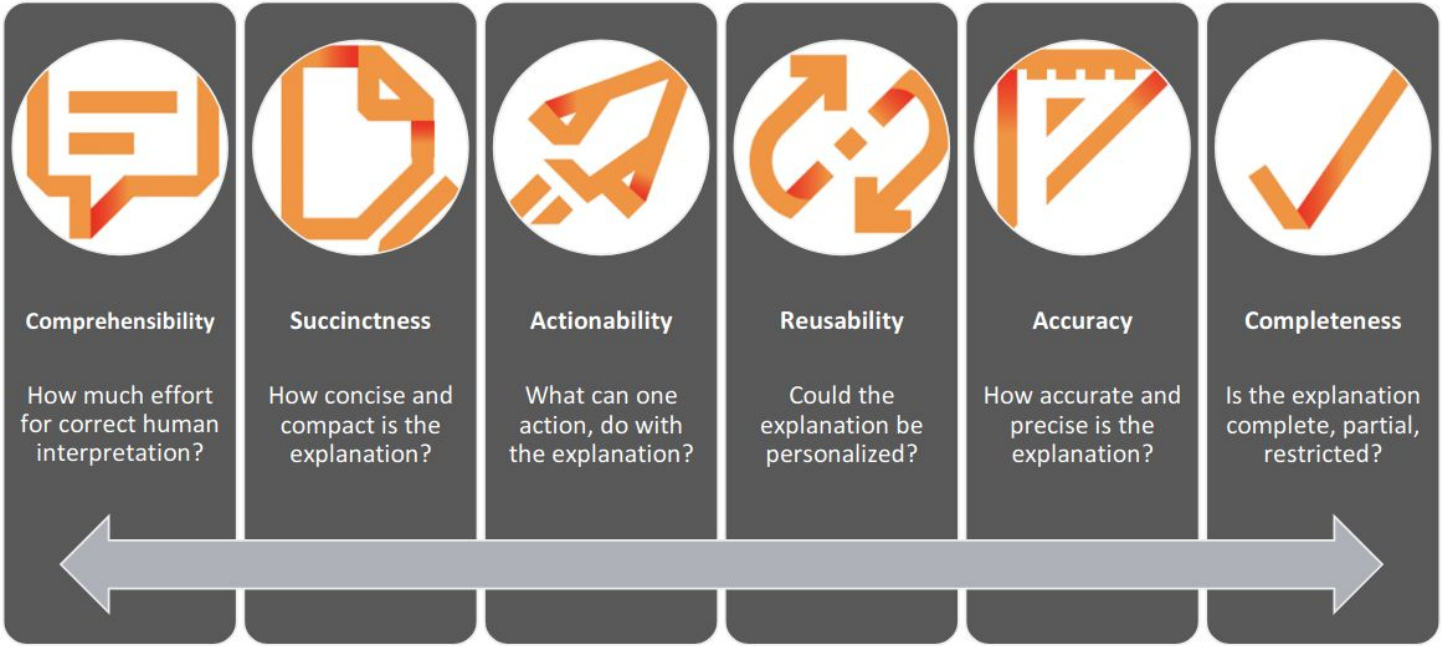
Tim Miller 2018



AI adoption: Requirements

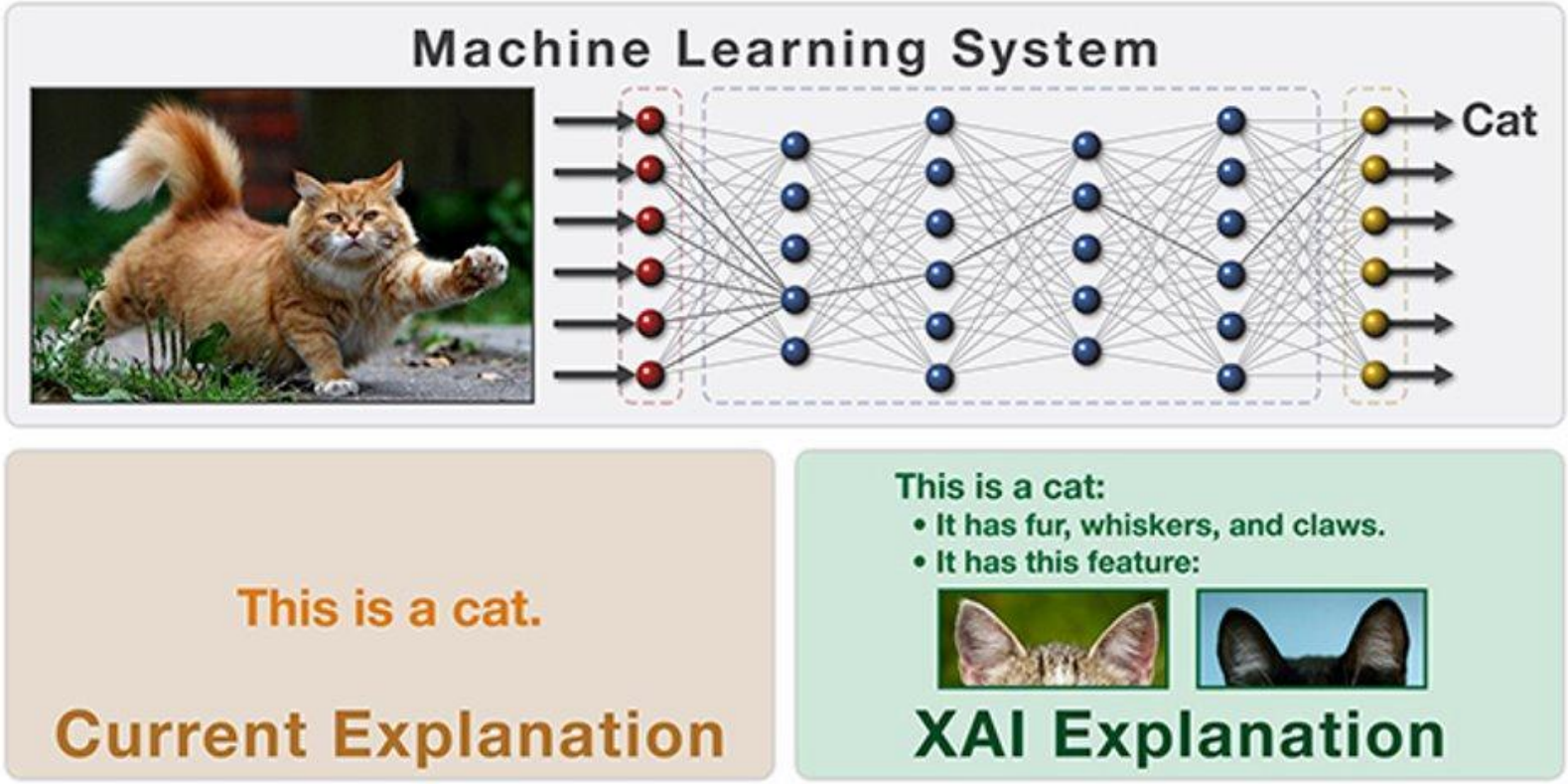


Evaluation of XAI methods



Visualisation credit: XAI tutorial AAAI22

XAI EXPLANATION

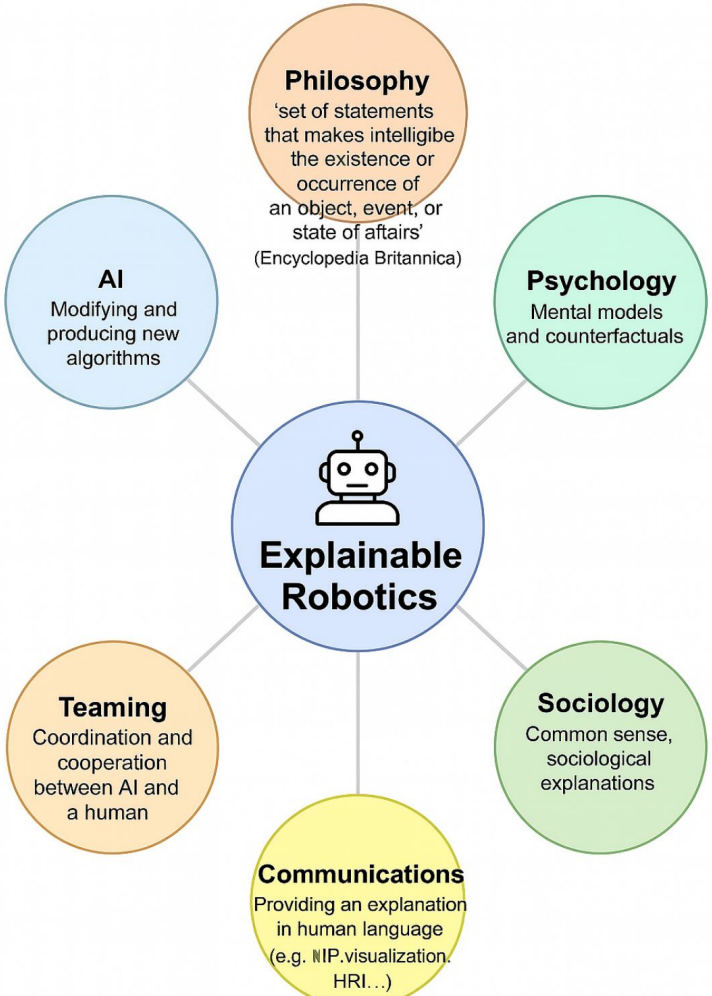


From XAI to Explainable robotics

- Embodied, interactive, real-time agents
- Need for context-aware and multimodal explanations
- Challenges of explainability in dynamic environments

What is Explainable Robotics?

- Making robotic behavior understandable to humans
- Covers **task and motion planning, perception, navigation, control**
- Supports trust, predictability, and team fluency

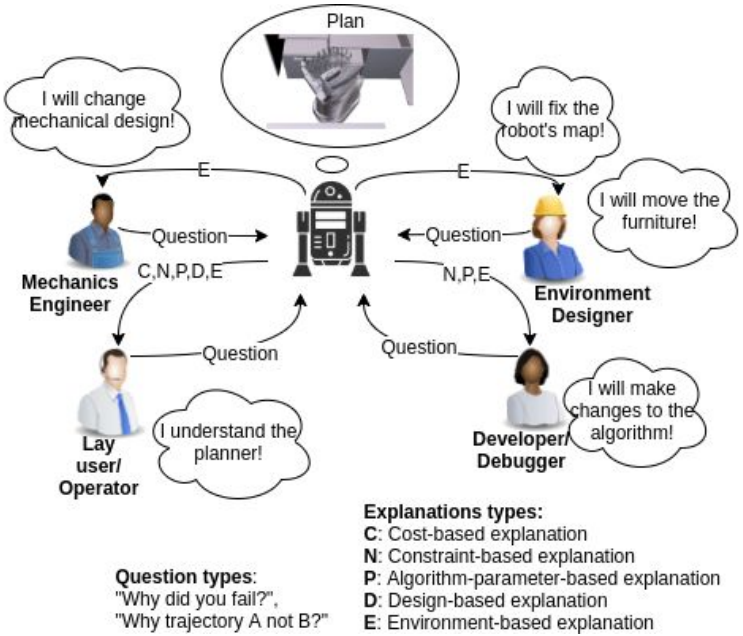


Different User Needs

Lay/End-users: actionable, simple explanations

Experts: detailed trace and confidence

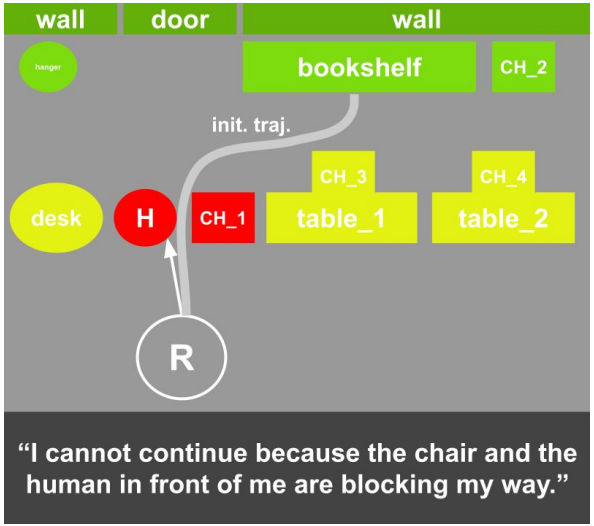
Supervisors: verification and accountability



M Brandao, G Canal, S Krivić, D Magazzeni, ICRA 2021

Explanation Modalities

- Natural language (voice, text)
- Visual overlays
- Simulations
- Symbolic representations

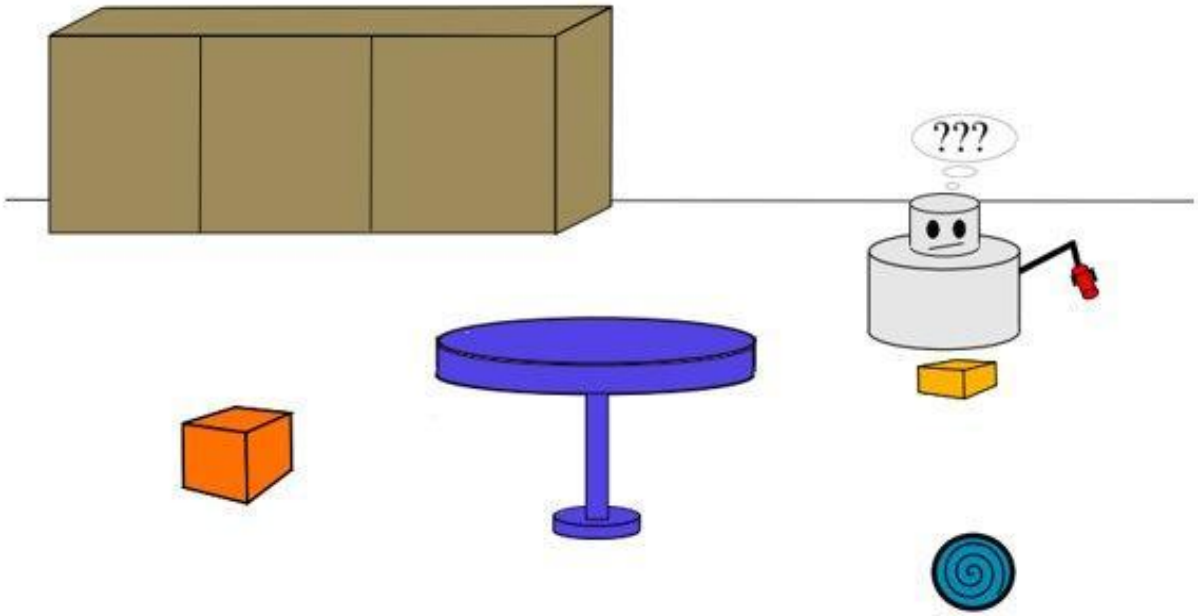


Explanations in Task Planning

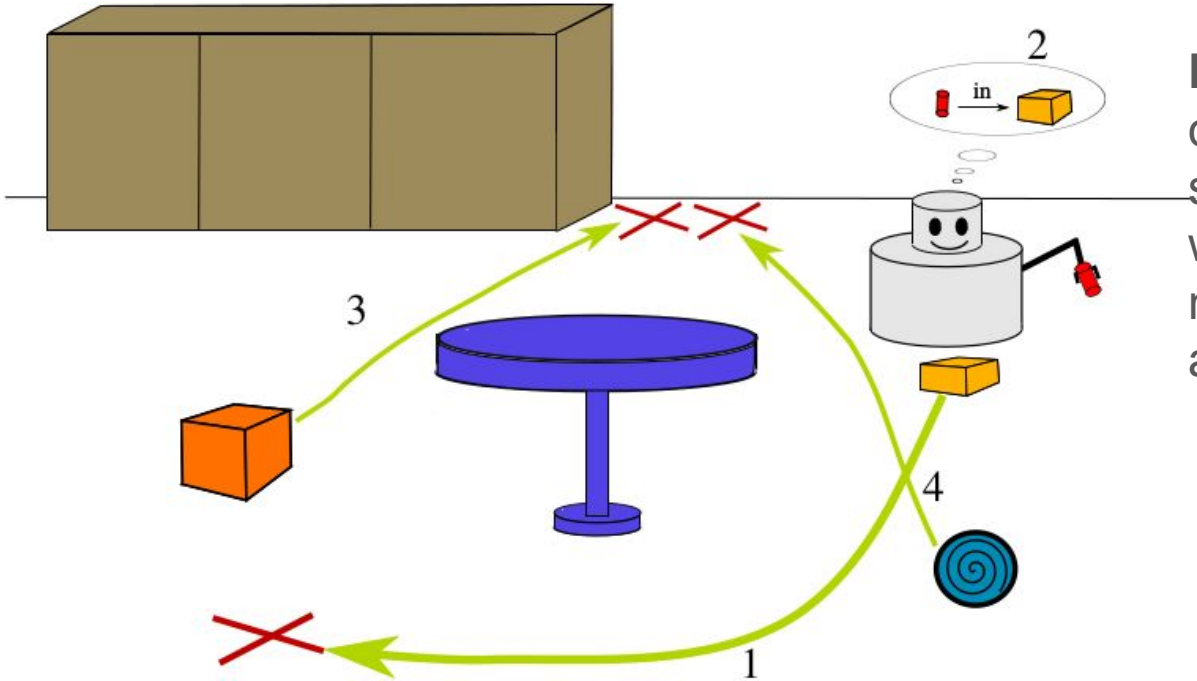
Examples:

- Explain plan steps and goals
- Visualize plan structure and alternatives
- Link symbolic goals to environment

Problems of AI planning in robotics

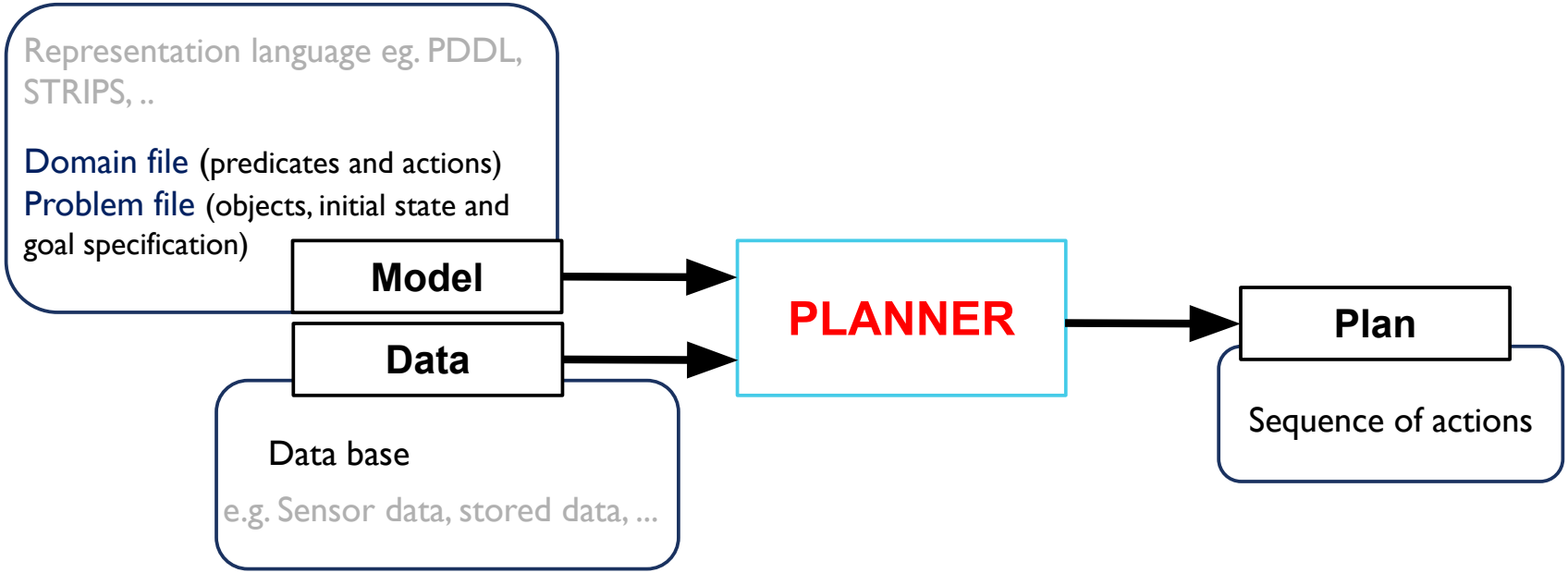


Problems of AI planning in robotics



Planning is a process of determining a sequence of actions π which an agent needs to take to achieve a goal G.

Automated planning and scheduling



Type of planners

Planner Type	Key Feature	Example Planners
State-space search	Forward/backward search	FF, Fast Downward, HSP
Plan-space	Partial-order planning	UCPOP, SNLP
HTN	Task decomposition	SHOP2, Pyhop
SAT-based	Logic encoding of planning	SATPlan, Blackbox
Planning graphs	Graph-based plan extraction	Graphplan, LPG
Heuristic planners	Heuristic-driven search	Fast Downward, FF
Multi-agent planners	Distributed symbolic planning	FMAP, MA-STRIPS

PLANNERS

International Planning Competition

- Zeno
- FF
- MetricFF
- VHPOP
- Marvin
- Crikey
- POPF
- MIPS-XXL

- LPGP
- LPRPG
- CoLin
- Fast Downward
- UPMurphi
- Gamer
- MBP
- ...
- ...

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

(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 you need to replan at this point?
- Q6: Why do you not need to replan at this point?

Explainable Planning, Fox et al., 2017

(Some) Things to be explained

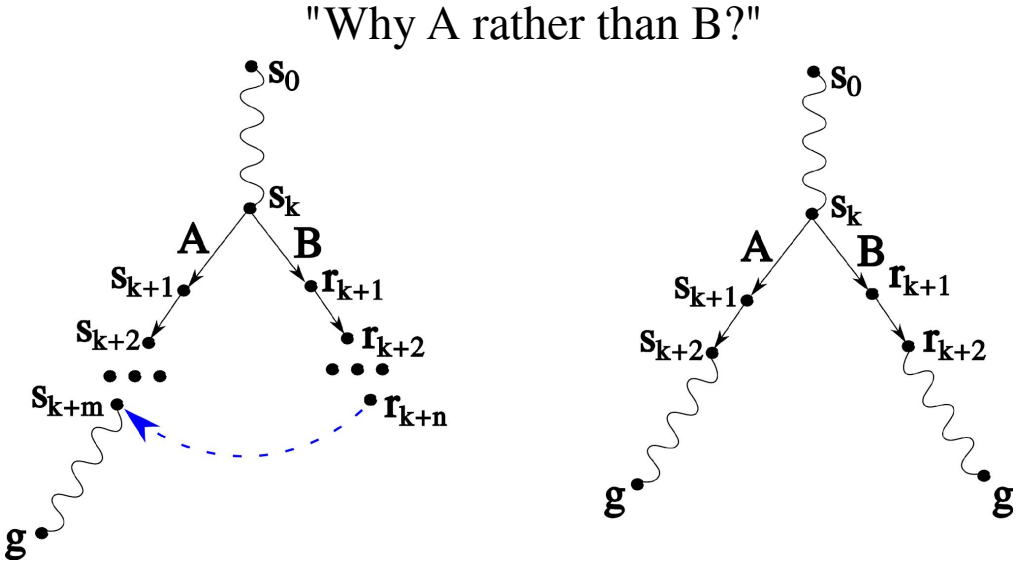
- Q1: Why did you do that?
- Q2: **Why didn’t you do something else? (that I would have done)**
- Q3: Why is what you propose to do more efficient/safe/cheap than something else? (that I would have done)
- Q4: Why can’t you do that?
- Q5: Why do you need to replan at this point?
- Q6: Why do you not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

Explainable Planning, Fox et al., 2017

Why didn’t you do something else?

Providing a contrastive example by taking the suggestion of a user into account

The hypothetical alternative would be a plan that is not better than the one found by the planner or a plan which is better than the original one



s - planning state

g - the goal state

r - alternative state

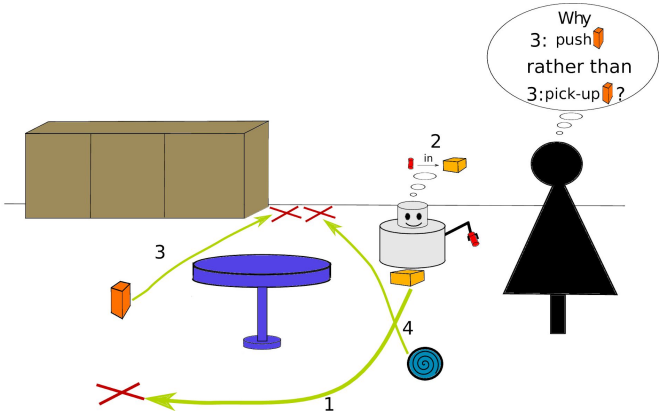
A,B - actions

Contrastive questions

"Why A rather than B?"

- A is the *fact (i.e. what occurred in the plan)*
- B is the *foil* (i.e. the hypothetical alternative expected by the user)

Question
"Why did the plan include action a_A (rather than not including action a_A)?"
"Why did the plan not include action a_A (rather than including action a_A)?"
"Why is action a_A used rather than action a_B?"
"Why is action a_A used before/after action a_B (rather than after/before)?"
"Why is action a_A used at this time (rather than at another time)?"

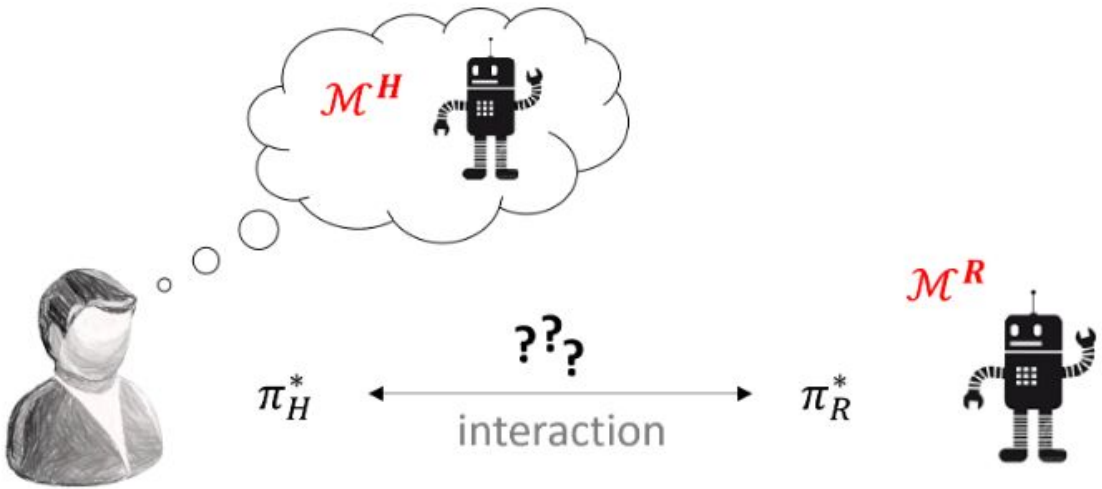


"*Why push the orange block rather than pick up the orange block*?"
Fact - '*push the orange block*'
Foil - 'pick up the orange block'
Comparator - 'rather than'

Interaction between humans and AI systems (robots)

Humans have expectations!

Differing mental models between humans and agents

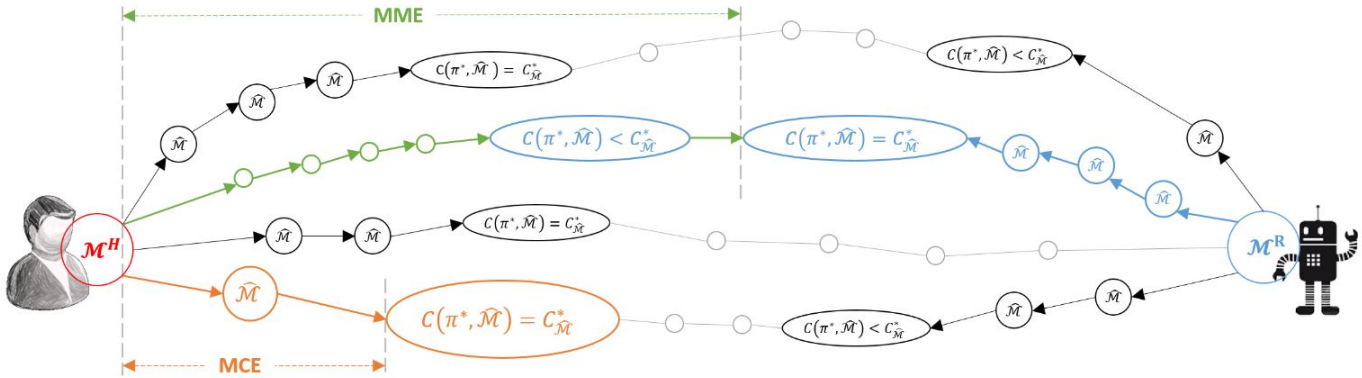


T. Chakraborti Plan Explanations as Model Reconciliation:Moving Beyond Explanation as Soliloquy, 2018

Model reconciliation

Aligning or adjusting a human's mental model by providing explanations that resolve discrepancies

Example explanation: "*I avoided that shorter path because my sensors detected a slippery floor there, which you might not have known*."

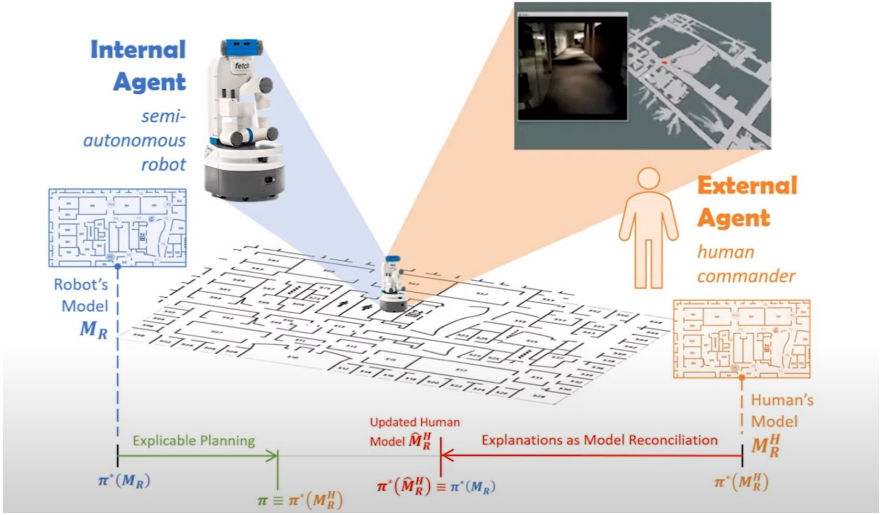


T. Chakraborti Plan Explanations as Model Reconciliation:Moving Beyond Explanation as Soliloquy, 2018

Model reconciliation

Aligning or adjusting a human's mental model by providing explanations that resolve discrepancies

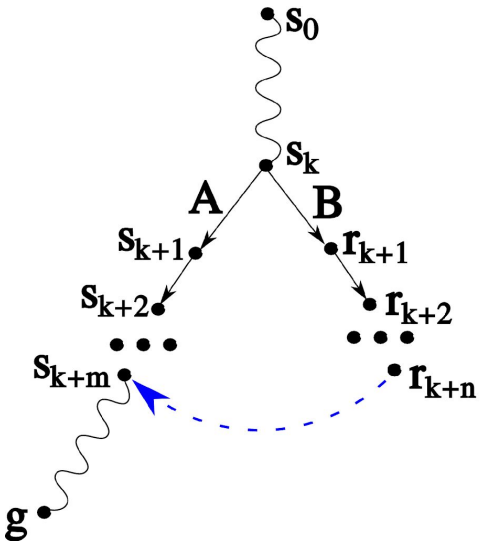
Example explanation: "*I avoided that shorter path because my sensors detected a slippery floor there, which you might not have known*."

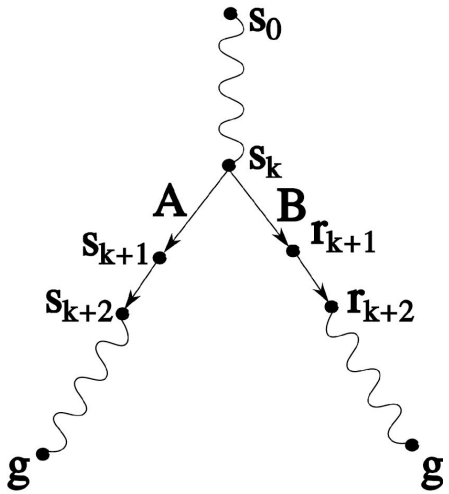


Kambhampati, Subbarao. "Challenges of Human-Aware AI Systems"

Generating contrastive explanation (CE)

Providing a contrastive example by taking suggestion of a user into account.

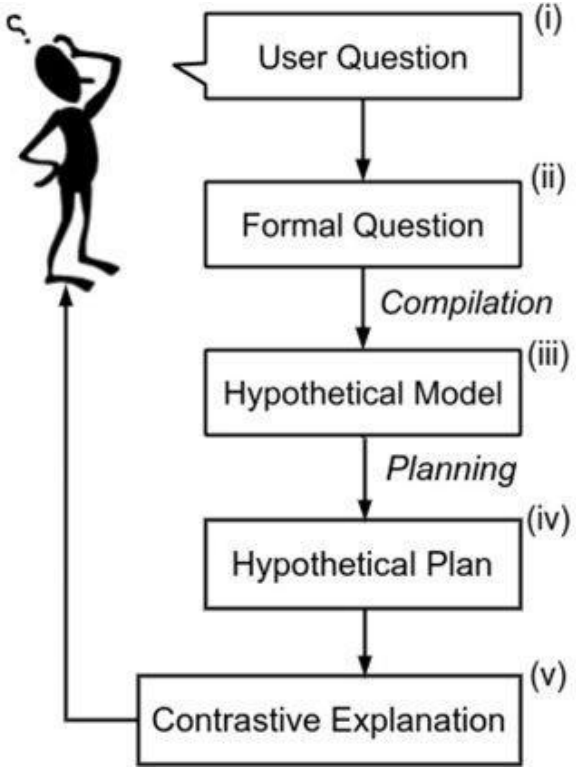




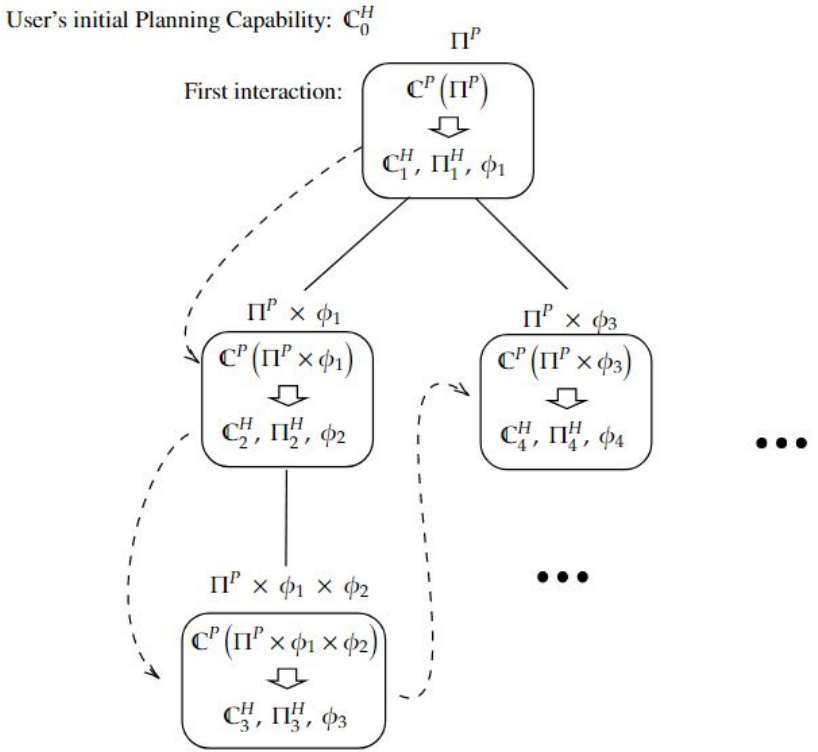
- s - planning state
- g - the goal state
- r - alternative state
- A,B - actions

Iterative process to explainable planning (XAIP)

Forming contrastive explanations in iterative way through model restrictions



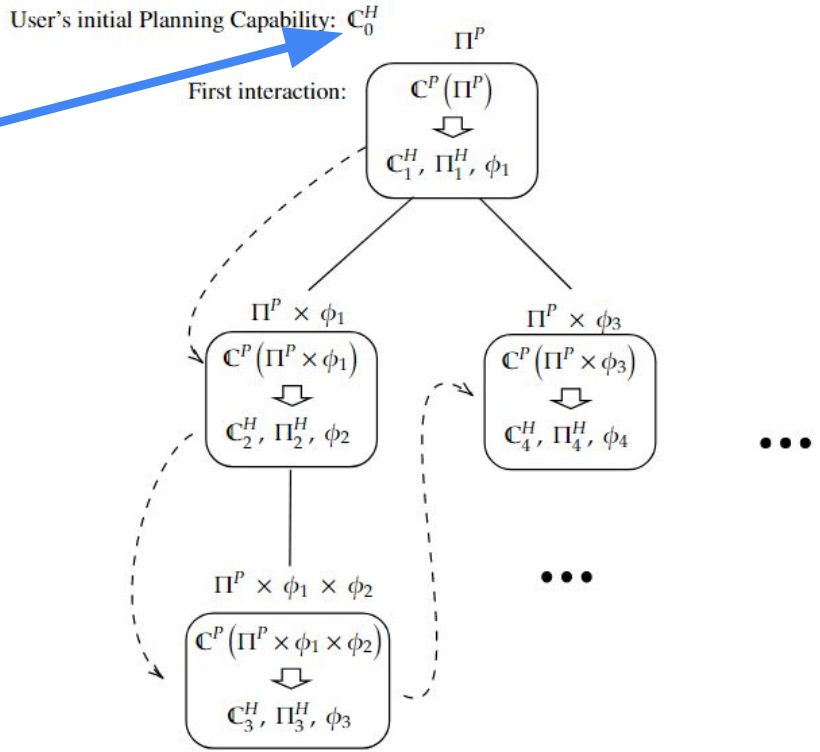
Sequence of interactions between user and planner



Contrastive Explanations of Plans Through Model Restrictions (JAIR) - Krarup et al.

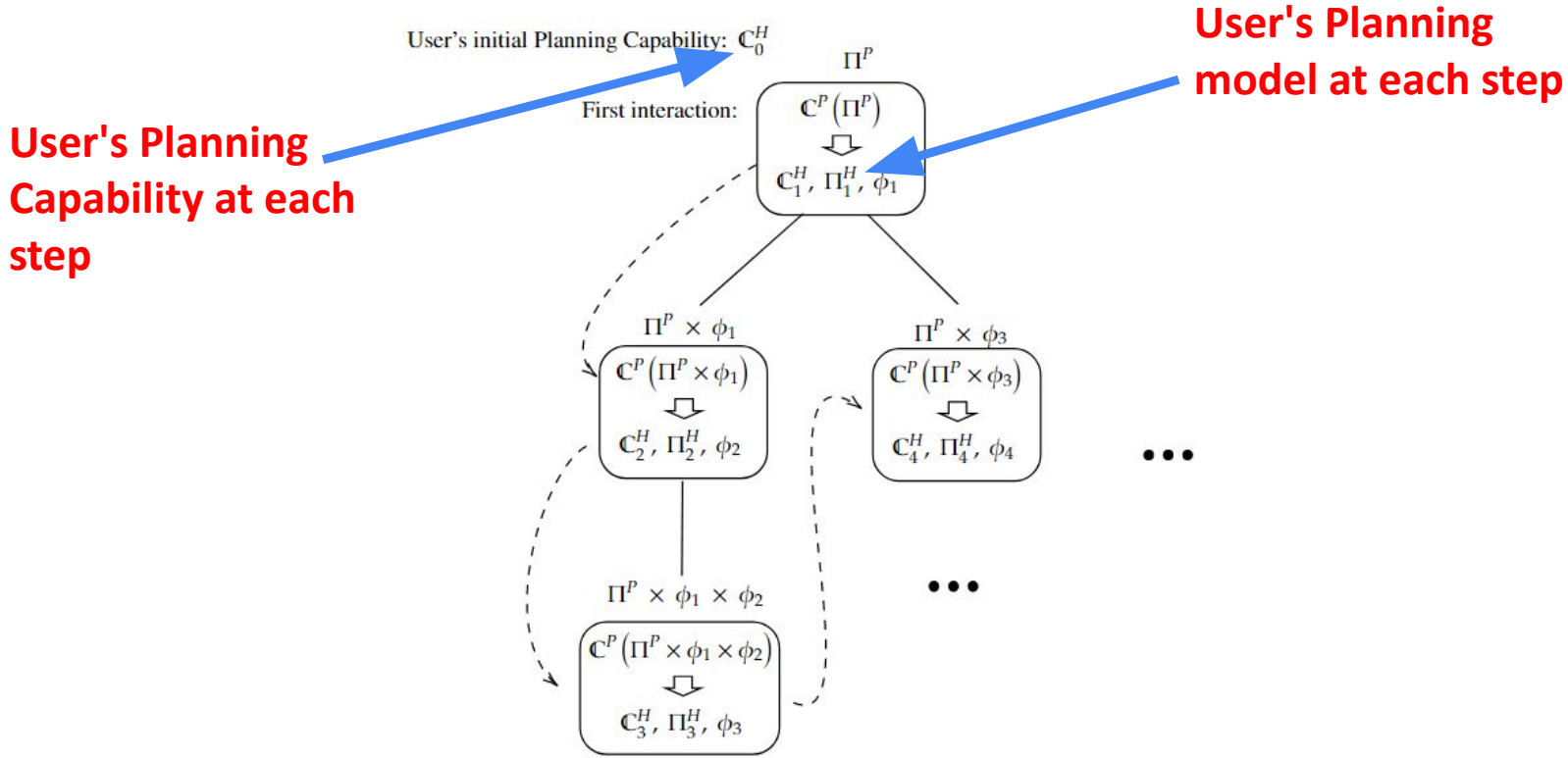
Sequence of interactions between user and planner

User's Planning Capability at each step



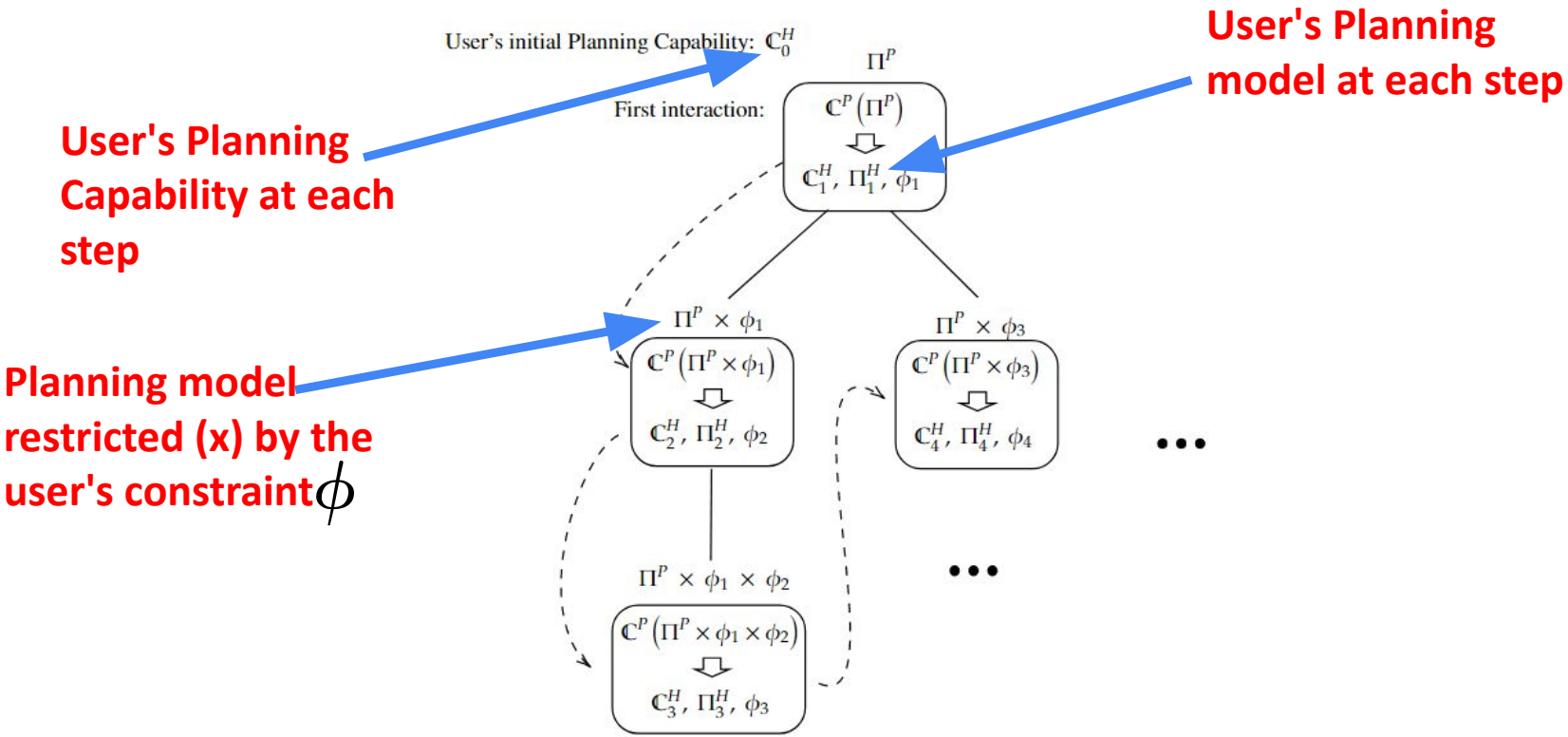
Contrastive Explanations of Plans Through Model Restrictions (JAIR) - Krarup et al.

Sequence of interactions between user and planner



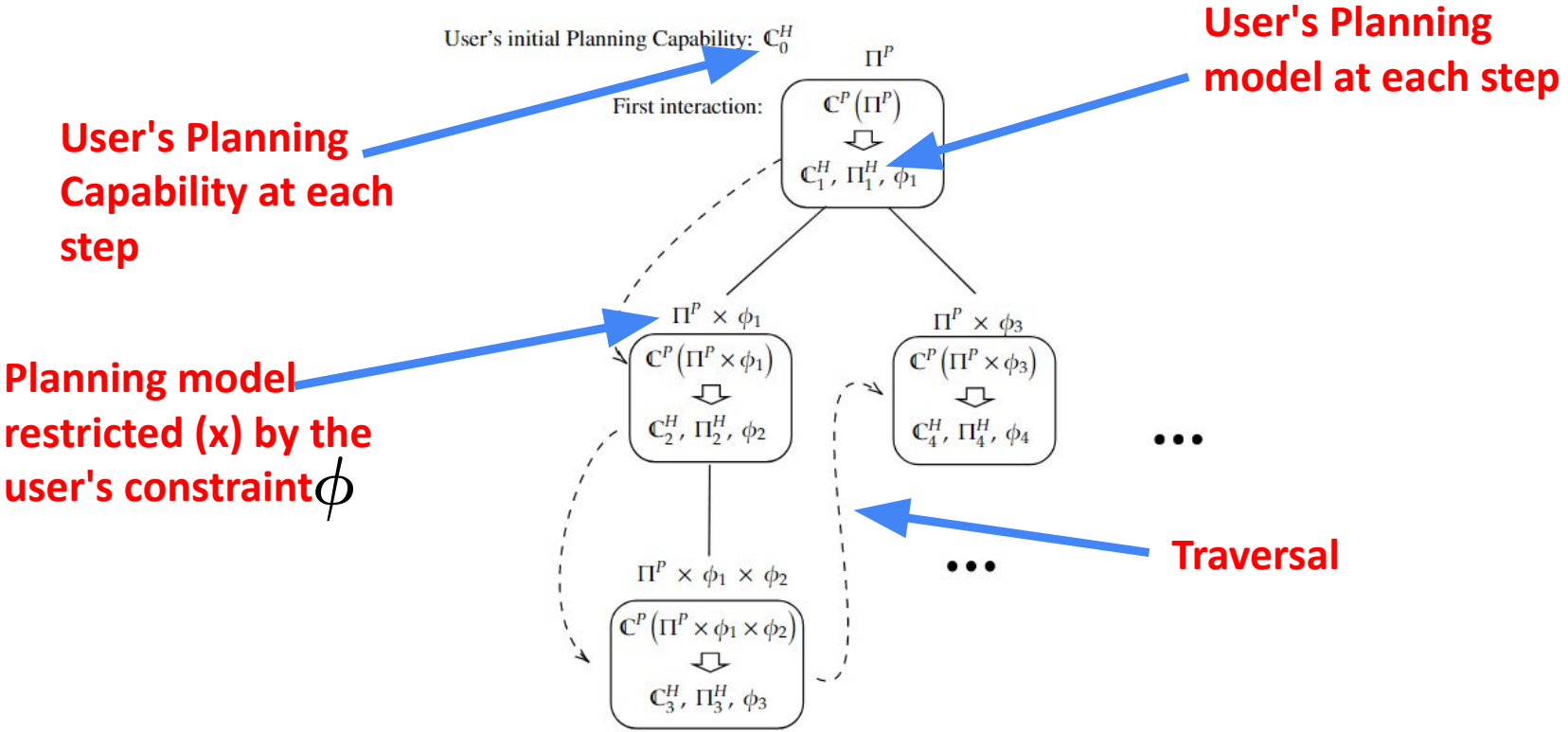
Contrastive Explanations of Plans Through Model Restrictions (JAIR) - Krarup et al.

Sequence of interactions between user and planner



Contrastive Explanations of Plans Through Model Restrictions (JAIR) - Krarup et al.

Sequence of interactions between user and planner



Contrastive Explanations of Plans Through Model Restrictions (JAIR) - Krarup et al.

Composition of Compilations

- Can compose multiple constraints to produce more complex constraints
- q1, q2, …, qn -> $\phi_1, \phi_2, \ldots, \phi_n$

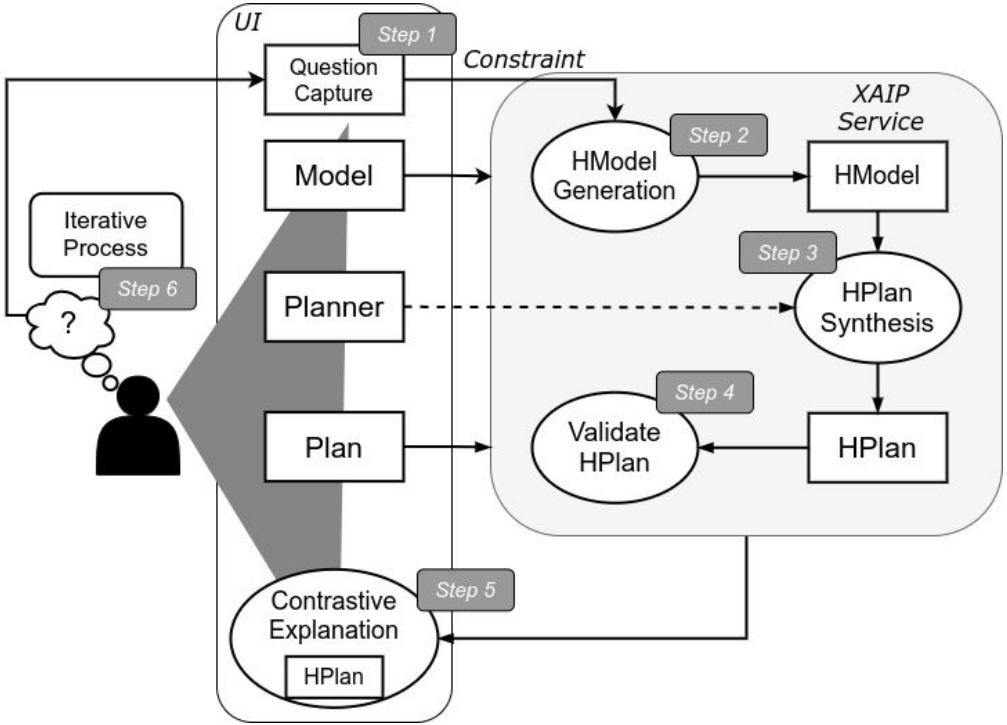
$$\Pi \times \phi_x = ((\Pi \times \phi_1) \times \phi_2) \ldots \times \phi_n$$

- qx -> ϕ_x

Composition of Compilations

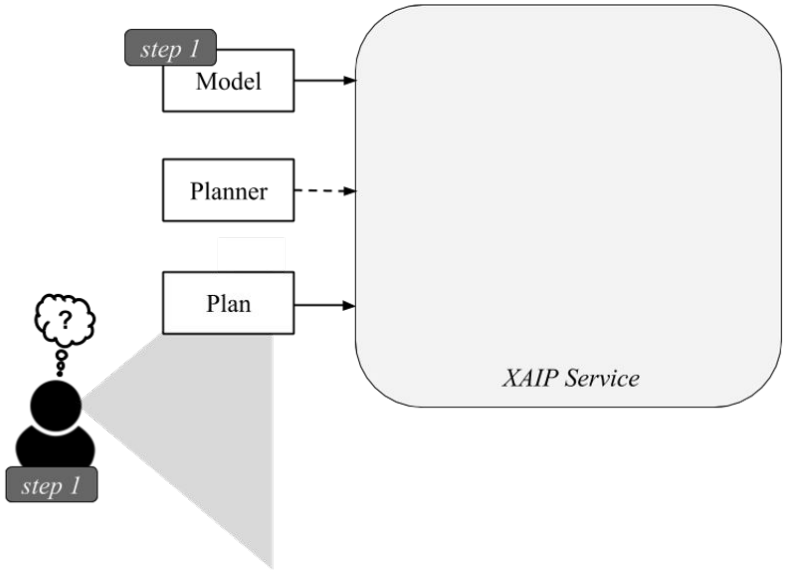
- qx can be hard to form
- Can be easier to form qx through intermediary questions q1, …, qn
- User study shows that by using a variety of questions, users converged quickly on desired plans

System Framework



XAIP as a service

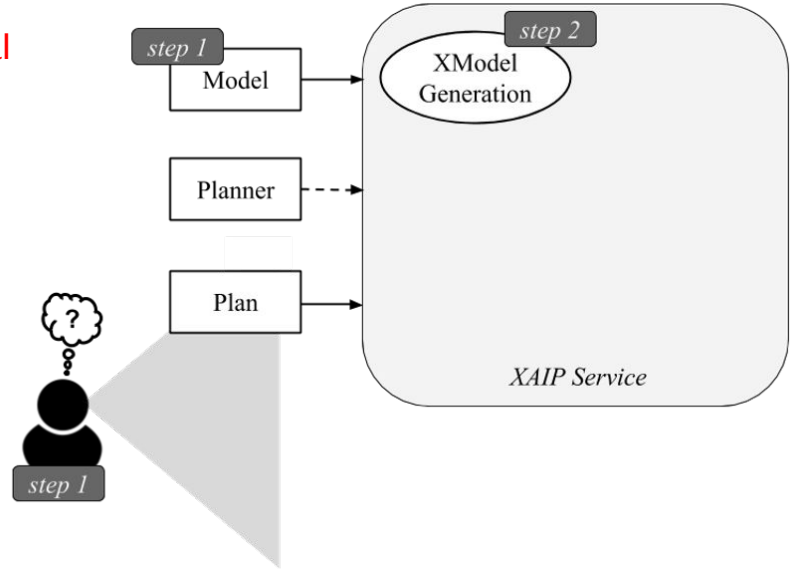
Step 1: Questioning the plan



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

XAIP as a service

Step 2: Deriving the hypothetical model



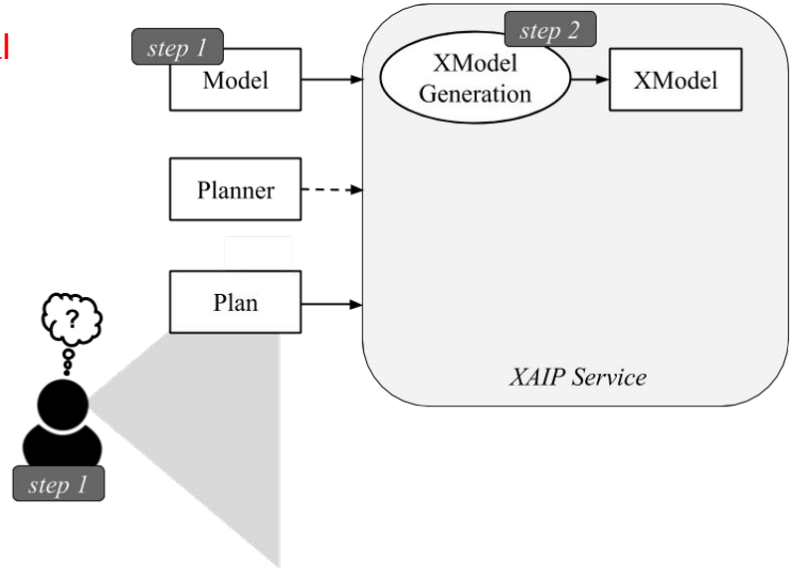
Compilation process:
Adding constraints in the domain and problem descriptions

Compilation type
RemoveAction()
AddAction()
ReplaceAction()
ReorderAction()
RescheduleAction()

The question is translated into constraints

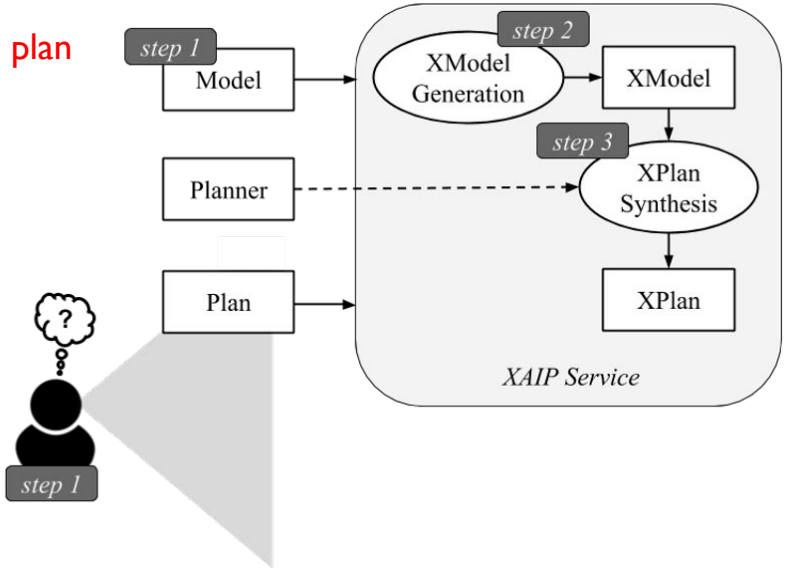
XAIP as a service

Step 2: Deriving the hypothetical model



XAIP as a service

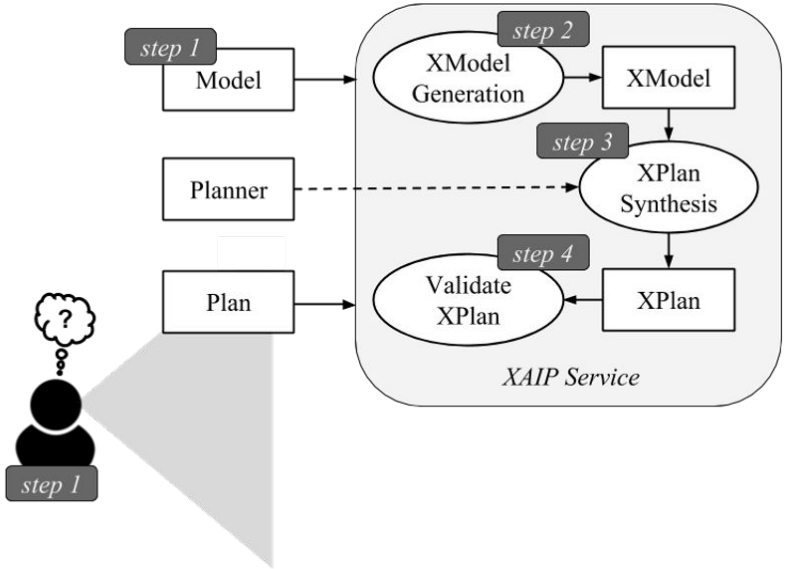
Step 3: Producing the alternative plan (XPlan)



The original planner must be used

XAIP as a service

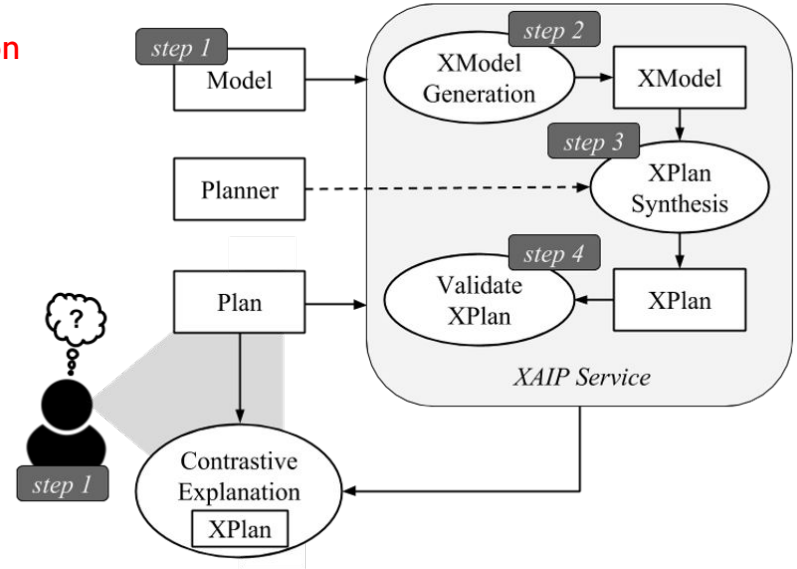
Step 4: Validation of the XPlan



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

XAIP as a service

Forming Contrastive Explanation

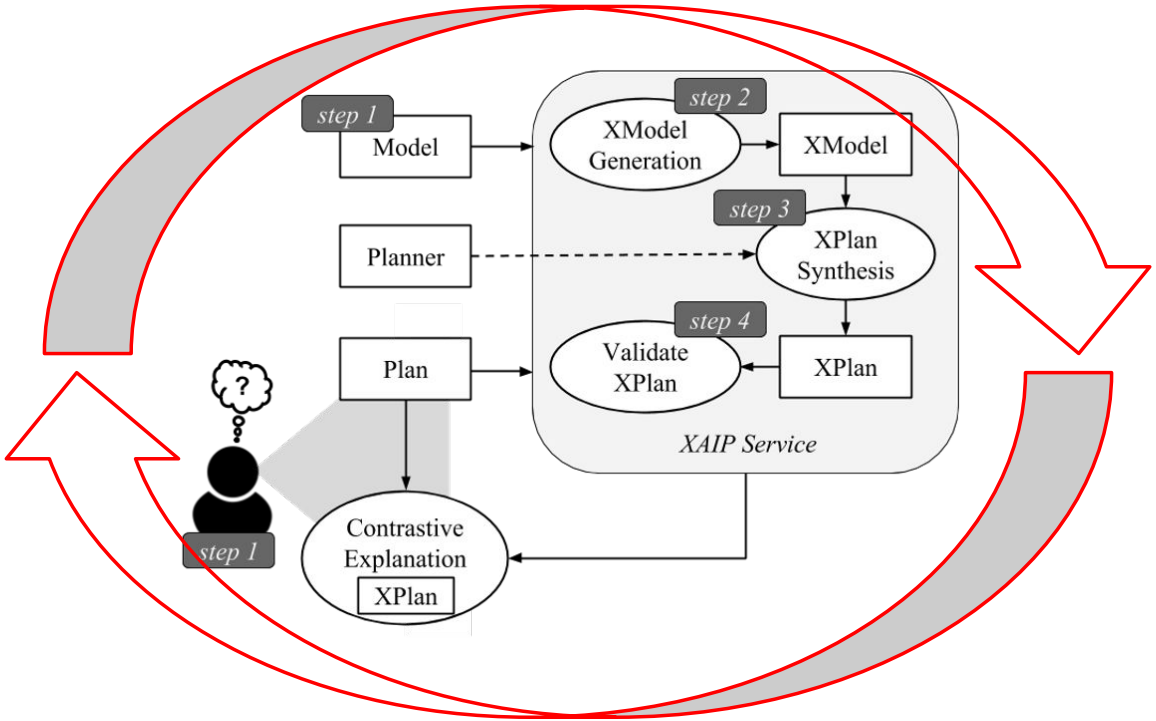


$$CE = \langle comparison, report, Q \rangle$$

$$compare(\pi, \pi_x) = \langle existing, removed, added, c_{diff} \rangle$$
$$existing = \pi \cap \pi_x$$
$$removed = \pi \setminus \pi \cap \pi_x$$
$$added = \pi_x \setminus \pi \cap \pi_x$$
$$c_{diff} = c(\pi) - c(\pi_x)$$

XAIP as a service

Iterative Process



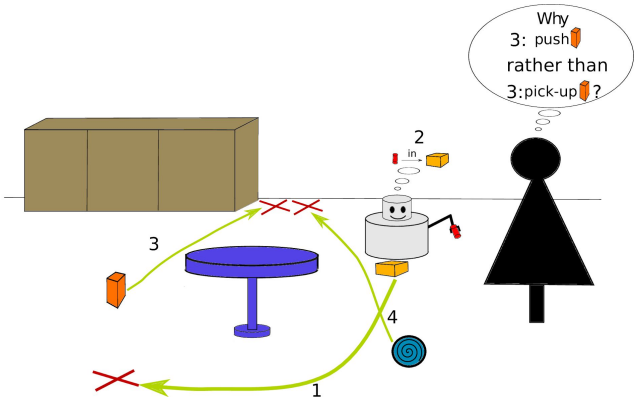
XAIP as a service

Original plan

```
0: (pick-up robot red-cylinder) [0.001]
1: (push robot yellow-box wp1) [0.001]
2: (put-in robot wp1 red-cylinder yellow-box) [0.001]
3: (goto robot wp1 orange-box-loc) [0.001]
4: (push robot orange-box wp3) [0.001]
5: (goto robot wp3 ball-push-loc) [0.001]
6: (push robot red-ball wp2) [0.001]
plan cost: 0.007
```

XPlan

```
0: (pick-up robot red-cylinder) [0.001]
1: (push robot yellow-box wp1) [0.001]
2: (put-in robot wp1 red-cylinder yellow-box) [0.001]
3: (goto robot wp1 orange-box-loc) [0.001]
4: (pick-up robot orange-box ) [0.001]
5: (goto robot orange-box-loc wp3) [0.001]
6: (place-at robot wp3 orange-box) [0.001]
7: (goto robot wp3 ball-push-loc) [0.001]
8: (push robot red-ball wp2) [0.001]
plan cost: 0.009
```



Contrastive explanation

$$existing = \{\{a_0, a_1, a_2, a_3\}, \{a_5, a_6\}\}$$
$$removed = \{a_4\}$$
$$added = \{a_4^x, a_5^x, a_6^x\}$$
$$c_{diff} = 0.002$$

Example

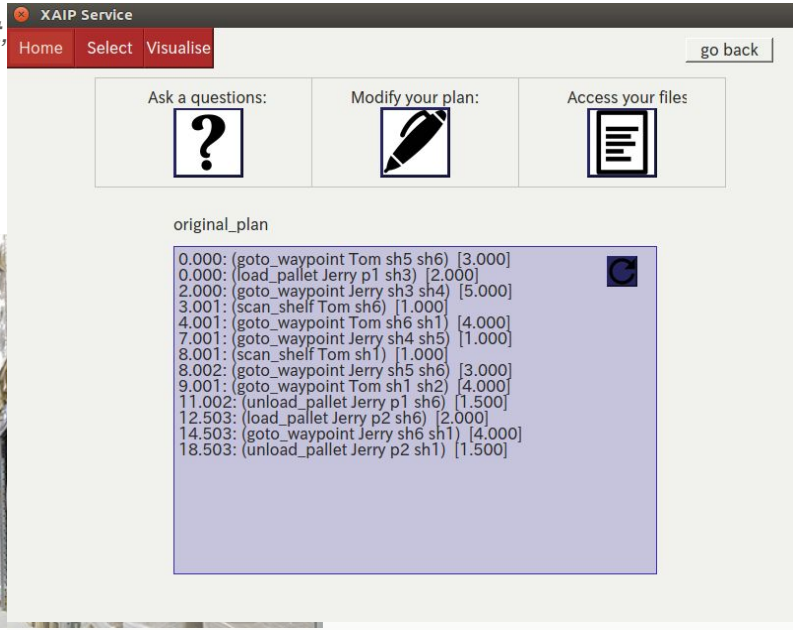
- **Warehouse organisation delivery system**
- **Domain**
  - **Types:** *pallet, robot, waypoint*
  - **Temporal actions:** *goto_waypoint, scan_shelf, load_pallet, unload_pallet*
- **Problem**
  - *2 robots (Tom and Jerry)*
  - *2 pallets (p1, p2)*
  - *6 waypoints (shelfs - sh1, sh2, sh3, sh4, sh5, sh6)*

www.dsv.com

Example

- **Warehouse organisation delivery system**
- **Domain**
  - **Types:** *pallet, robot, waypoint*
  - **Temporal actions:** *goto_waypoint, scan_shelf, load_pallet,*
- **Problem**
  - *2 robots (Tom and Jerry)*
  - *2 pallets (p1, p2)*
  - *6 waypoints (shelfs - sh1, sh2, sh3, sh4, sh5, sh6)*

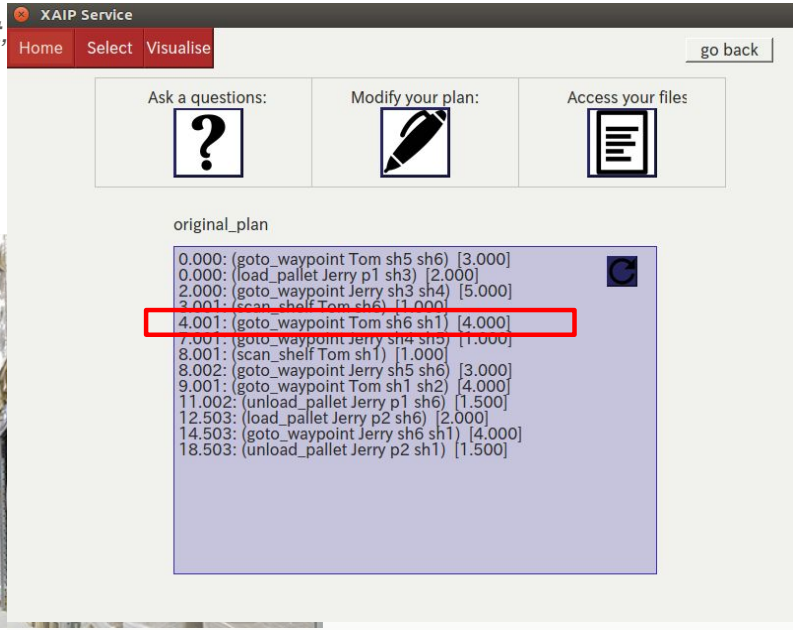
www.dsv.com



Example

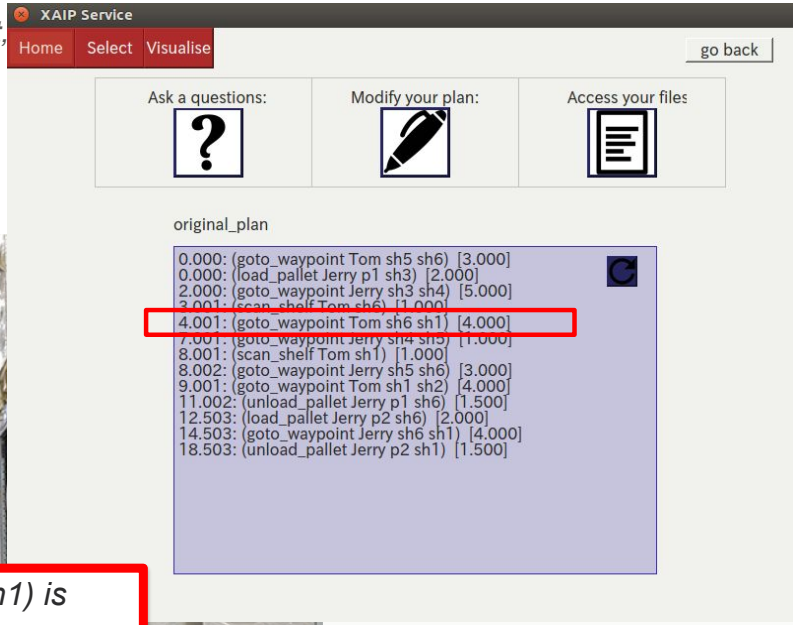
- **Warehouse organisation delivery system**
- **Domain**
  - **Types:** *pallet, robot, waypoint*
  - **Temporal actions:** *goto_waypoint, scan_shelf, load_pallet,*
- **Problem**
  - *2 robots (Tom and Jerry)*
  - *2 pallets (p1, p2)*
  - *6 waypoints (shelfs - sh1, sh2, sh3, sh4, sh5, sh6)*

www.dsv.com



Example

- **Warehouse organisation delivery system**
- **Domain**
  - **Types:** *pallet, robot, waypoint*
  - **Temporal actions:** *goto_waypoint, scan_shelf, load_pallet,*
- **Problem**
  - *2 robots (Tom and Jerry)*
  - *2 pallets (p1, p2)*
  - *6 waypoints (shelfs - sh1, sh2, sh3, sh4, sh5, sh6)*

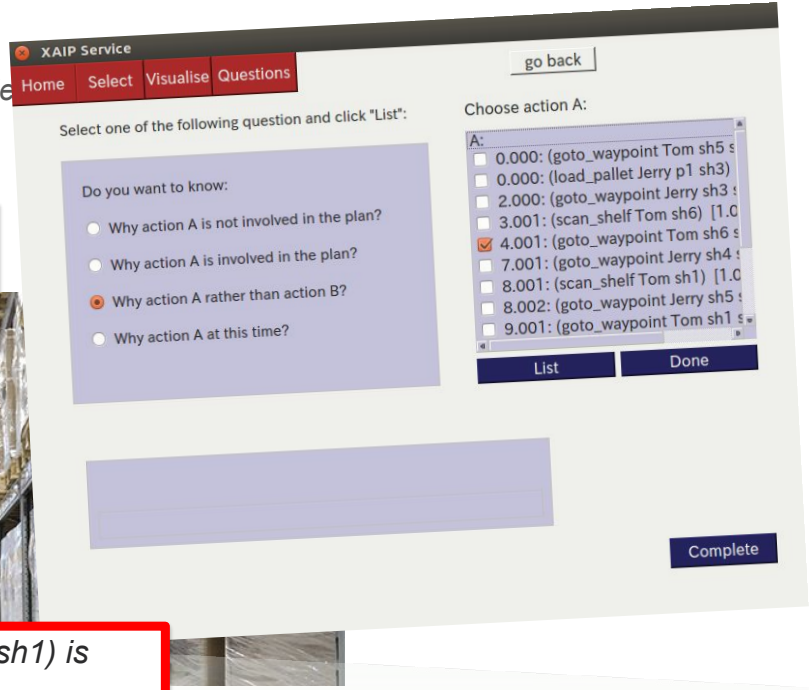


At the point in the plan where action (goto_waypoint Tom sh6 sh1) is used, there's a pallet, so why doesn't Tom pick it up?

Example

- **Warehouse organisation delivery system**
- **Domain**
  - **Types:** *pallet, robot, waypoint*
  - **Temporal actions:** *goto_waypoint, scan_shelf, load_palle*
- **Problem**
  - *2 robots (Tom and Jerry)*

Formal question:
Why is the action A used in state S, rather than action B?



At the point in the plan where action (goto_waypoint Tom sh6 sh1) is used, there's a pallet, so why doesn't Tom pick it up?

Example

- **Warehouse organisation delivery system**
- **Domain**
  - **Types:** *pallet, robot, waypoint*
  - **Temporal actions:** *goto_waypoint, scan_shelf, load_palle*
- **Problem**
  - *2 robots (Tom and Jerry)*

Formal question:
Why is the action A used in state S, rather than action B?

Constructing the HModel
- executing action B at state S
- new initial state I' in the HModel (effects of each happening are used up to the replacement action)

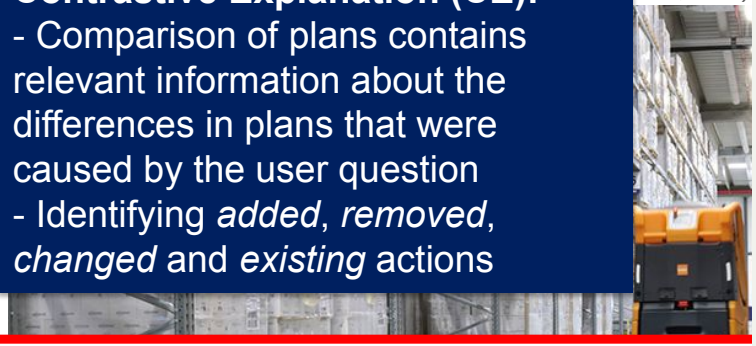
At the point in the plan where action (goto_waypoint Tom sh6 sh1) is used, there's a pallet, so why doesn't Tom pick it up?

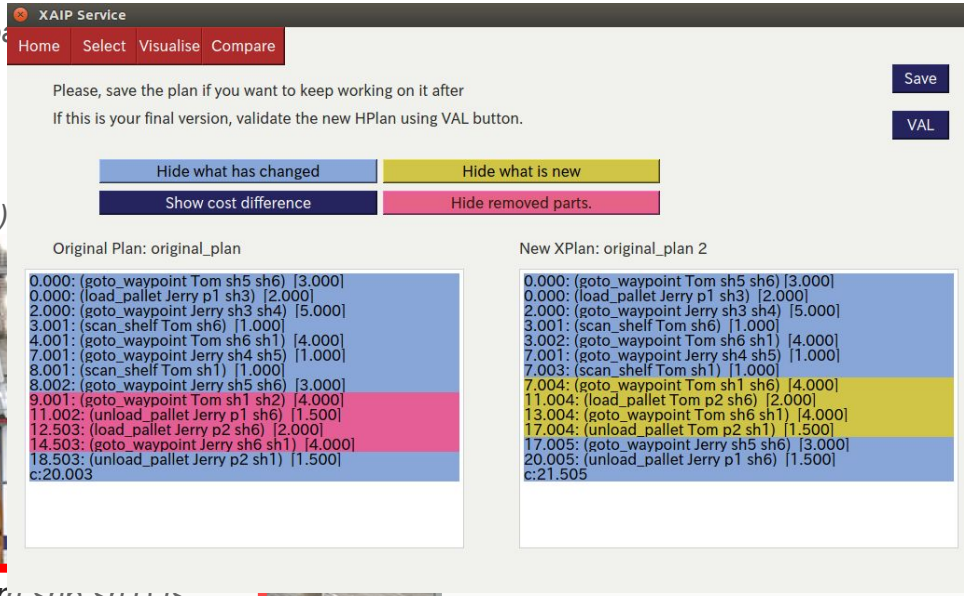
Example

- **Warehouse organisation delivery system**
- **Domain**
  - **Types:** *pallet, robot, waypoint*
  - **Temporal actions:** *goto_waypoint, scan_shelf, lo*
- **Problem**
  - *2 robots (Tom and Jerry)*
  - *2 pallets (p1, p2)*

Contrastive Explanation (CE):
- Comparison of plans contains relevant information about the differences in plans that were caused by the user question
- Identifying *added*, *removed*, *changed* and *existing* actions

sh5, sh6)

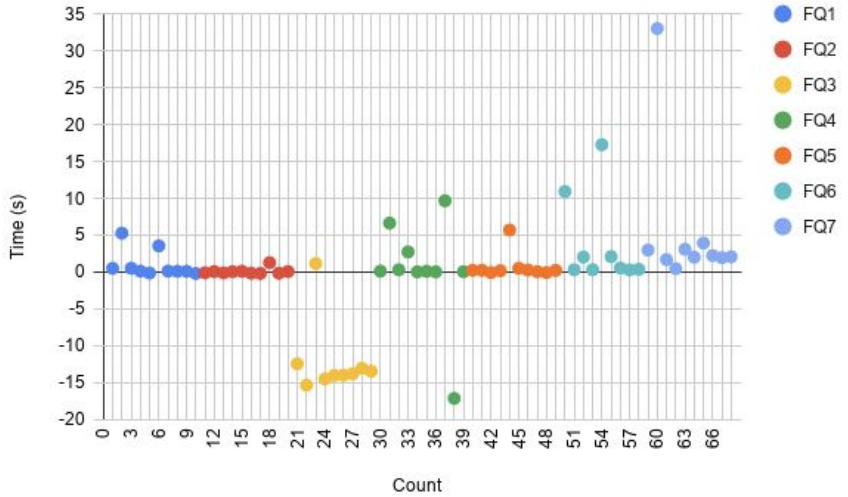




At the point in the plan where action (goto_waypoint Tom sh6 sh1) is used, there's a pallet, so why doesn't Tom pick it up?

Evaluation: performance

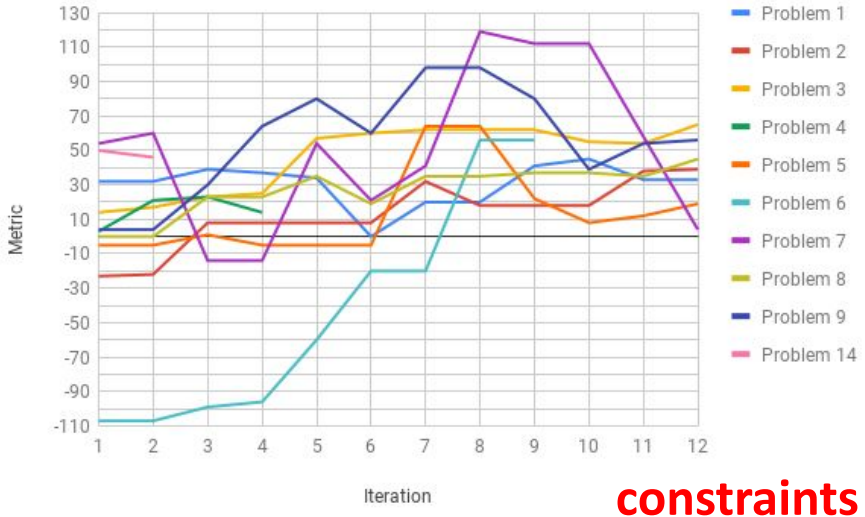
- Planner Performance
  - constraints had no systematic negative impact on planning time



Crew Planning #20
270 literals
2880 time units

Evaluation: quality

- Plan Quality
  - constraints often improved plan quality!



constraints helped focus search towards better plans

Evaluation: satisfaction

User Study

- 20 participants
  - group 1: full system
  - group 2: no highlighting
- Time
- # of questions
- Explanation Satisfaction scale
  - helpful
  - increased understanding
  - deeper comparison

ROADMAP

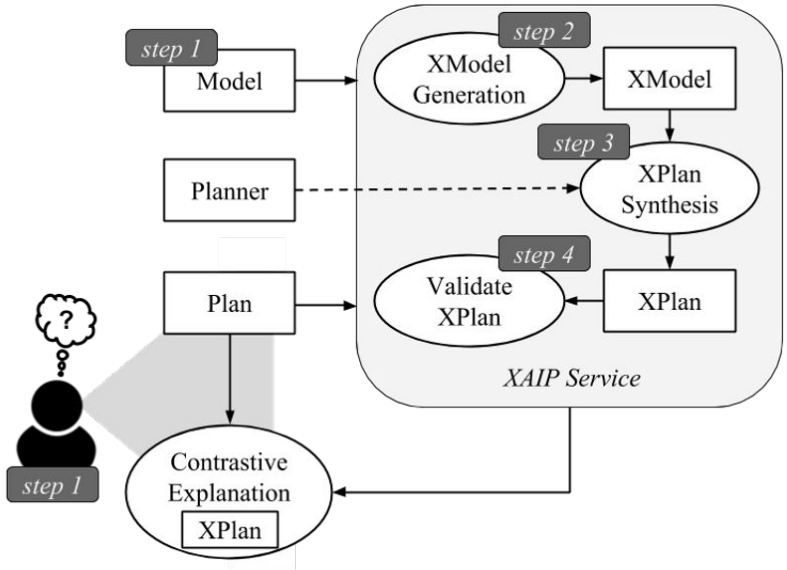
How the user question can be understood, properly taking into account the context in which it was asked?

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

How constraints can be formally encoded in the XModel?

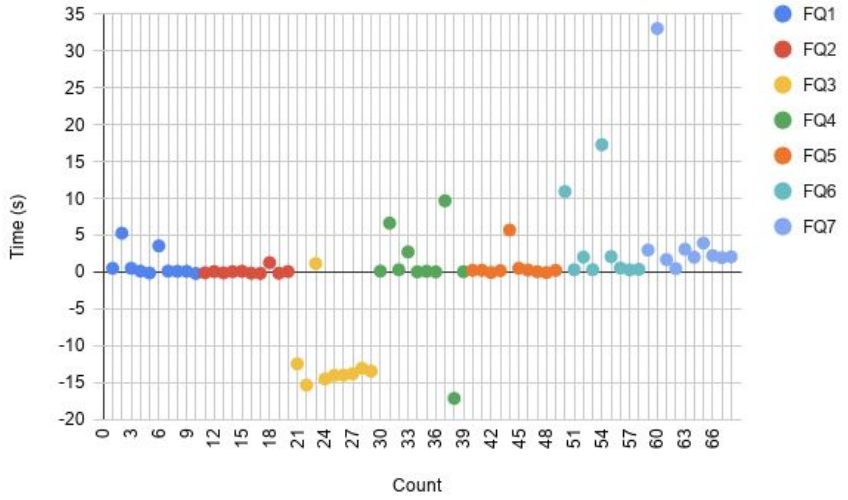
How to present explanations to the users?

How to assess the effectiveness of the provided explanations?



Evaluation: performance

- Planner Performance
  - constraints had no systematic negative impact on planning time

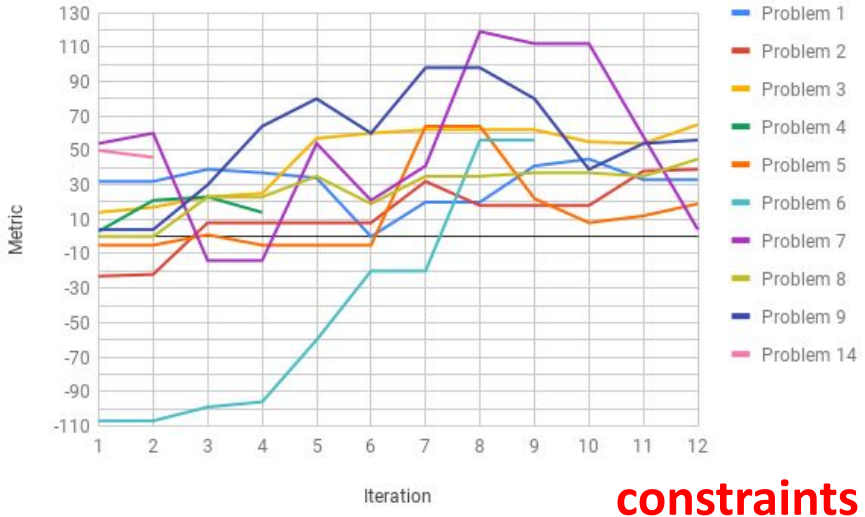


Crew Planning #20
270 literals
2880 time units

Contrastive Explanations of Plans Through Model Restrictions (JAIR) - Krarup et al.

Evaluation: quality

- Plan Quality
  - constraints often improved plan quality!



constraints helped focus search towards better plans

Contrastive Explanations of Plans Through Model Restrictions (JAIR) - Krarup et al.

Evaluation: satisfaction

User Study

- 20 participants
  - group 1: full system
  - group 2: no highlighting
- Time
- # of questions
- Explanation Satisfaction scale
  - helpful
  - increased understanding
  - deeper comparison

Contrastive Explanations of Plans Through Model Restrictions (JAIR) - Krarup et al.

What if our robot has an ethical dilemma?

The robot wants to motivate Frank so that he will do some exercise and keep healthy.

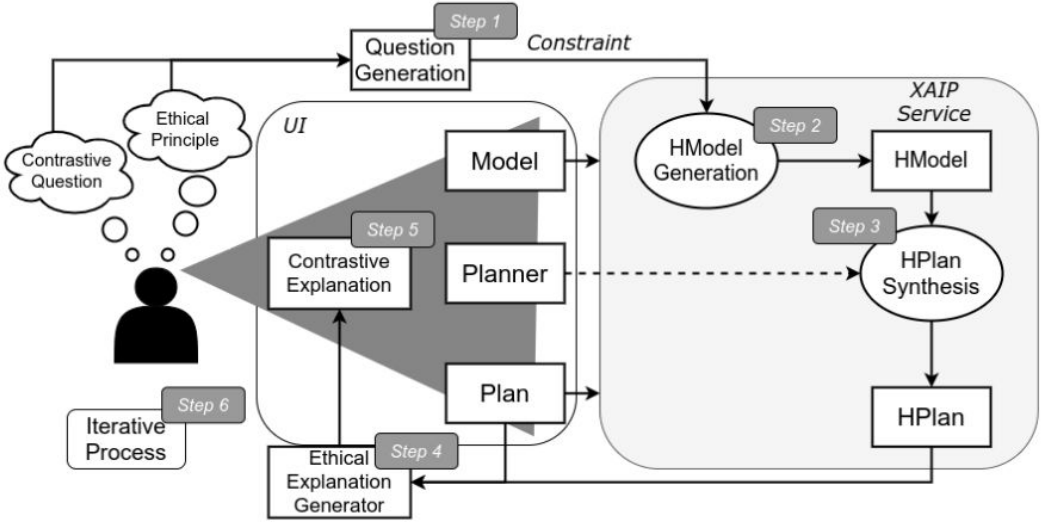
The robot has two choices, it can either beg Frank to exercise, or it can lie and deceive him.

Robot and Frank, 2012

Explainable and Ethical AI and Robotics

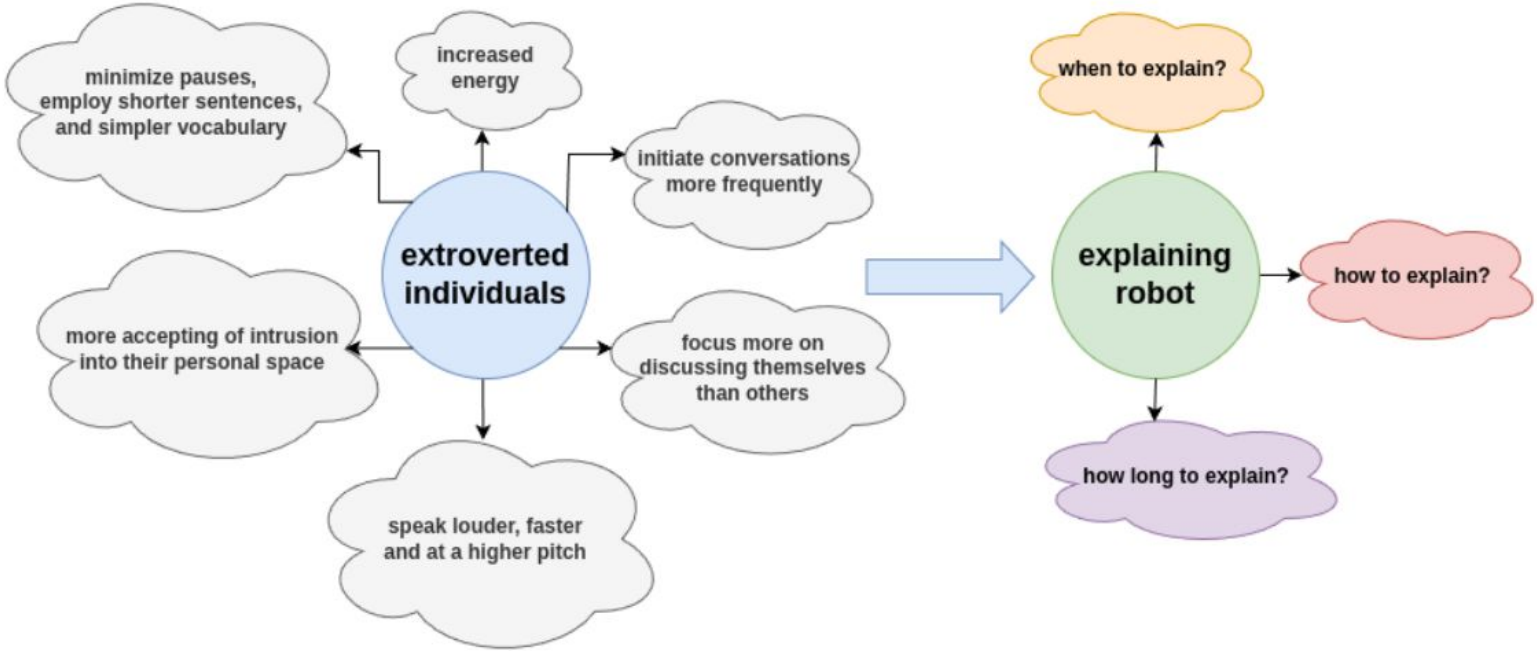
Why a particular plan is permissible or not?

Explanations generated by the system can improve humans' understanding of a robot's ethical principle.



Benjamin Krarup, Felix Lindner,, Senka Krivic, and Derek Long, CASE 22

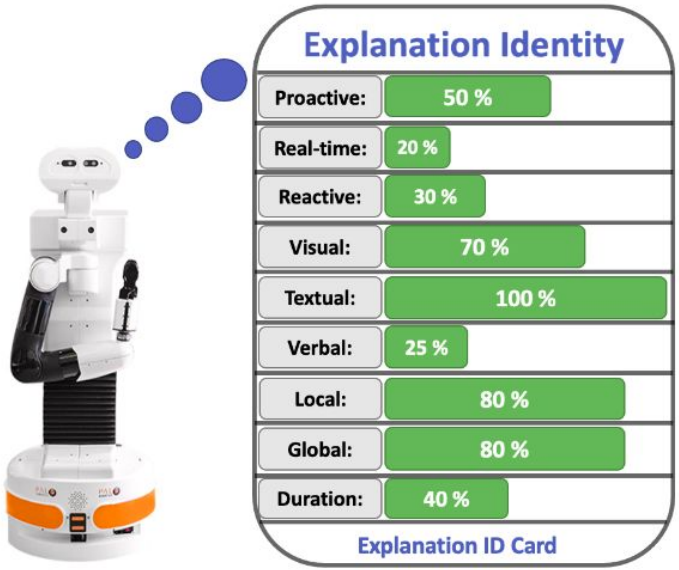
From human to robot personality



The Influence of a Robot's Personality on
Real-Time Explanations of Its Navigation, Halilovic & Krivic, ICSR 2023

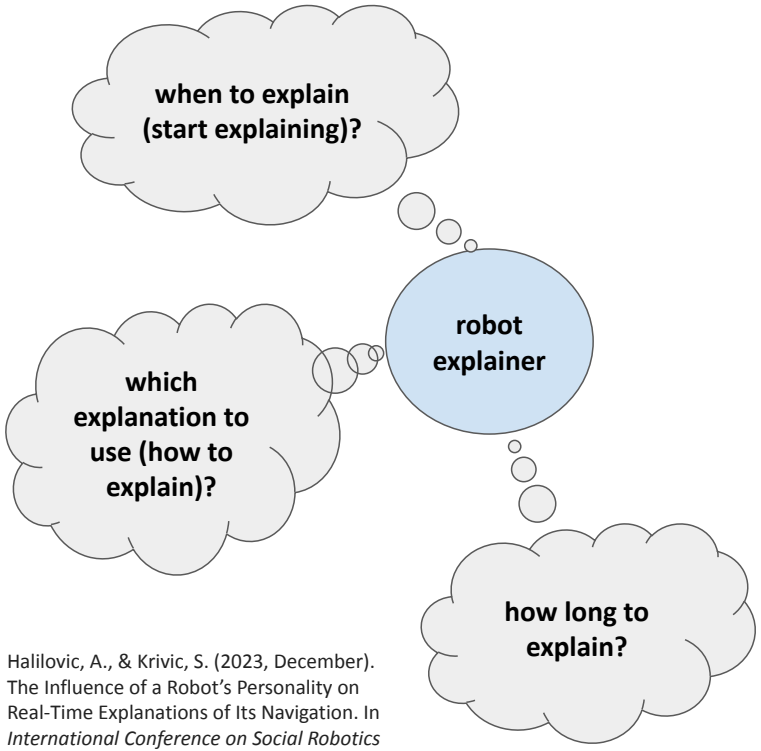
Robot explanation identity

- Unique identity characteristics
- Multi-faceted
- Fluidity
- Parametrized



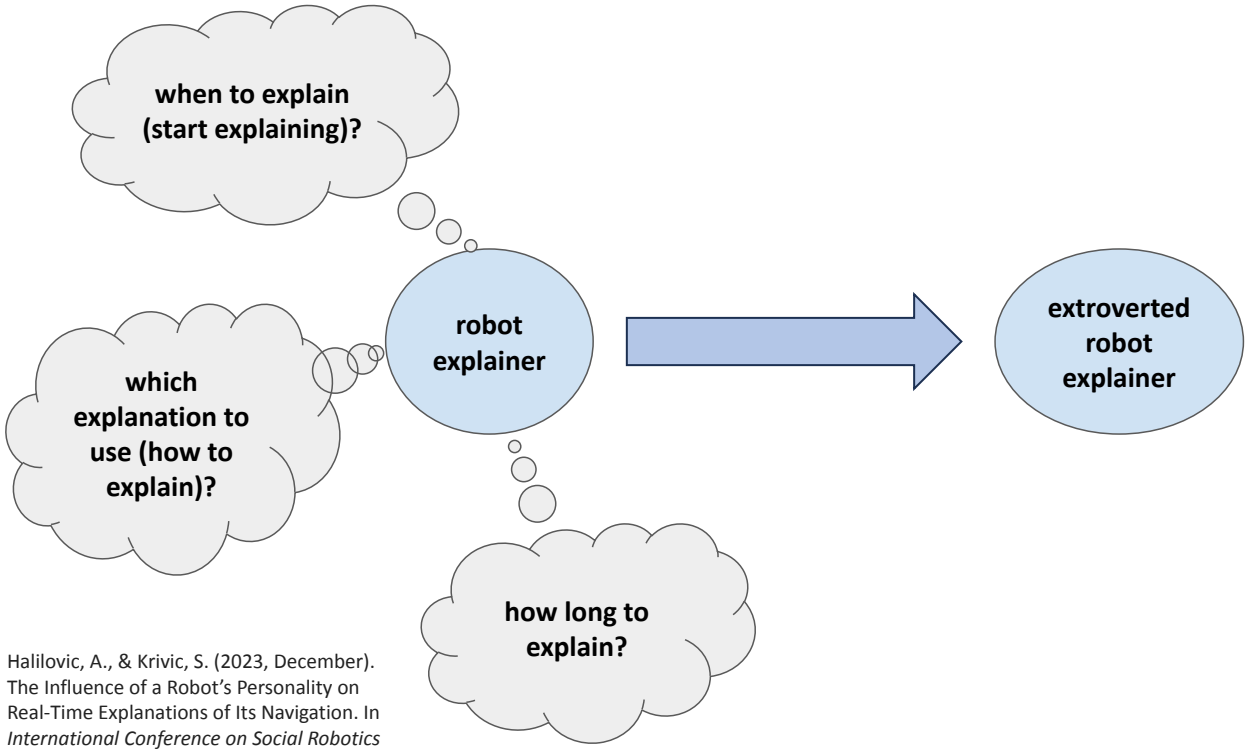
Halilovic, Amar, and Senka Krivic. "Robot Explanation Identity." *arXiv preprint arXiv:2405.13841* (2024).

Influence of robot personality on explanations



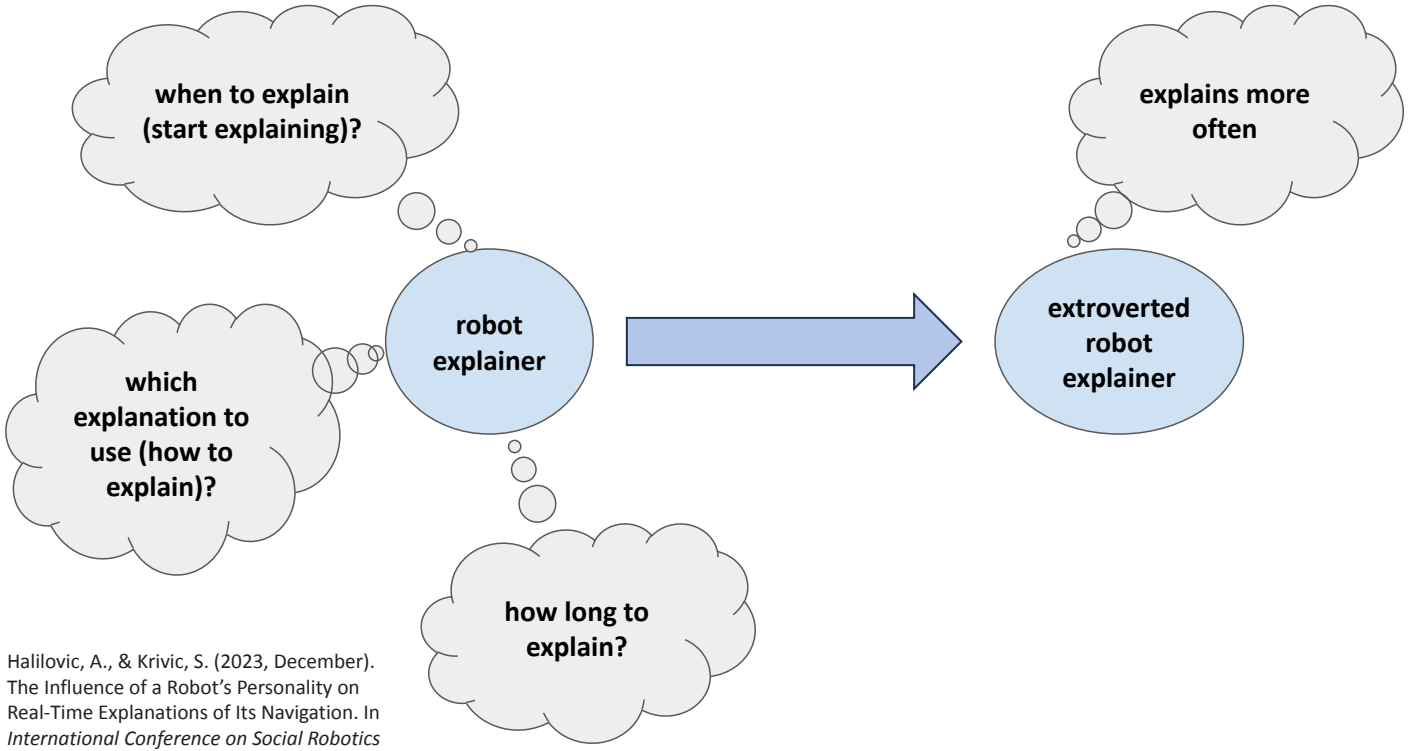
Halilovic, A., & Krivic, S. (2023, December). The Influence of a Robot's Personality on Real-Time Explanations of Its Navigation. In *International Conference on Social Robotics* (pp. 133-147). Singapore: Springer Nature

Influence of robot personality on explanations



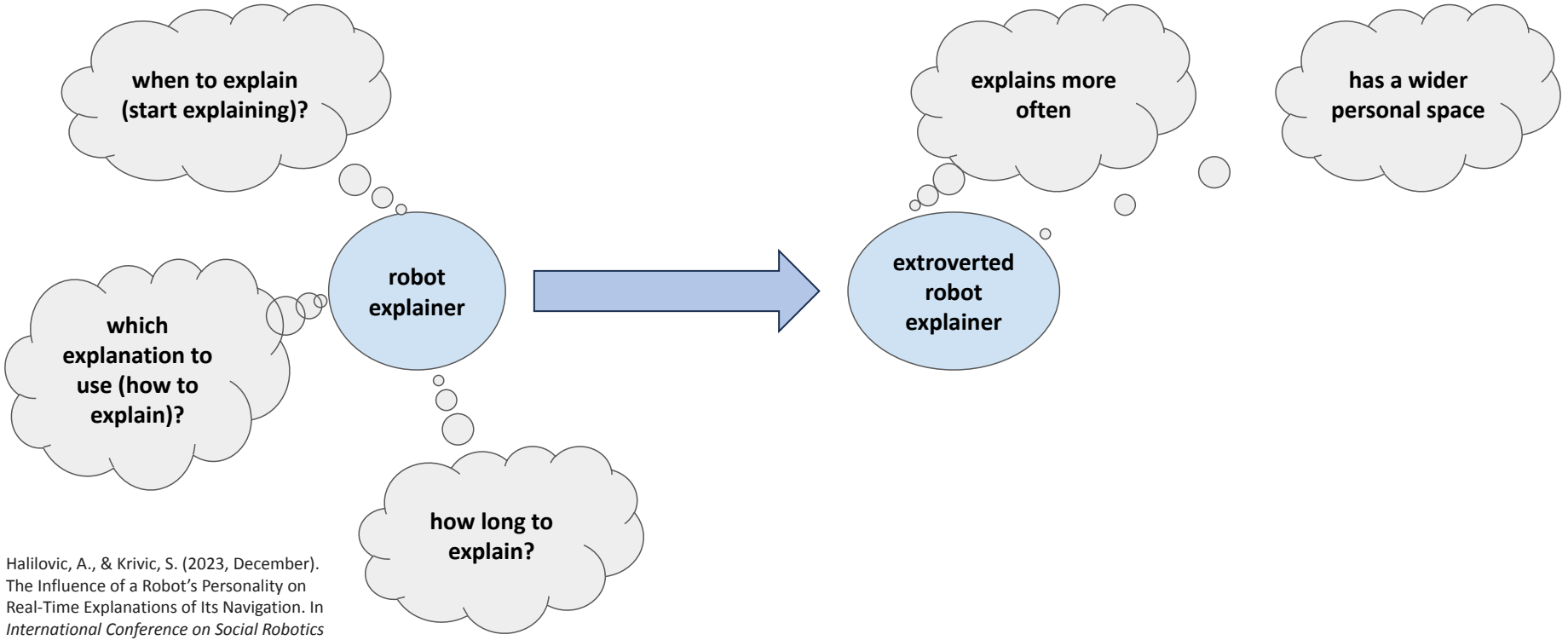
Halilovic, A., & Krivic, S. (2023, December). The Influence of a Robot's Personality on Real-Time Explanations of Its Navigation. In *International Conference on Social Robotics* (pp. 133-147). Singapore: Springer Nature

Influence of robot personality on explanations



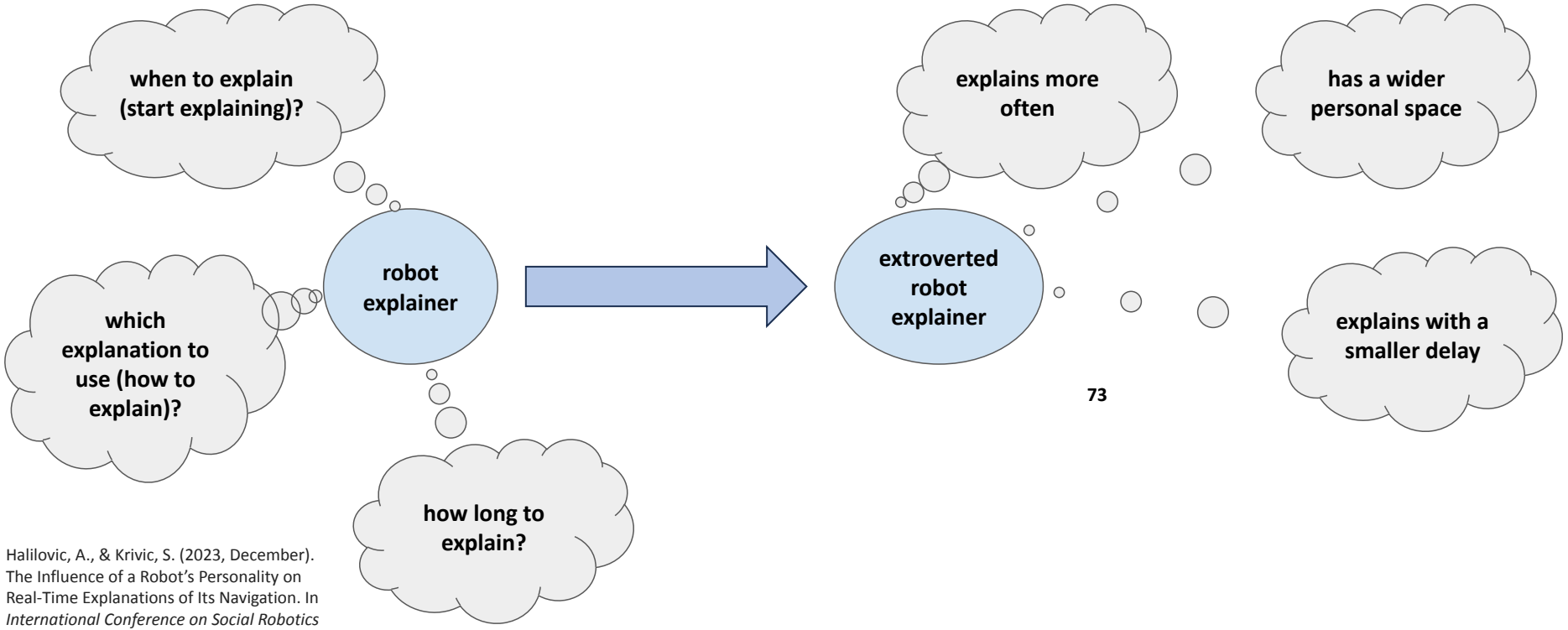
Halilovic, A., & Krivic, S. (2023, December). The Influence of a Robot's Personality on Real-Time Explanations of Its Navigation. In *International Conference on Social Robotics* (pp. 133-147). Singapore: Springer Nature

Influence of robot personality on explanations



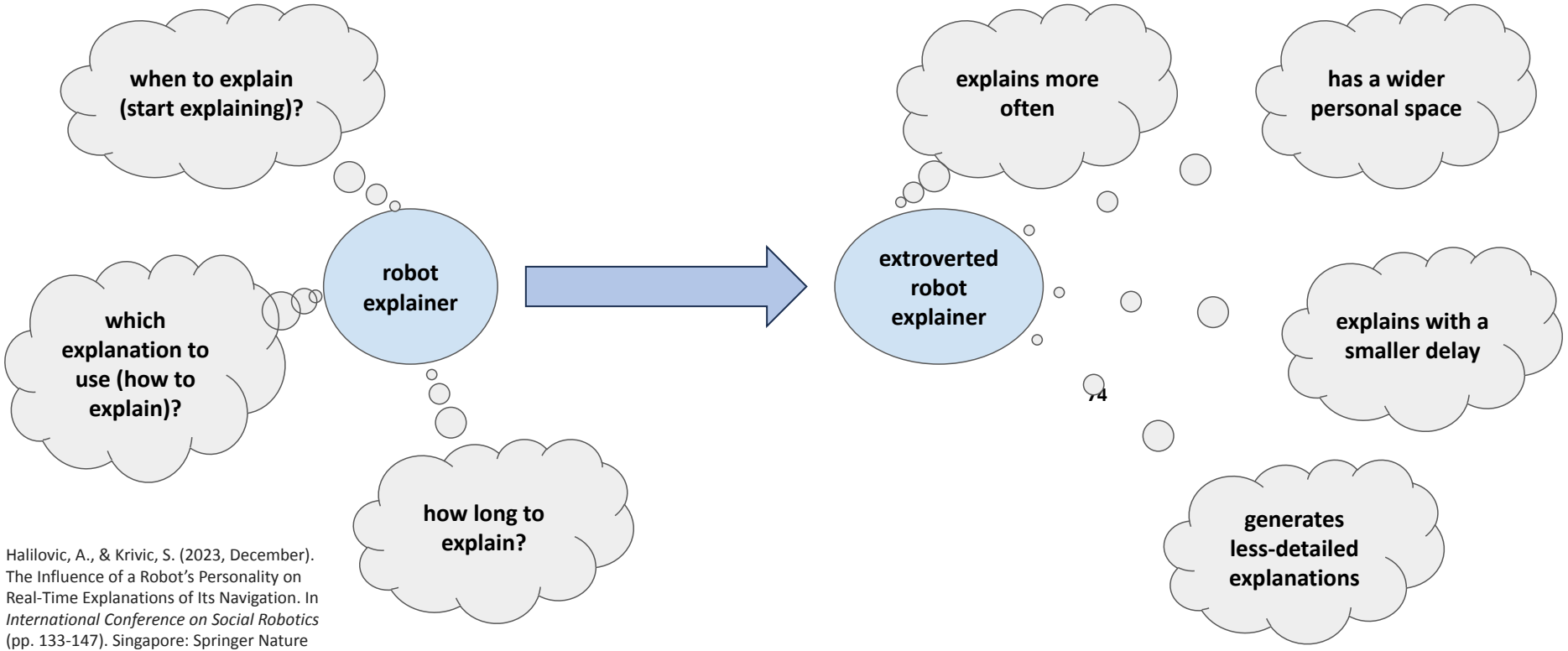
Halilovic, A., & Krivic, S. (2023, December). The Influence of a Robot's Personality on Real-Time Explanations of Its Navigation. In *International Conference on Social Robotics* (pp. 133-147). Singapore: Springer Nature

Influence of robot personality on explanations



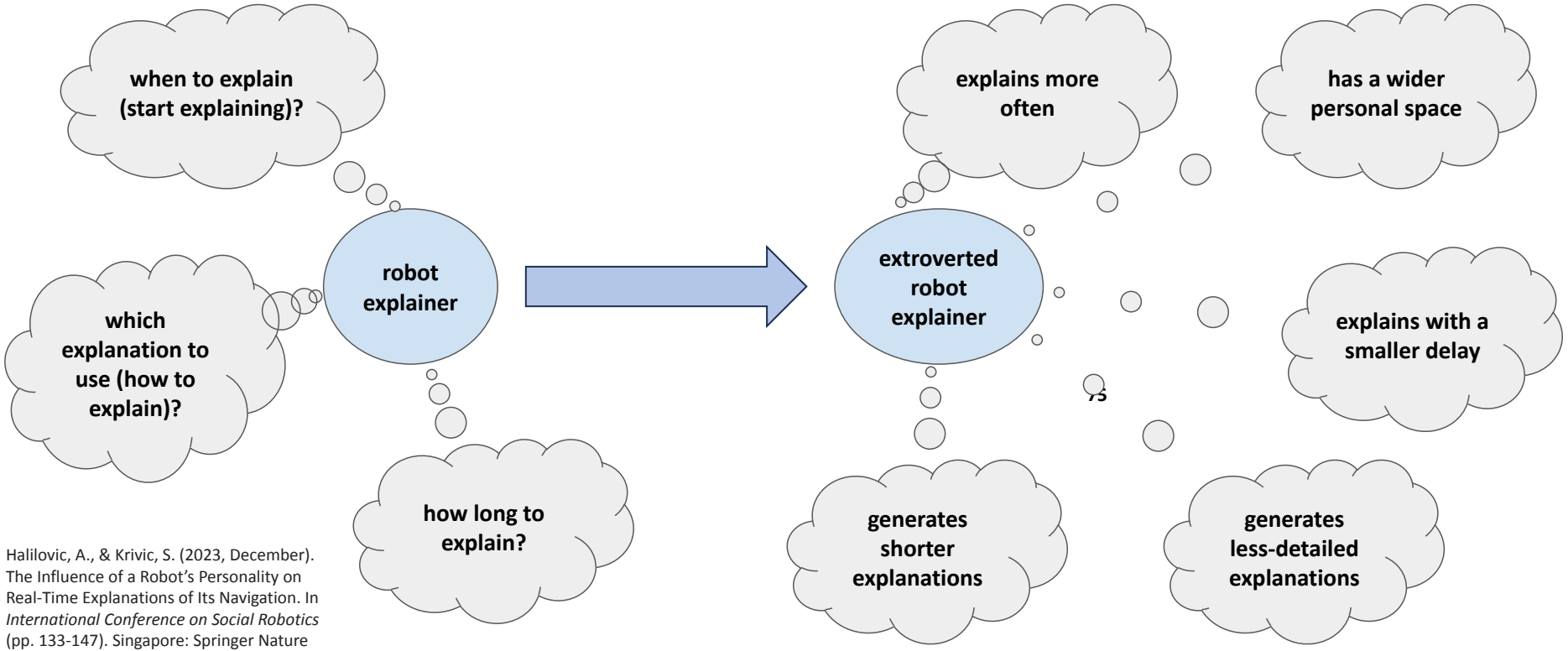
Halilovic, A., & Krivic, S. (2023, December). The Influence of a Robot's Personality on Real-Time Explanations of Its Navigation. In *International Conference on Social Robotics* (pp. 133-147). Singapore: Springer Nature

Influence of robot personality on explanations



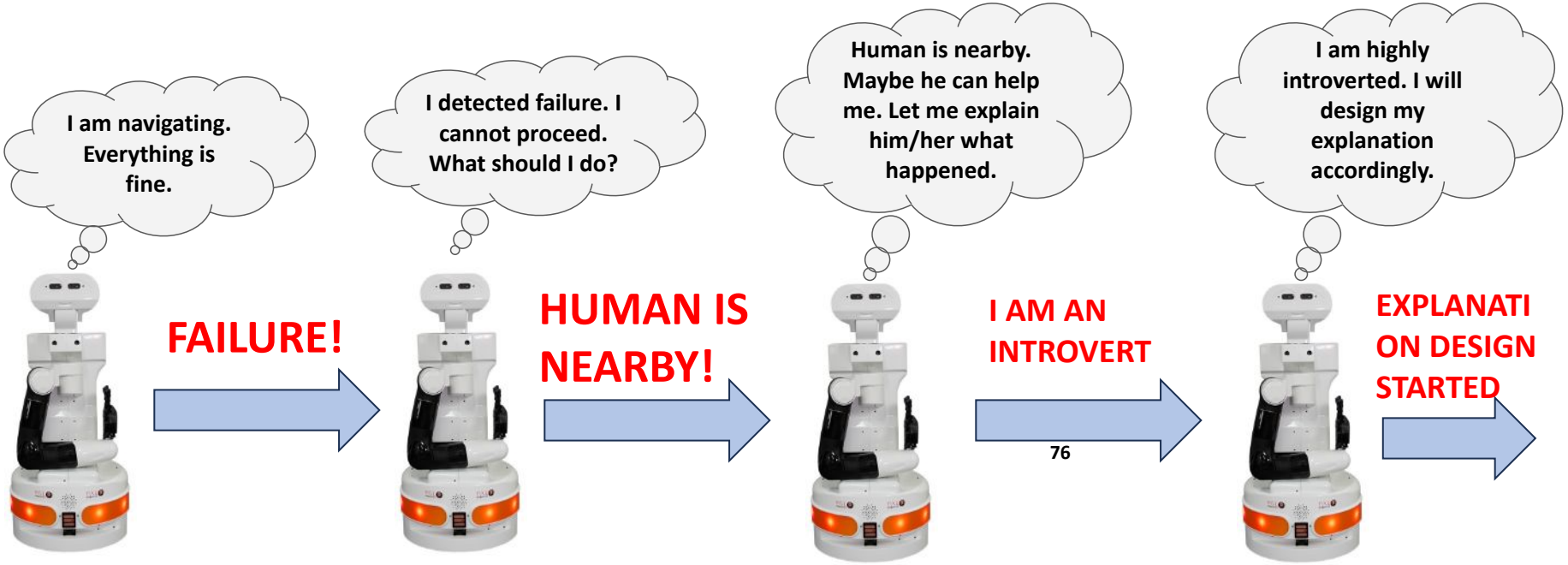
Halilovic, A., & Krivic, S. (2023, December). The Influence of a Robot's Personality on Real-Time Explanations of Its Navigation. In *International Conference on Social Robotics* (pp. 133-147). Singapore: Springer Nature

Influence of robot personality on explanations

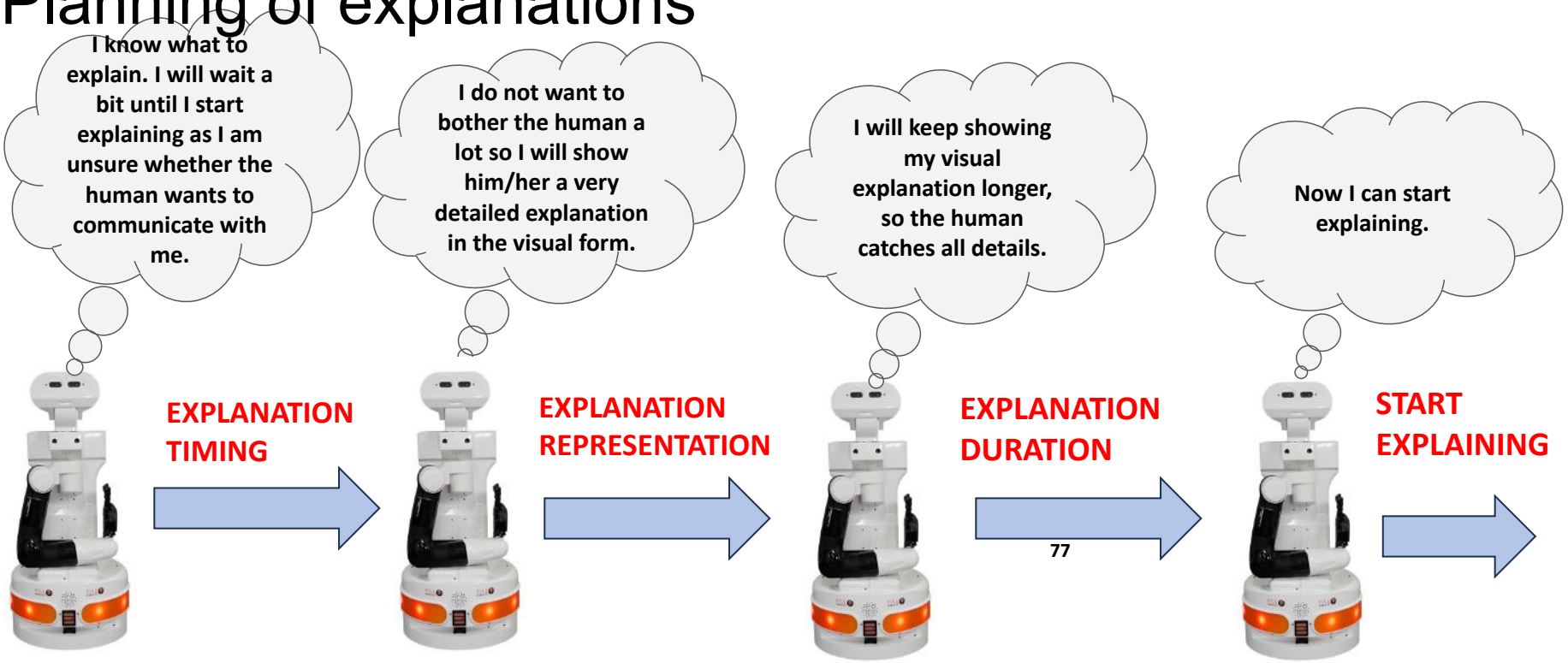


Halilovic, A., & Krivic, S. (2023, December). The Influence of a Robot's Personality on Real-Time Explanations of Its Navigation. In *International Conference on Social Robotics* (pp. 133-147). Singapore: Springer Nature

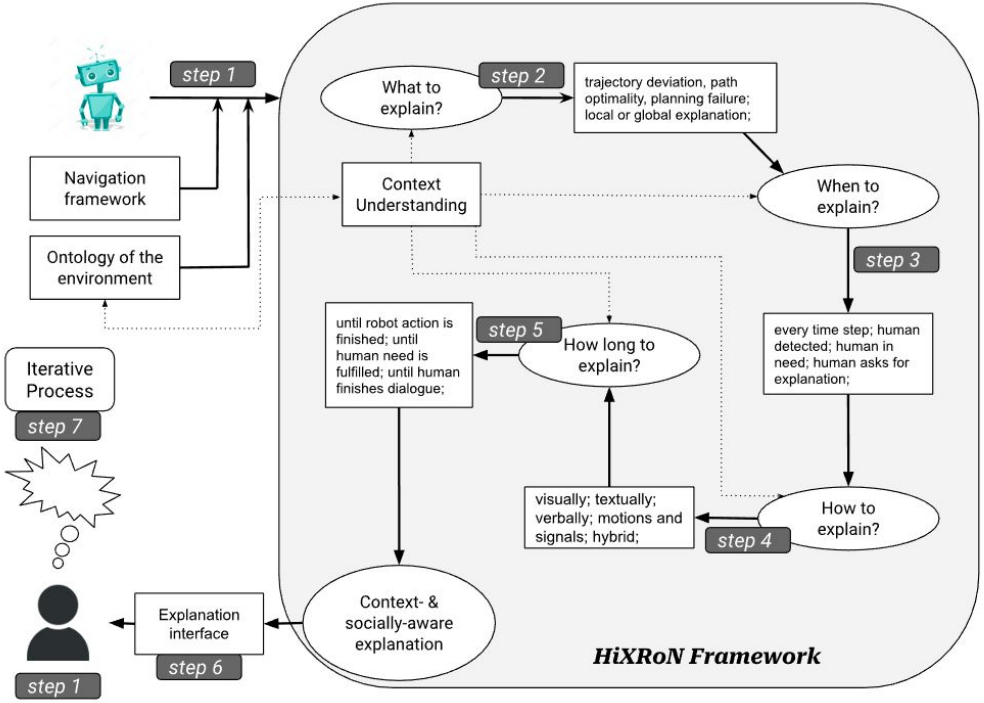
Planning of explanations



Planning of explanations

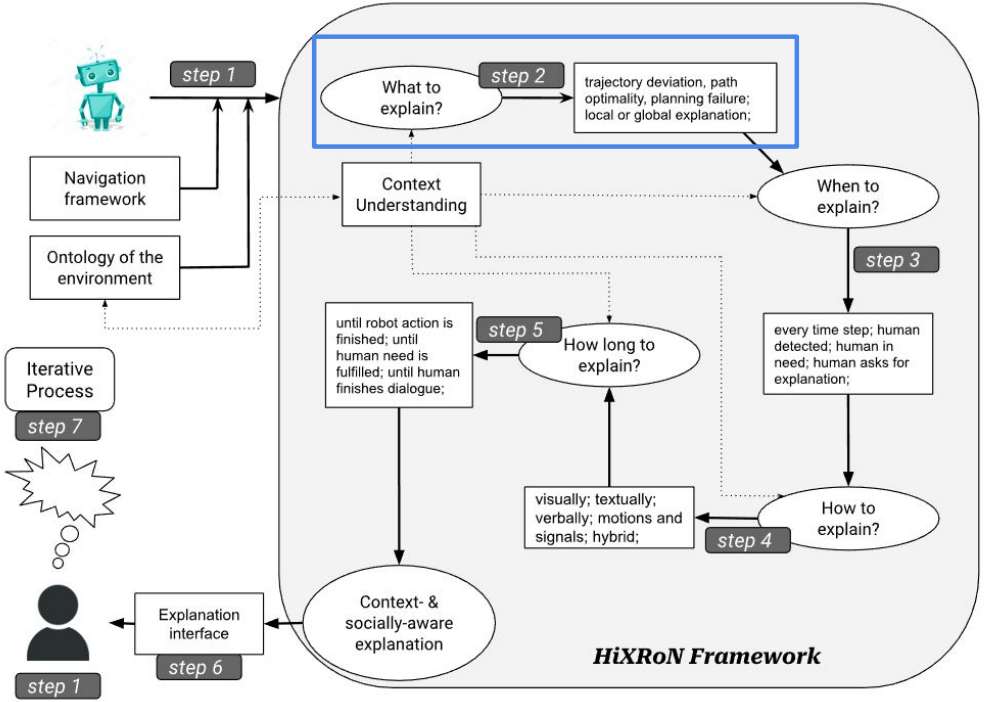


HiXRoN - **Hi**erarchical Framework for e**X**plainable **Ro**bot **N**avigation



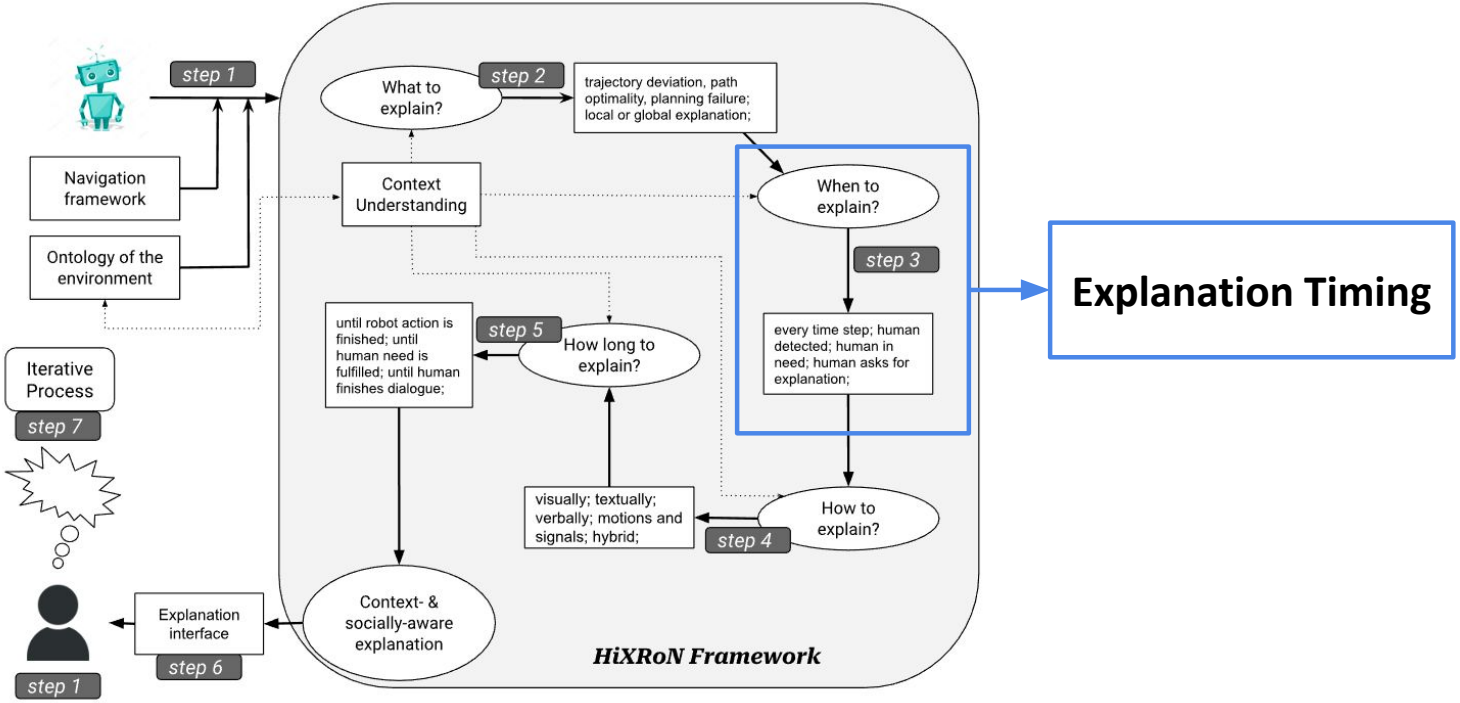
Halilovic, A., & Krivic, S. (2023, September). Towards a Holistic Framework for Explainable Robot Navigation. In *International Workshop on Human-Friendly Robotics* (pp. 213-228). Cham: Springer Nature Switzerland.

HiXRoN - **Hi**erarchical Framework for e**X**plainable **Ro**bot **N**avigation



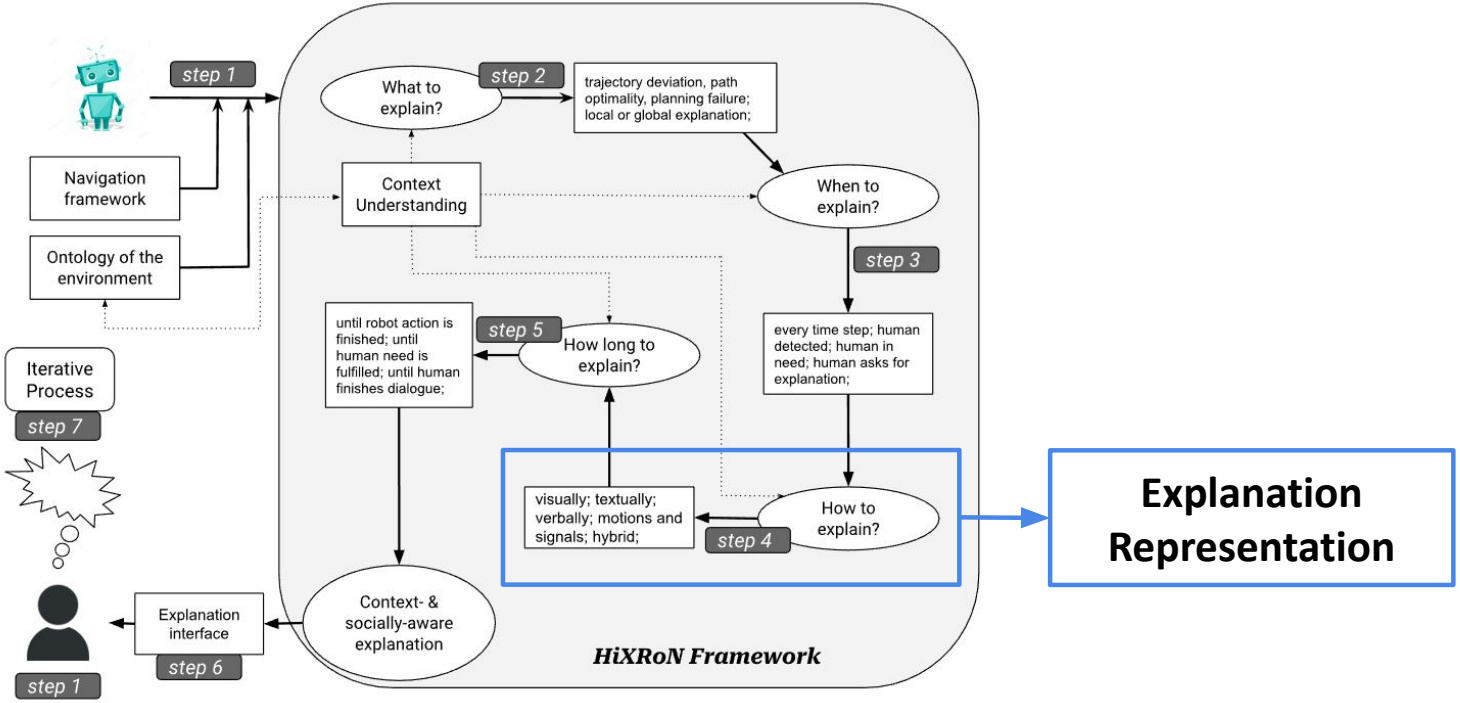
Halilovic, A., & Krivic, S. (2023, September). Towards a Holistic Framework for Explainable Robot Navigation. In *International Workshop on Human-Friendly Robotics* (pp. 213-228). Cham: Springer Nature Switzerland.

HiXRoN - **Hi**erarchical Framework for e**X**plainable **Ro**bot **N**avigation



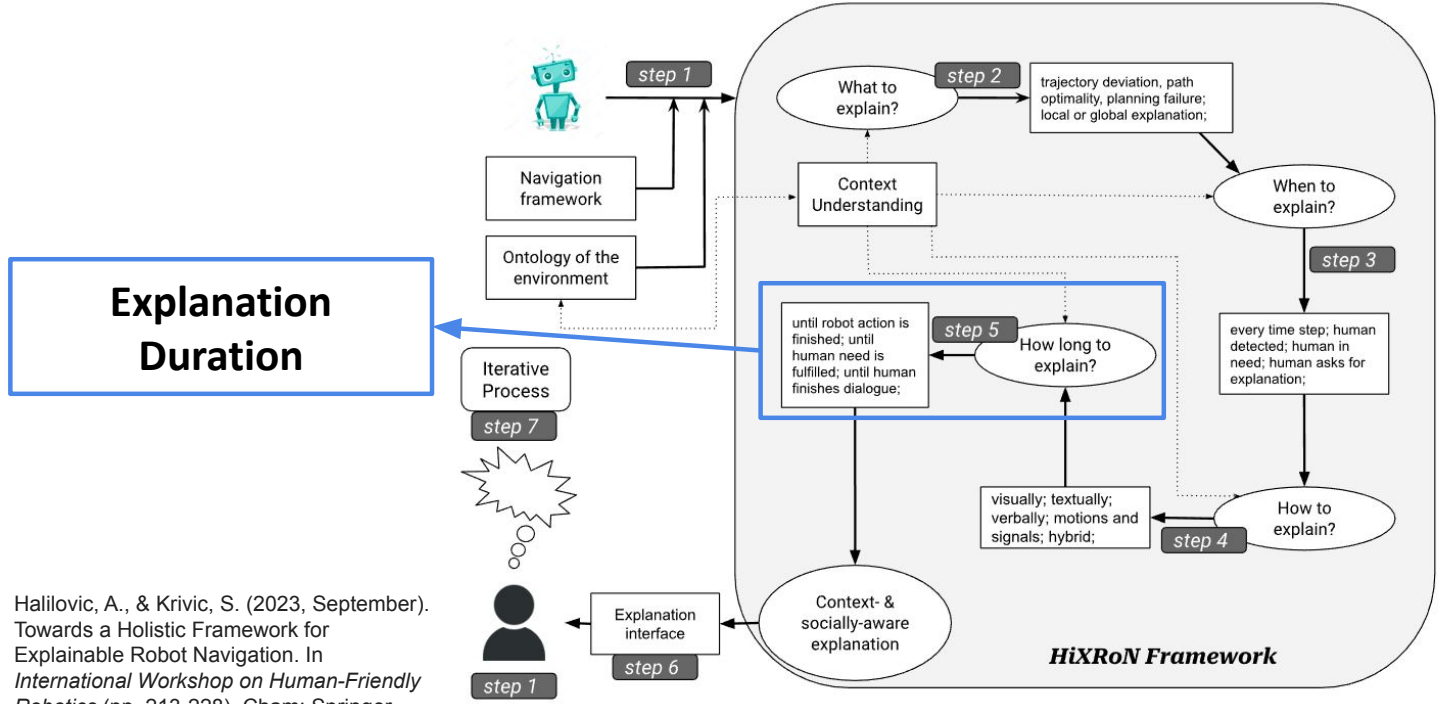
Halilovic, A., & Krivic, S. (2023, September). Towards a Holistic Framework for Explainable Robot Navigation. In *International Workshop on Human-Friendly Robotics* (pp. 213-228). Cham: Springer Nature Switzerland.

HiXRoN - **Hi**erarchical Framework for e**X**plainable **Ro**bot **N**avigation



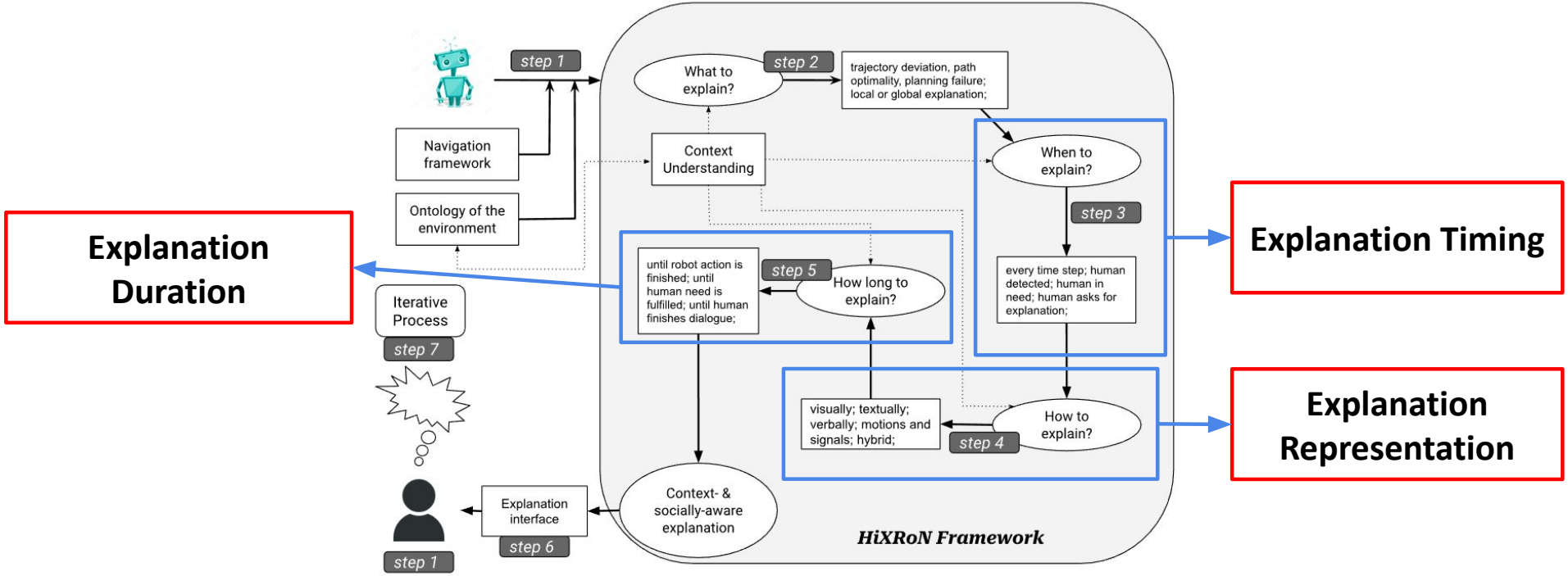
Halilovic, A., & Krivic, S. (2023, September). Towards a Holistic Framework for Explainable Robot Navigation. In *International Workshop on Human-Friendly Robotics* (pp. 213-228). Cham: Springer Nature Switzerland.

HiXRoN - **Hi**erarchical Framework for e**X**plainable **Ro**bot **N**avigation

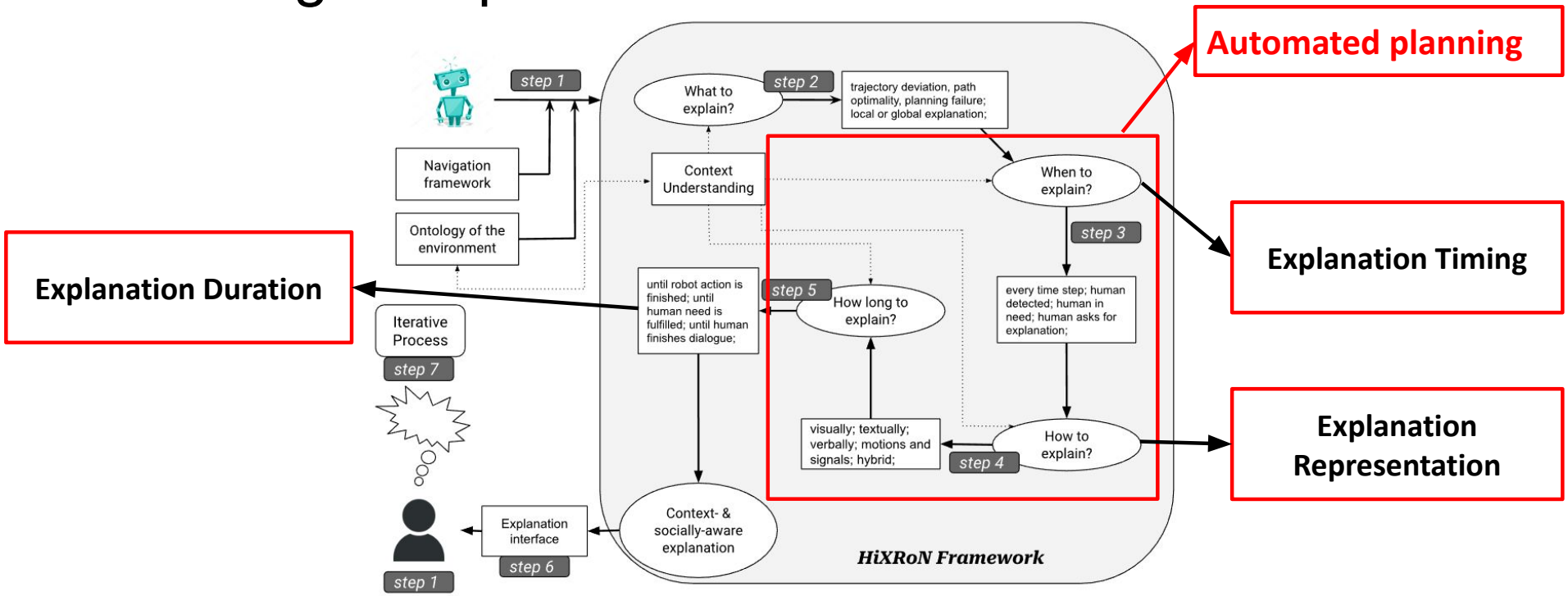


Halilovic, A., & Krivic, S. (2023, September). Towards a Holistic Framework for Explainable Robot Navigation. In *International Workshop on Human-Friendly Robotics* (pp. 213-228). Cham: Springer Nature Switzerland.

Planning of explanations



Planning of explanations

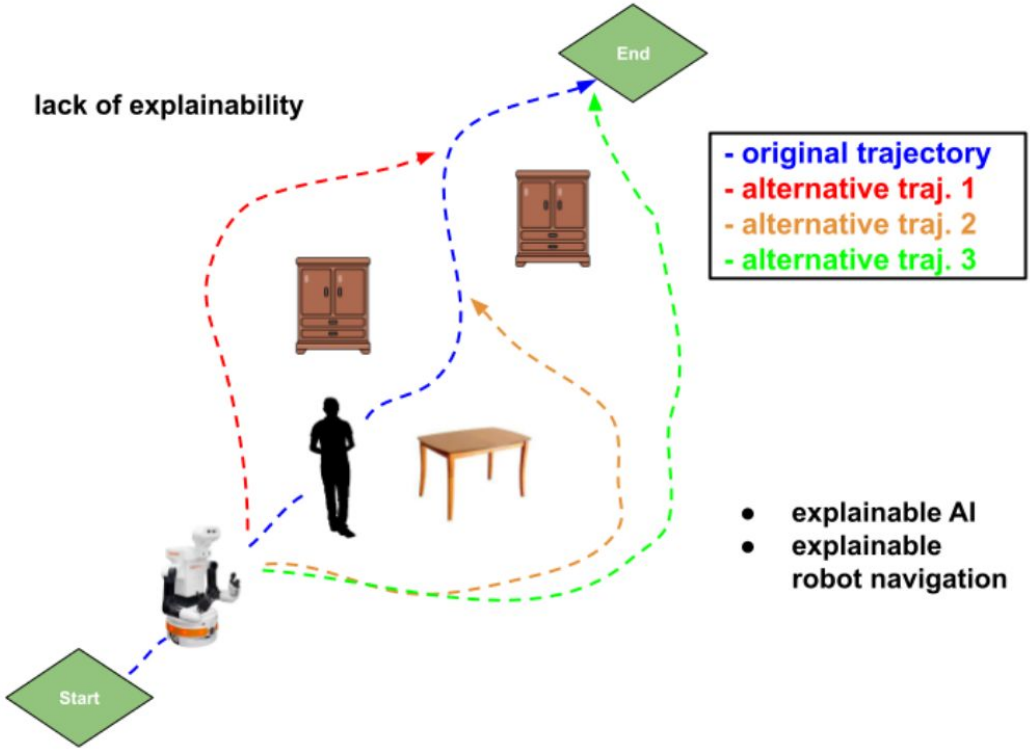


Explanations in Robot Motion Planning

- Explaining trajectory selection
- Environment-based justification
- Deviations and failure recovery

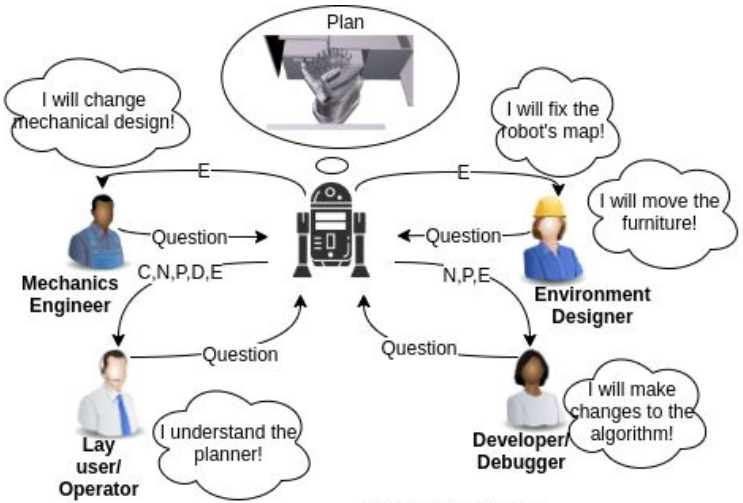
Motivation

- lack of explainability



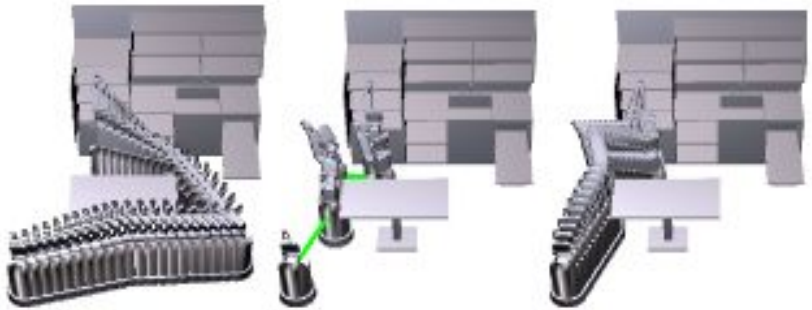
- explainable AI
- explainable robot navigation

Types of explanations and user needs



Question types:
"Why did you fail?",
"Why trajectory A not B?"

Explanations types:
C: Cost-based explanation
N: Constraint-based explanation
P: Algorithm-parameter-based explanation
D: Design-based explanation
E: Environment-based explanation

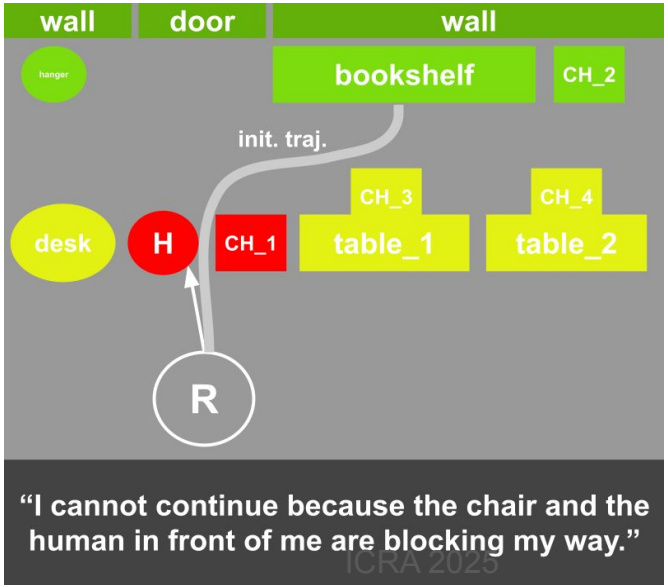


(a) Original plan ξ_A (10sec budget) (b) User's expected trajecotory ξ_B (c) New plan, using 30sec budget

M Brandao, G Canal, S Krivić, D Magazzeni, ICRA 2021

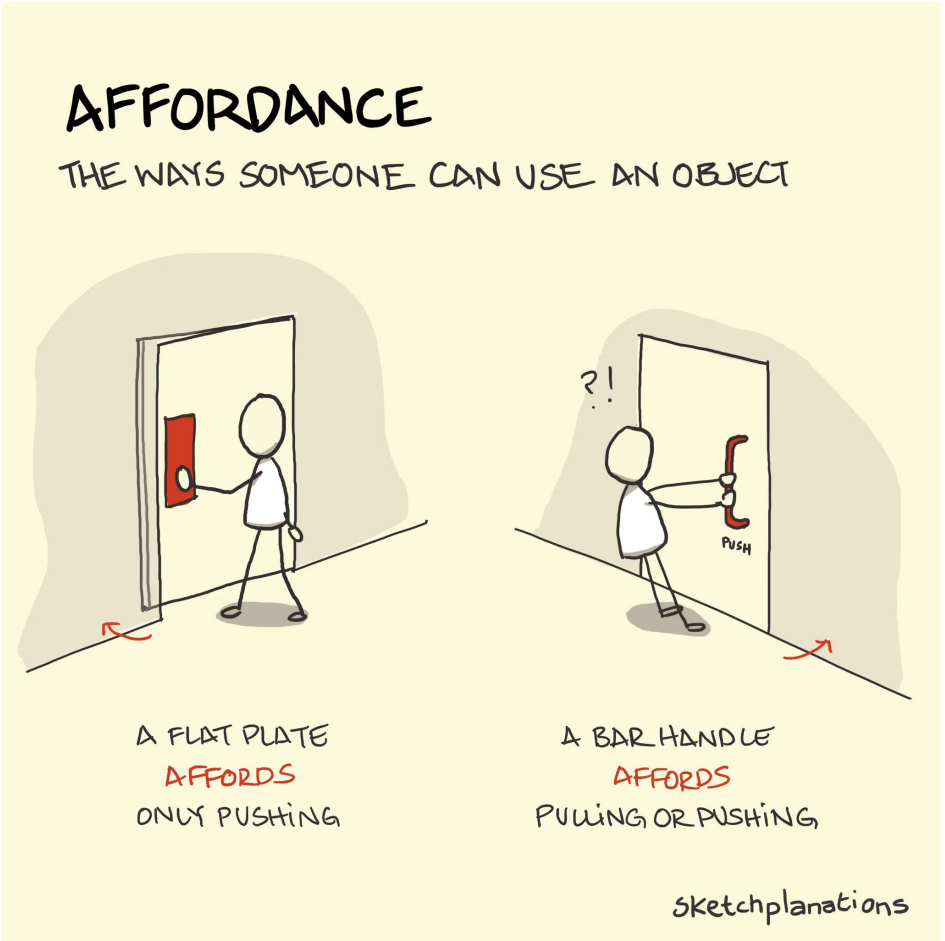
Explanation of robot navigation

- Explanations in terms of objects in the robot environment
- Explanations by perturbing object positions in the environment and replanning in each perturbation

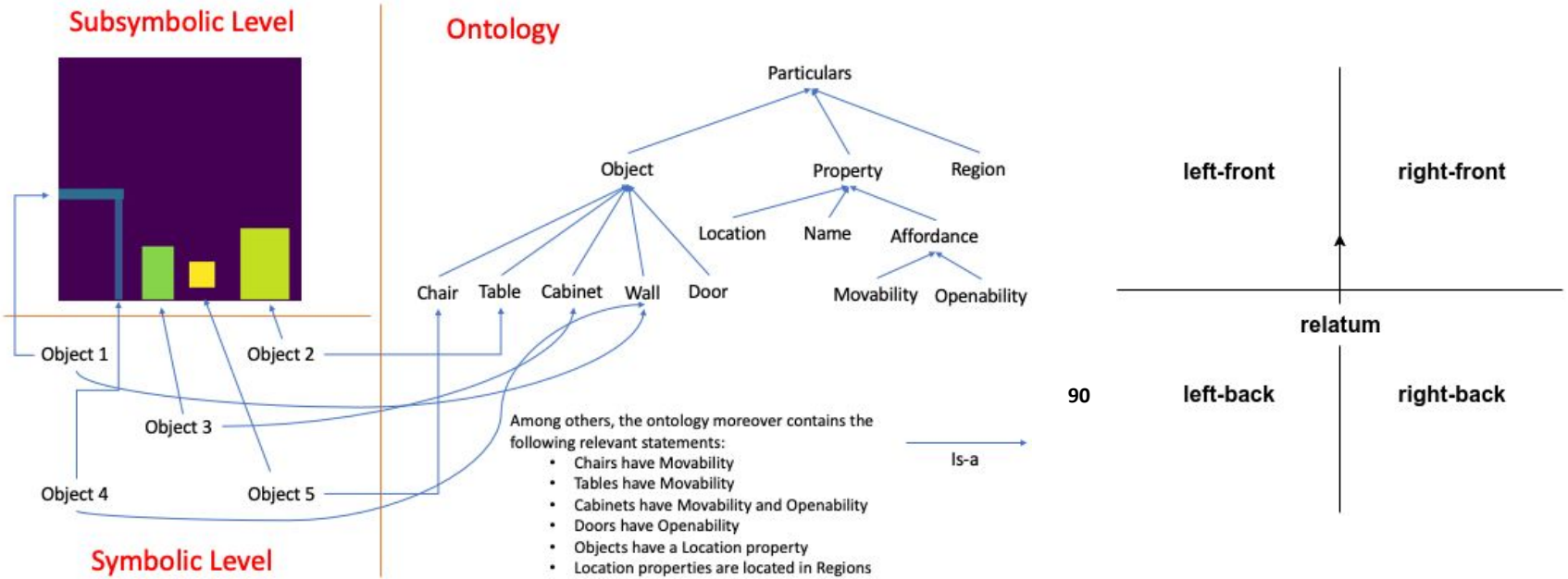


Affordances of objects

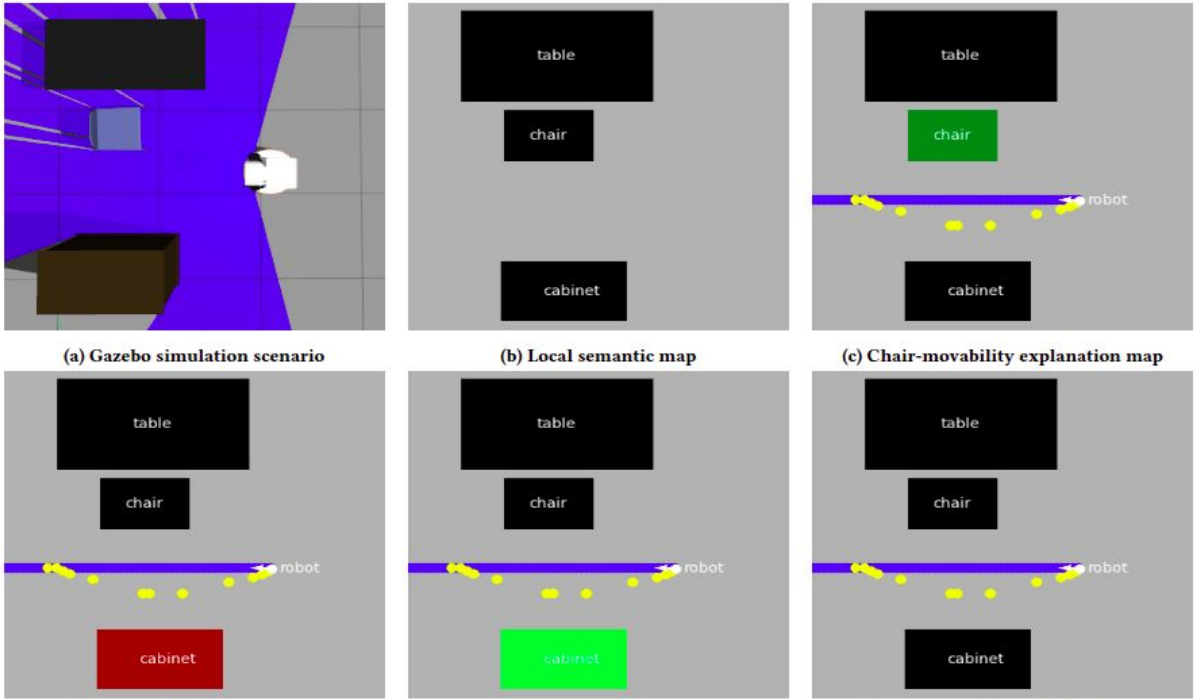
By Makito Nagawa



Visual-textual environmental explanations



Halilovic, Amar, and Felix Lindner. "Visuo-Textual Explanations of a Robot's Navigational Choices." *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 2023.



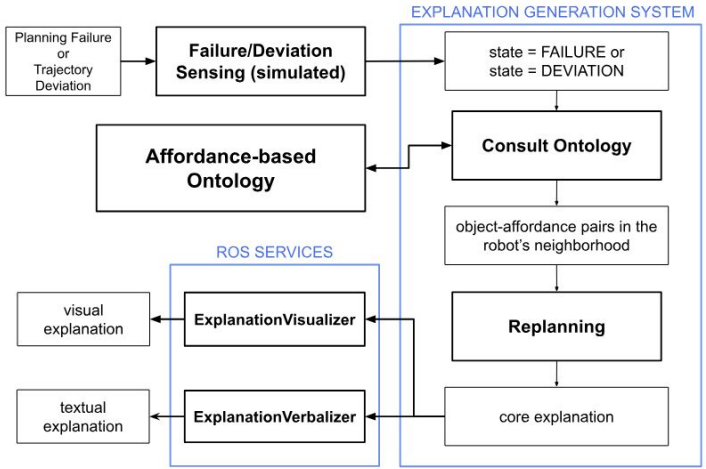
(a) Gazebo simulation scenario (b) Local semantic map (c) Chair-movability explanation map

(d) Cabinet-movability explanation map (e) Cabinet-openability explanation map (f) Table-movability explanation map

Figure 3: Along with the visual explanations come the textual explanations and suggestions: c) *"Because of the chair right-front of me, I deviate from the initial plan."*; *"Dear human, please move the chair, so I proceed more smoothly."*; d) *"If the cabinet left-front of me was not there, I would deviate more from the initial plan."*; e) *"If the cabinet left-front of me was open, I would deviate less from the initial plan."*

Affordance-based explanations

- Formalization of the approach
- Explanation generation framework



Algorithm 1: Generation process of affordance-based explanation

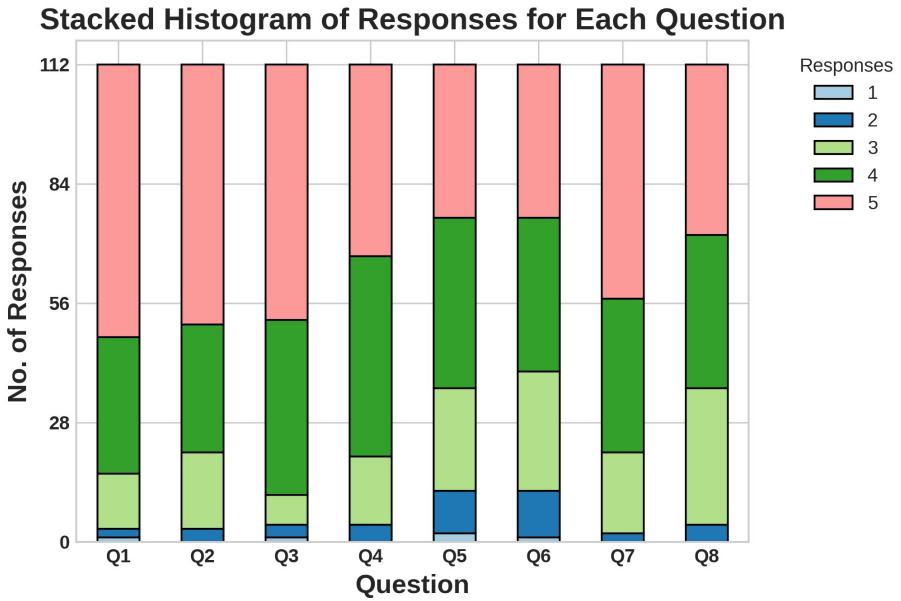
Input: robot, map, ontology, QSR
Output: visual_explanation, textual_explanation

```
repeat
    state = GetRobotState(robot)
    if state == FAILURE or state == DEVIATION then
        objects = ScanNeighborhood(robot, map)
        affordances = ontology[objects]
        if affordances ≠ ∅ then
            core_explanation = Replanning(objects, affordances, map)
            if core_explanation ≠ ∅ then
                visual_explanation = Visualize(core_explanation, map)
                textual_explanation = Verbalize(core_explanation, QSR)
until goal not reached
```

Halilovic, Amar, and Senka Krivic. "Affordance-Based Explanations of Robot Navigation." to appear at *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025.

Affordance-based explanations

- User study on explanation satisfaction
- Multimodal explanations (visual + textual)

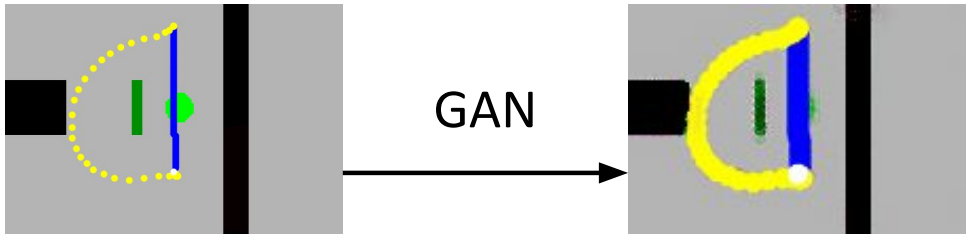


Halilovic, Amar, and Senka Krivic. "Affordance-Based Explanations of Robot Navigation." to appear at *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025.

Visual explanations of local path planning with LIME and GAN

Affordance-based explanations with generative AI (image-to-image translation and LLMs):

- The LIME explanation explains how obstacles and/or parts of obstacles contribute to the deviation.
- GAN model generates visual explanations of local path plans.
- faster than replanning



“Both obstacles increase the deviation, but the round one does so more significantly. If it were not there, the robot would follow the global plan. If the rectangular obstacle were not there, the robot would still deviate, but less.”

Halilović, Amar, and Senka Krivić. "Towards Fast Visual Explanations of Local Path Planning with LIME and GAN." (2024).

Local Interpretable Model-Agnostic Explanations (LIME)

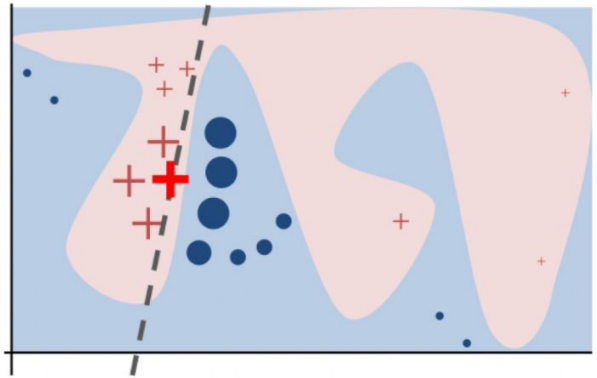


Illustration of the local approximation performed by LIME

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Examples of explanations produced by LIME

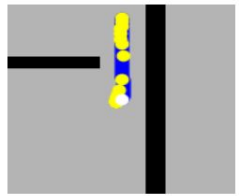
Both images taken from M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, Aug. 2016, 1135–1144.

- LIME identifies input features that are relevant for classification by **approximating a complex classification model f with a local linear approximator g**
- The approximating model is **trained with examples x' that are locally perturbed around the original example x for which we want an explanation**
- Given a function π_x that evaluates **the locality of examples x'**, a **loss function $\mathcal{L}$**, and a **complexity evaluation function Ω**, an explanation is produced by solving the following optimisation problem:

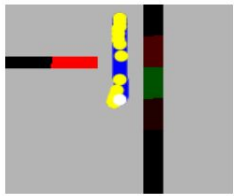
$$\xi(\boldsymbol{x}) = \underset{g \in G}{\arg\min}\, \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

- For images, explainable image patches are identified by using **super-pixels** as inputs to the local model g

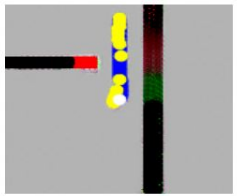
(a) **C1**: robot



(b) **C1**: costmap

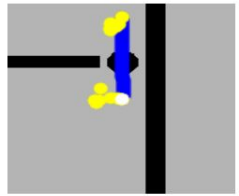


(c) **C1**: LIME expl.

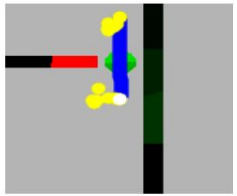


(d) **C1**: GAN expl.

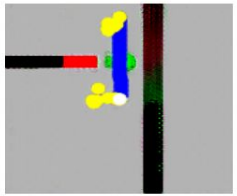
(e) **C2**: robot



(f) **C2**: costmap

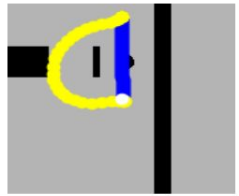


(g) **C2**: LIME expl.

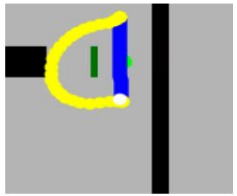


(h) **C2**: GAN expl.

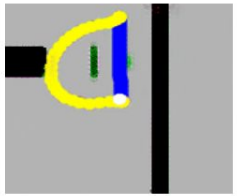
(i) **C3**: robot



(j) **C3**: costmap



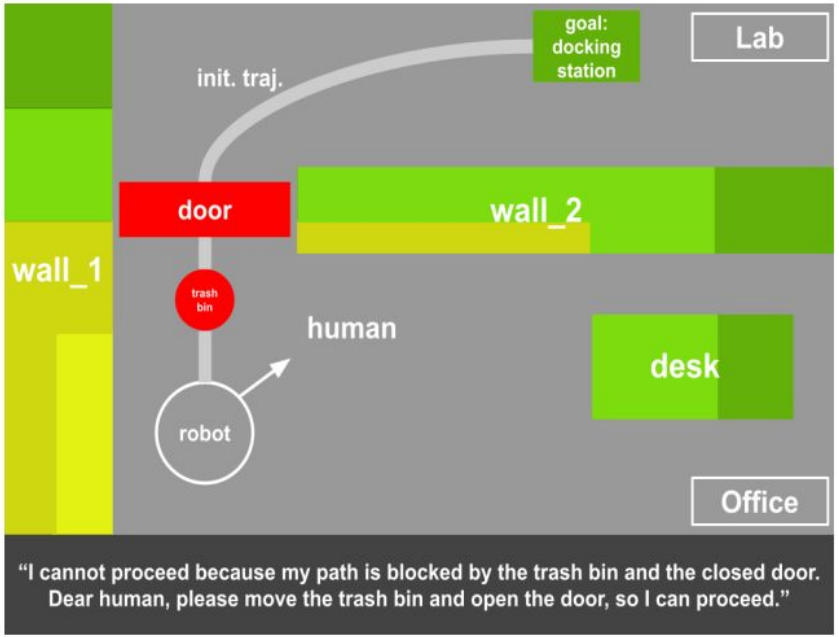
(k) **C3**: LIME expl.



(l) **C3**: GAN expl.

Using explanations to ask for help?

(a) Navigation scenario



(b) Corresponding explanation

Amar Halilovic, Vanchha Chandrayan, and Senka Krivic, TAHRI 2024

Conclusion – Rethinking Explainability in Robotics

- Explainable Robotics sits at the intersection of AI, HRI, philosophy, and design.
- Not all explanations are equal — effectiveness depends on *who*, *when*, and *why*.
- Embodiment and context shape what is perceived as *understandable behavior*.
- Real-world deployment still lags behind theoretical XAI frameworks.
- Transparency ≠ Trust: Avoid the illusion of explainability.
- Responsible design must consider user needs, limitations, and ethical boundaries.
- Robots need to be built that **not only act** intelligently — but also **explain** meaningfully.

Thank you for you attention!

Email: senka.krivic@etf.unsa.ba