Explainable Robot Navigation

Environment-Centered and Human-Centered Approach

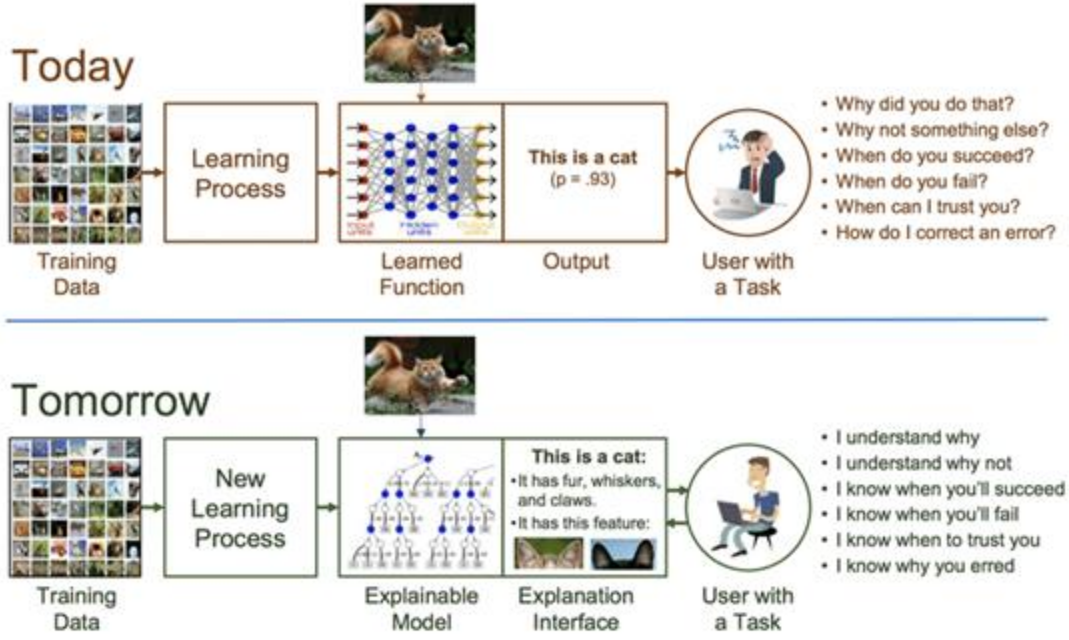
Dissertation Project Presentation

Amar Halilovic
Institute of Artificial Intelligence
Ulm University

amar.halilovic@uni-ulm.de

Explainable AI (XAI)

- Makes AI decisions transparent and understandable
- Builds trust and supports accountability
- Applicable to different domains



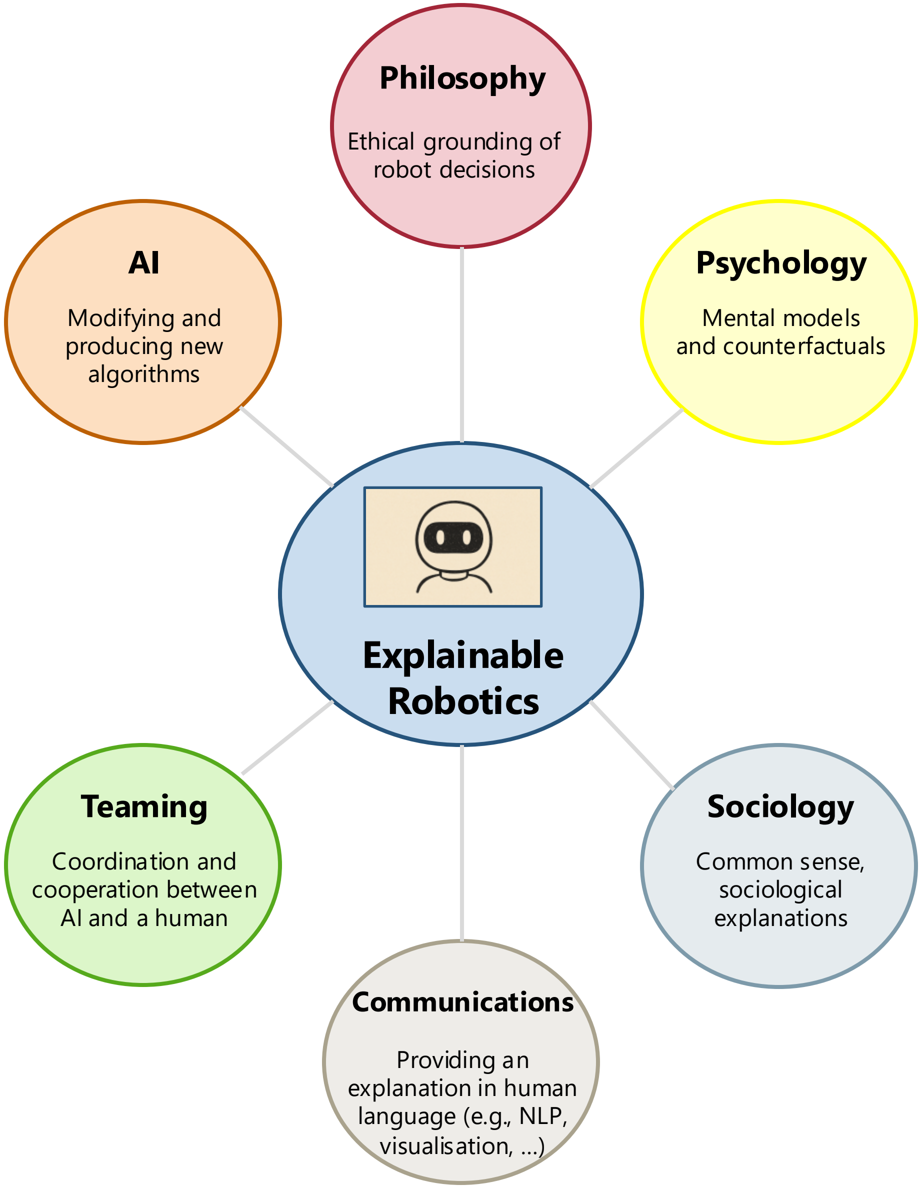
Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. AI magazine, 40(2), 44-58.

From XAI to Explainable Robotics

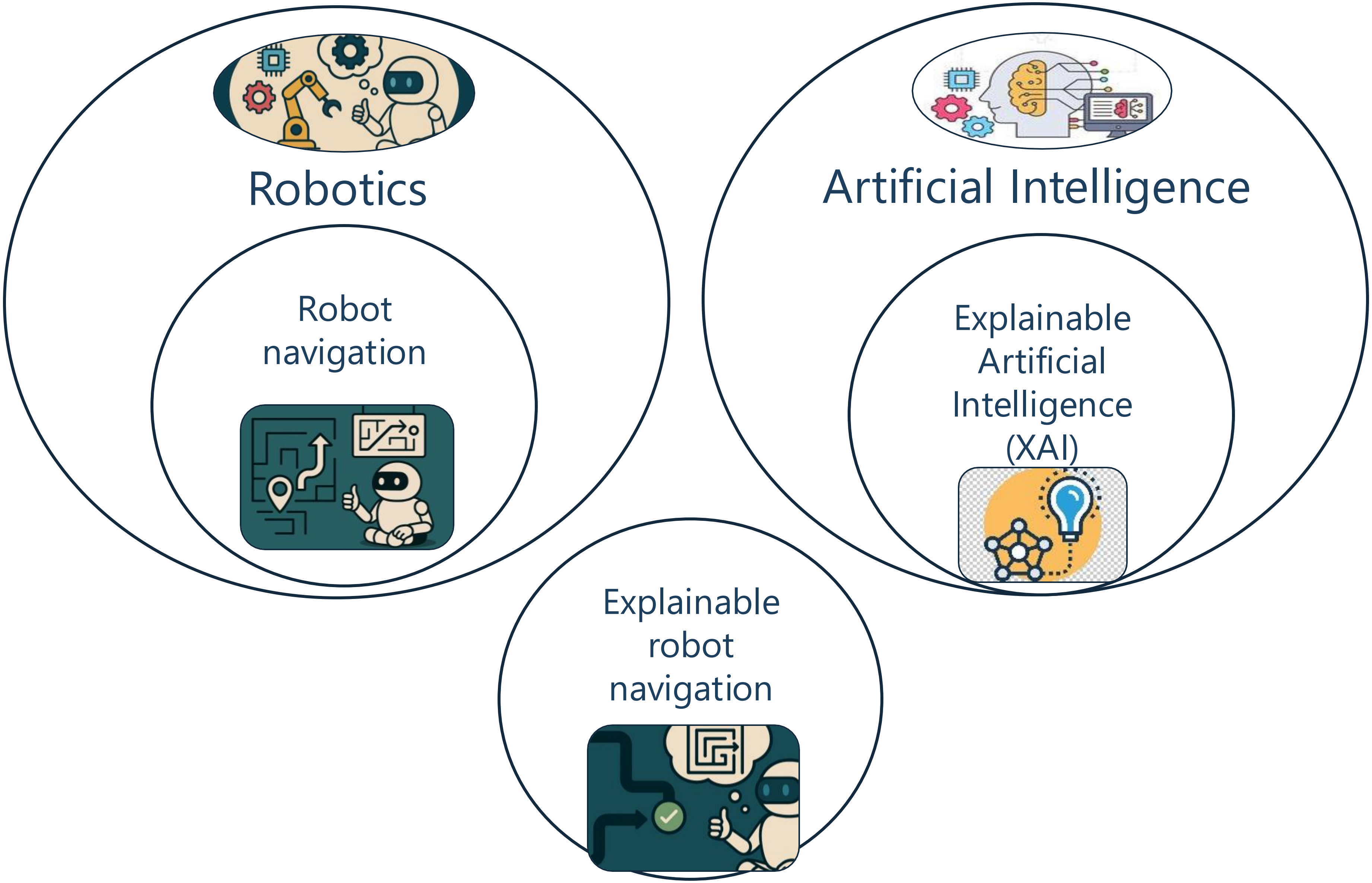
- Embodied, interactive, real-time agents
- Need for context-aware and multimodal explanations
- Challenges of explainability in dynamic environments

What is Explainable Robotics

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

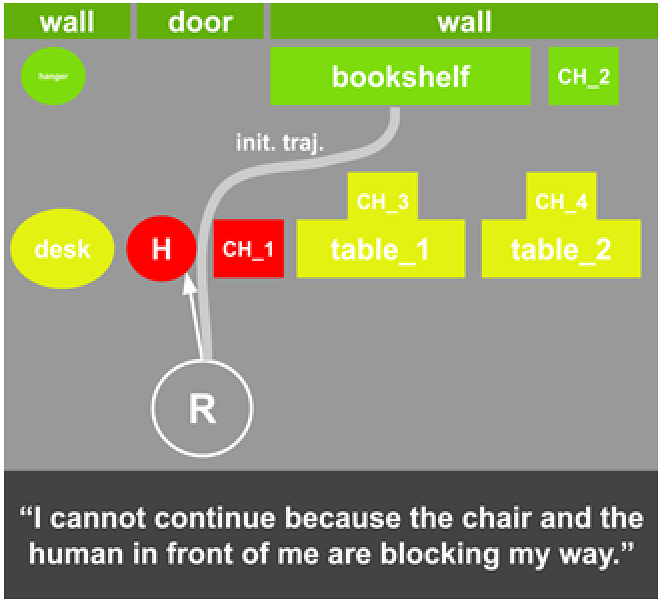


Explainable Robot Navigation



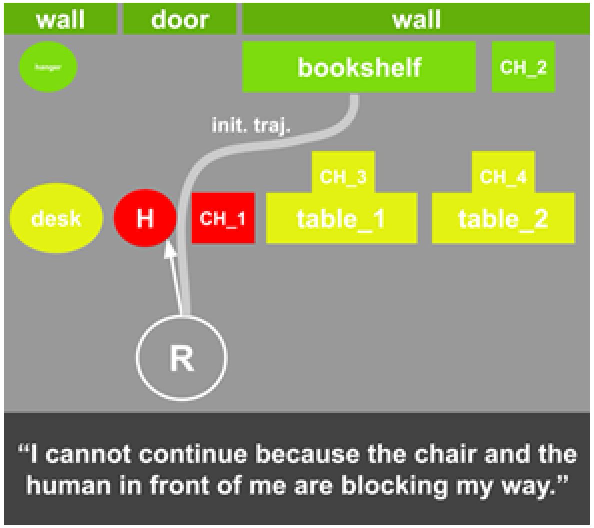
Motivational and Running Example

- Focused on service robotics
- Robot Librarian: delivers books to library visitors
- Has multimodal explanation capabilities

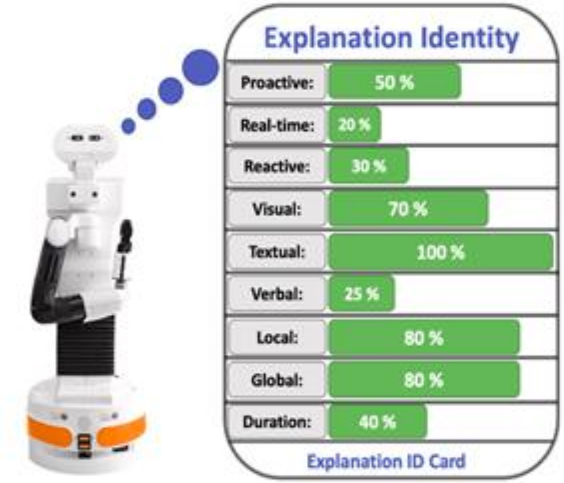


Research Challenges

Environment-centered explanation generation for robot navigation



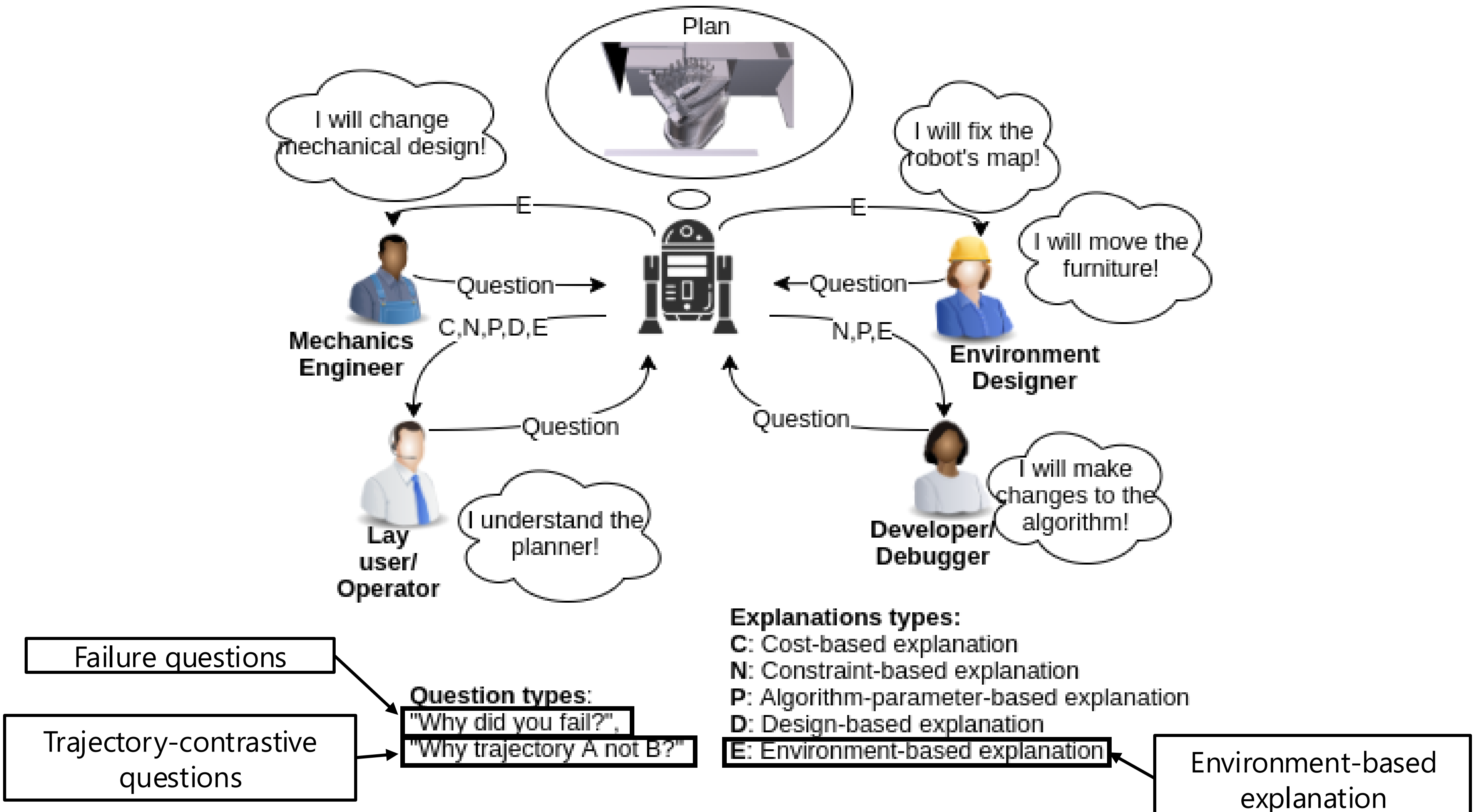
Human-centered explanation planning for robot navigation



Environment-centered explanations of robot navigation

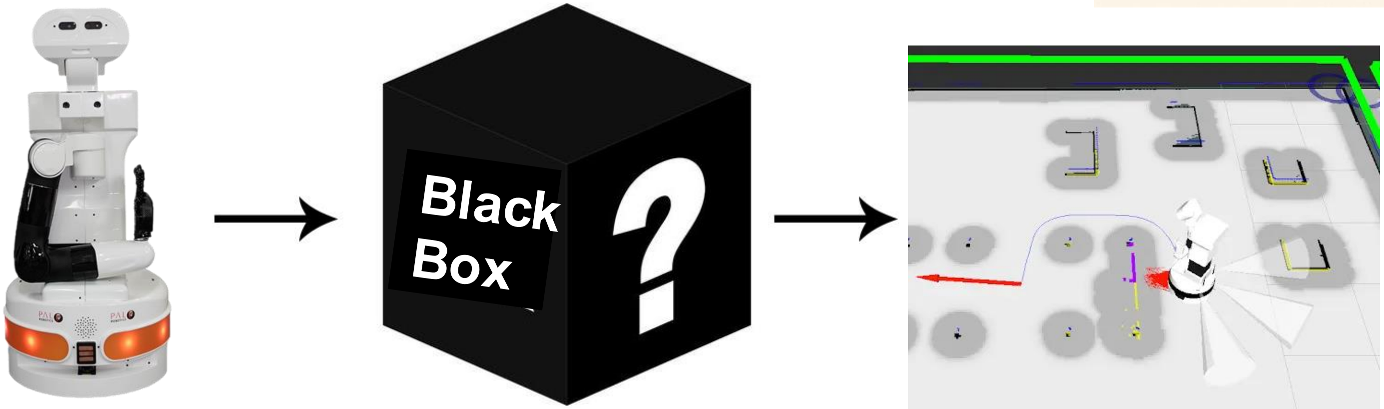
- Focus on environment, not robot internals

Types of explanations and user needs in robot motion planning

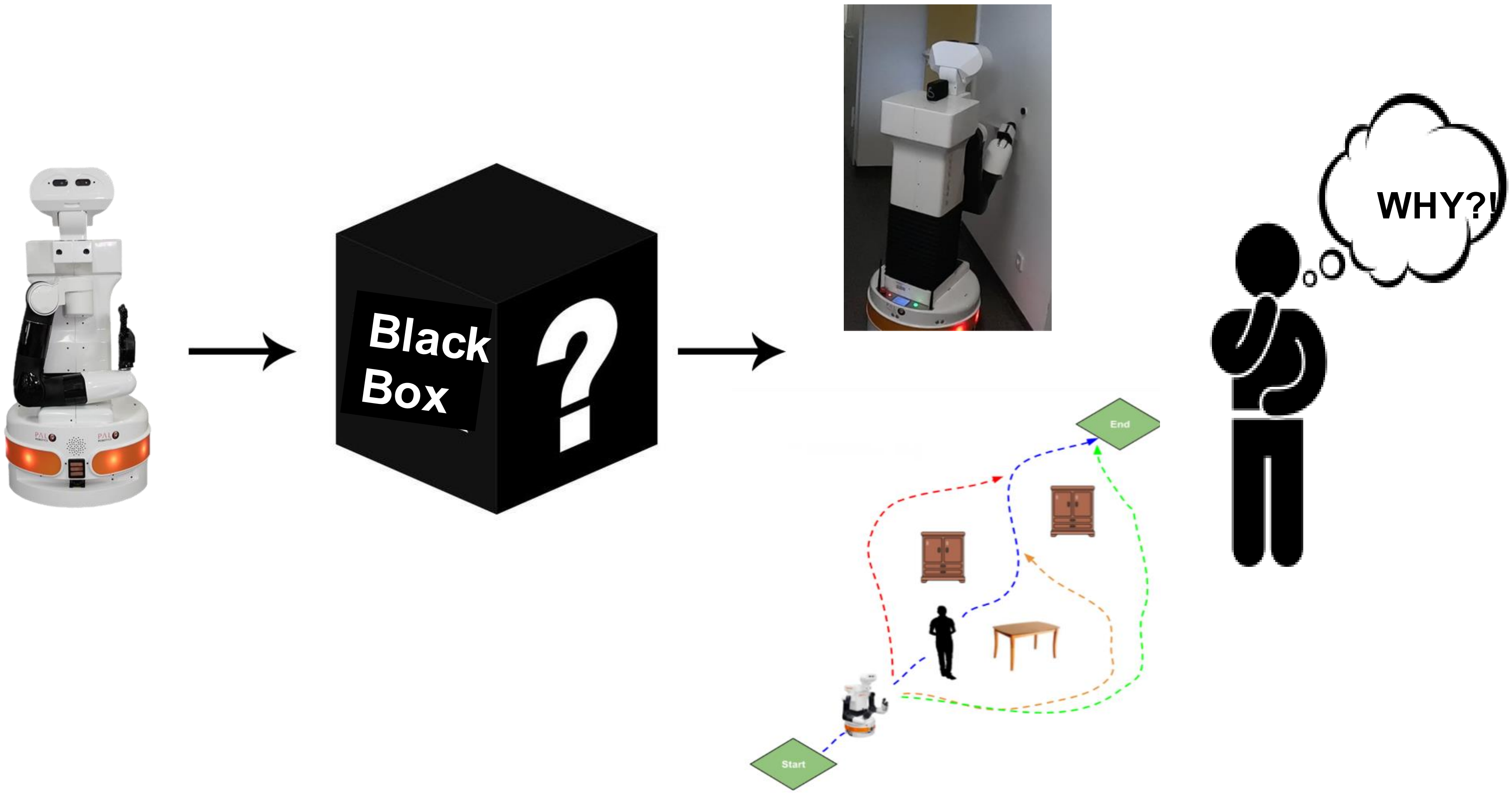


Brandao, Martim, et al. "Towards providing explanations for robot motion planning." 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021.

Motivation



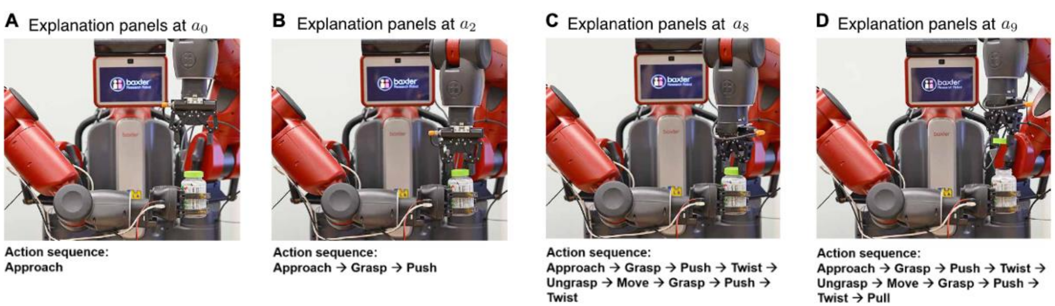
Motivation



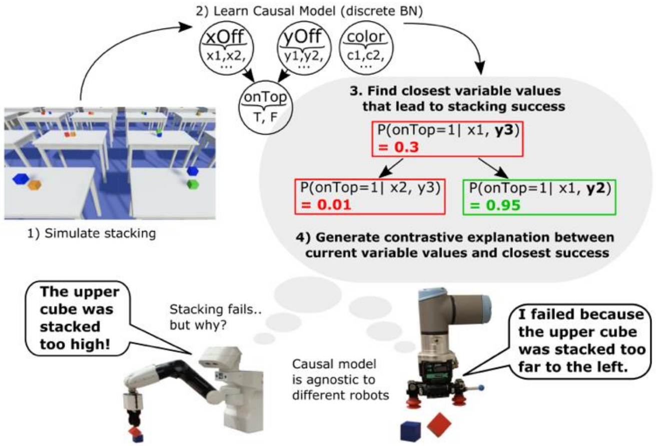
How can the robot use its environment to explain its navigational failures and decisions?

SOTA Overview

- Most of methods are constraint- or algorithm-based
- Focus on planner internals rather than environment
- Most methods explain task failures rather than planning failures
- Approaches focused on manipulation rather than navigation



Edmonds, M., Gao, F., Liu, H., Xie, X., Qi, S., Rothrock, B., ... & Zhu, S. C. (2019). A tale of two explanations: Enhancing human trust by explaining robot behavior. Science Robotics, 4(37), eaay4663.



Diehl, M., & Ramirez-Amaro, K. (2022). Why did i fail? a causal-based method to find explanations for robot failures. IEEE Robotics and Automation Letters, 7(4), 8925-8932.

Environment-centered explanations of robot navigation

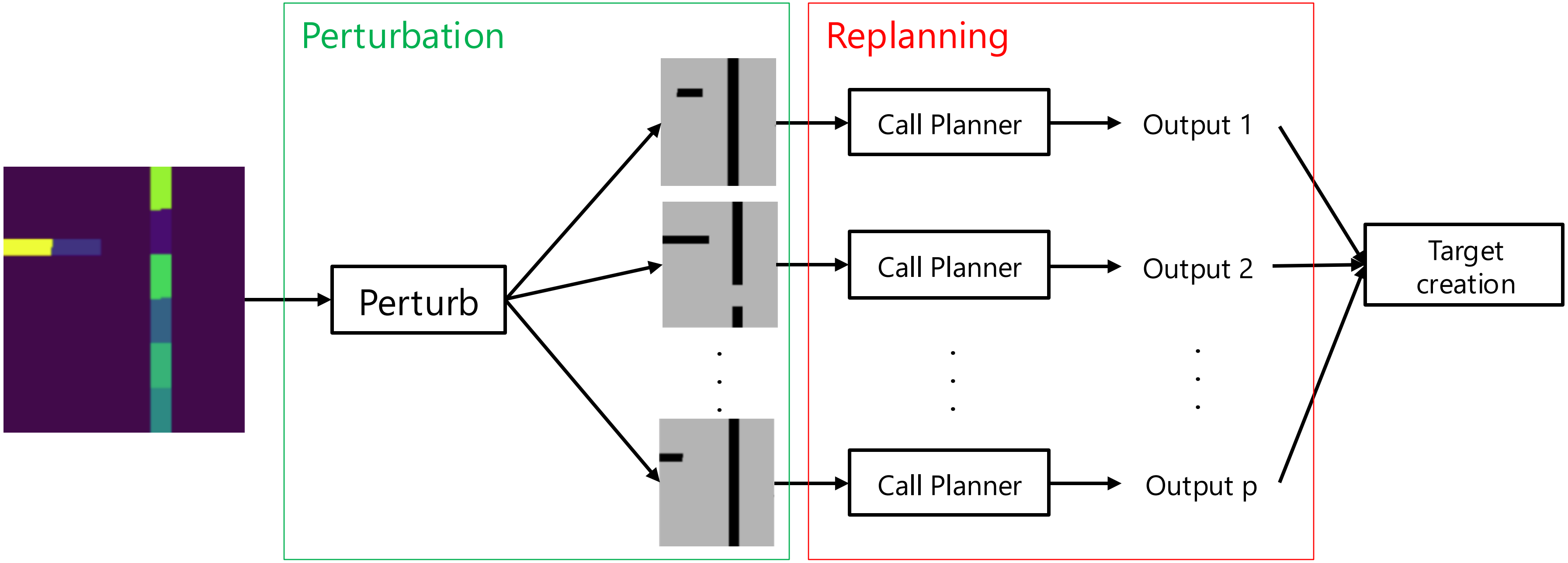
- Focus on environment, not robot internals
- Objects and their roles drive the explanation
- Replanning-based explanation generation

Environment clusterization (segmentation)

- Map (free space, obstacle) -> image representation
- Image clusterization (SLIC - Simple Linear Iterative Clustering)
- Removal of reduntant information (free space clusters)
- Manual clustering to the predefined number of segments

Halilovic & Lindner, Explaining local path plans using LIME, RAAD 2022

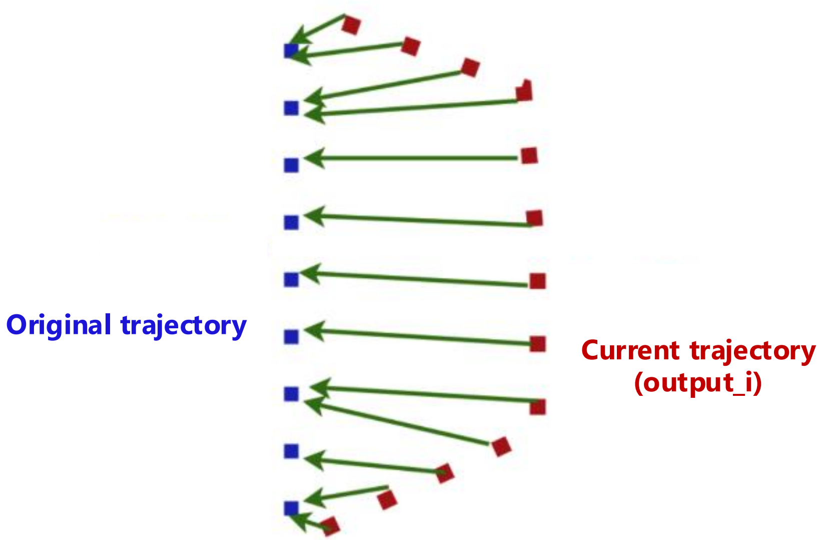
Environment perturbation and replanning

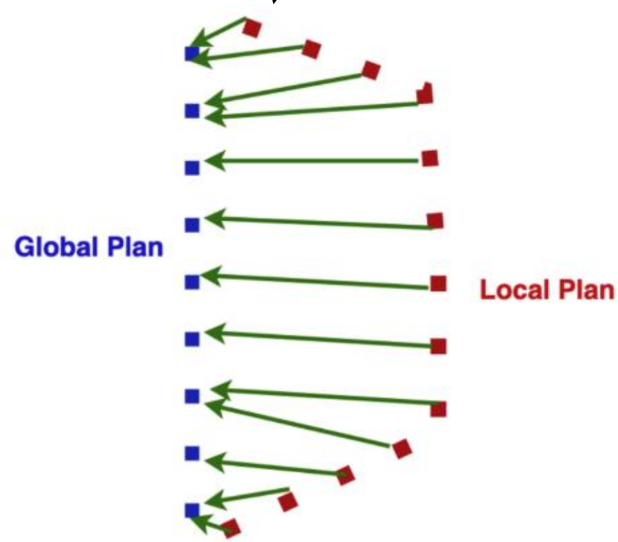


Halilovic & Lindner, Explaining local path plans using LIME, RAAD 2022

Target creation

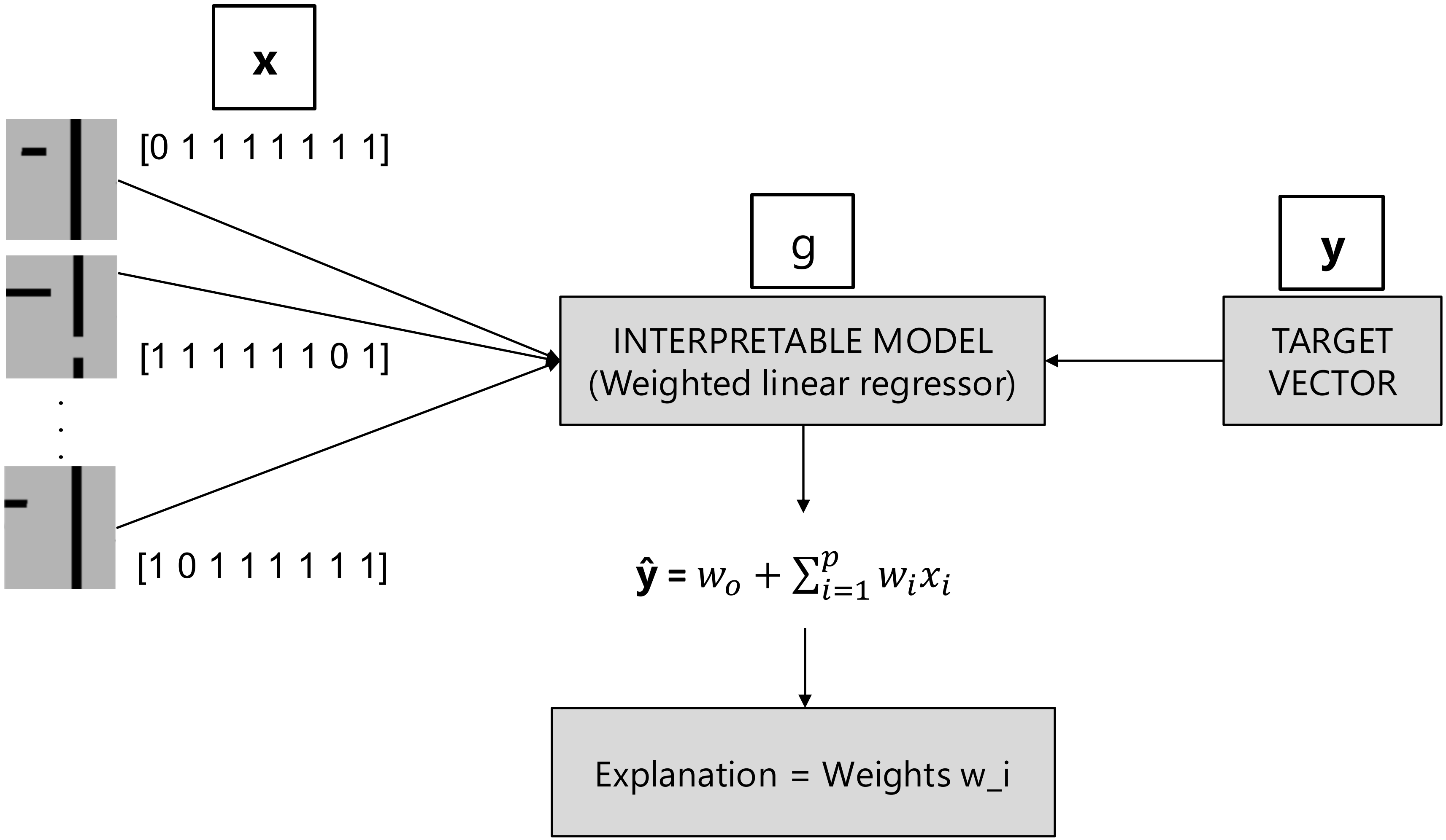
- For each perturbation
- L2 Euclidean distance between the current trajectory (output_i) from the original trajectory
- Original trajectory: from the same planner, from different planner





Halilovic & Lindner, Explaining local path plans using LIME, RAAD 2022

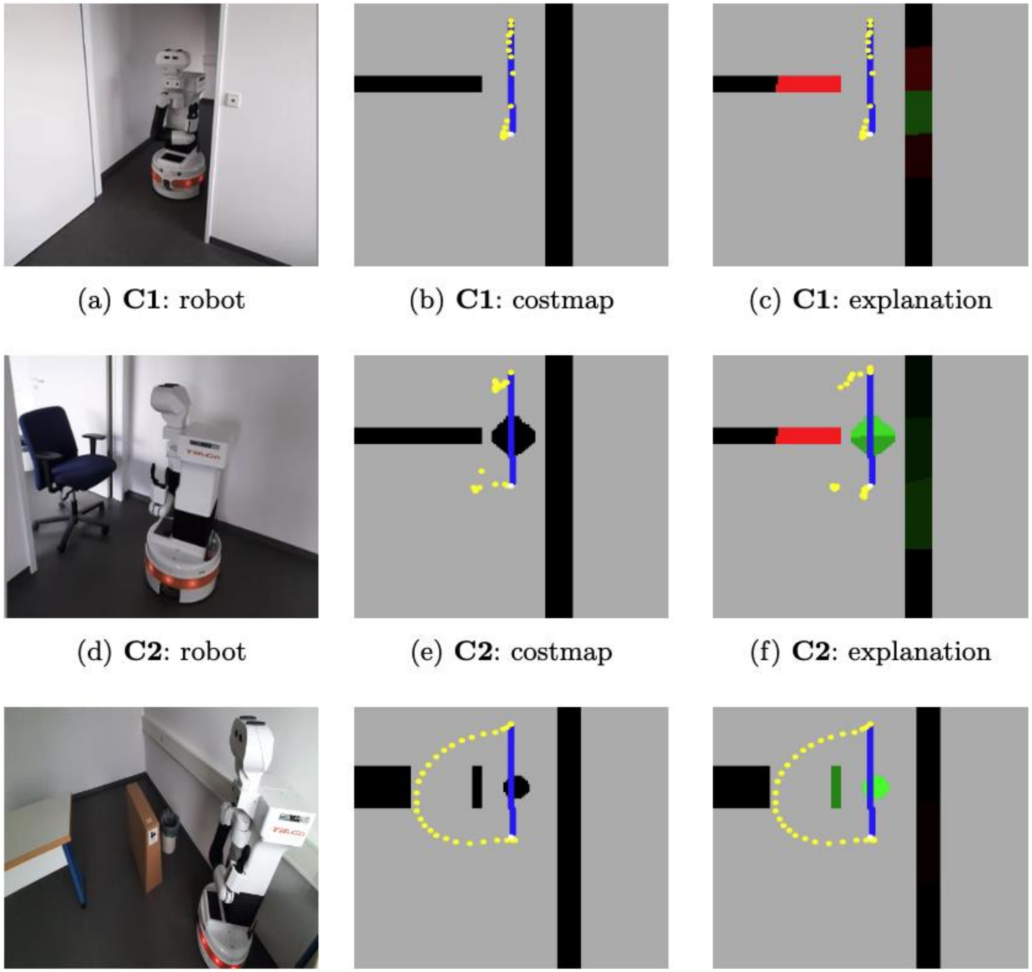
Explanation generation



Halilovic & Lindner, Explaining local path plans using LIME, RAAD 2022

Explanation Visualization

- Segment has positive weight
- Segment has negative weight



(a) **C1**: robot (b) **C1**: costmap (c) **C1**: explanation

(d) **C2**: robot (e) **C2**: costmap (f) **C2**: explanation

(g) **C3**: robot (h) **C3**: costmap (i) **C3**: explanation

Halilovic & Lindner, Explaining local path plans using LIME, RAAD 2022

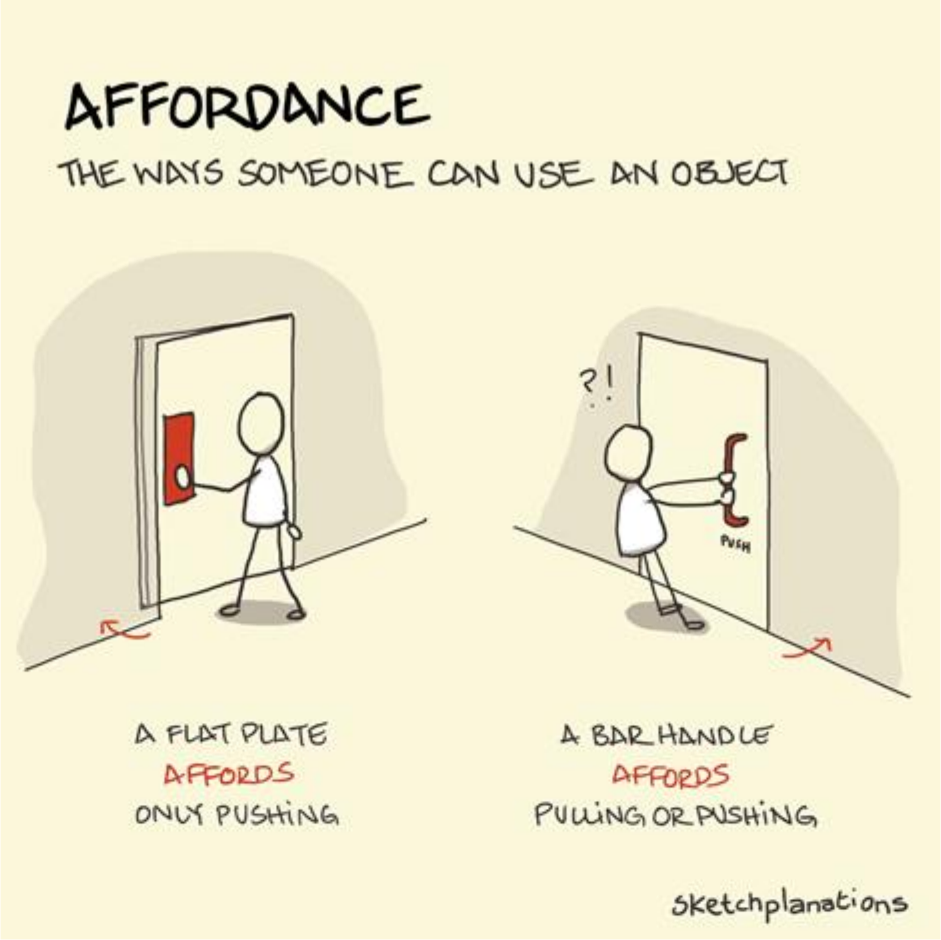
Semantics were missing!

Environment-centered explanations of robot navigation

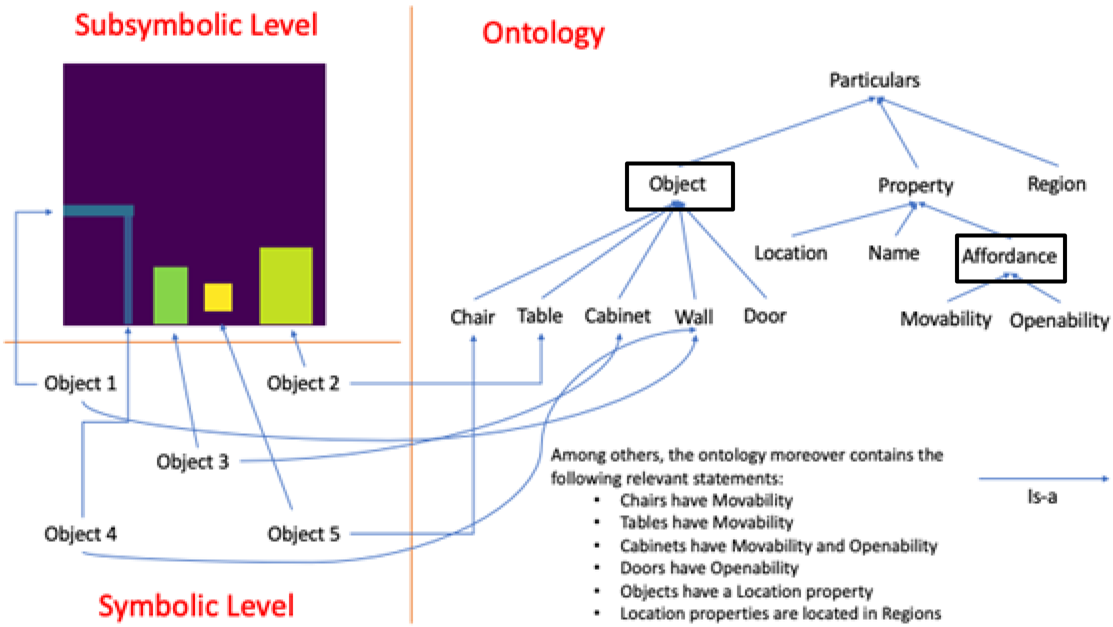
- Focus on environment, not robot internals
- Objects and their roles drive the explanation
- Perturbation-based explanation generation
- Ground explanations in human-understandable context

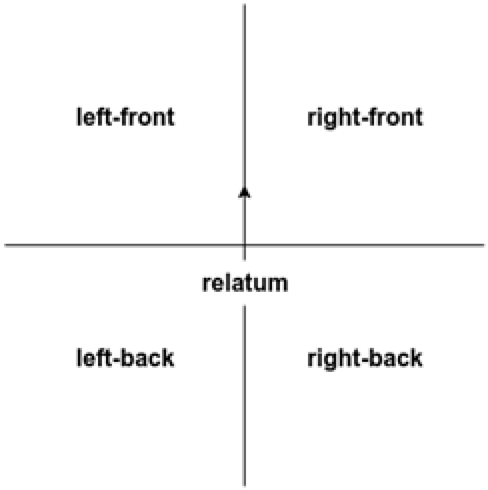
Affordances of objects

By Makito Nagawa



Affordance-based Ontology

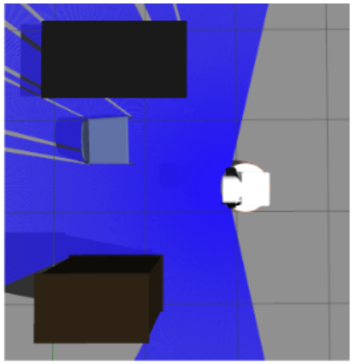




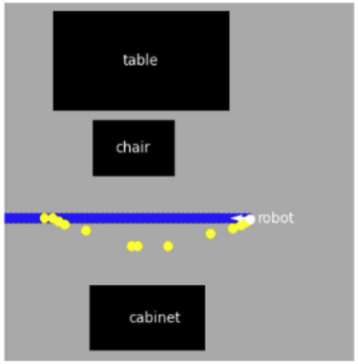
Object-affordance pair

Halilovic & Lindner, Visuo-textual explanations of a robot's navigational choices, HRI LBR 2023.

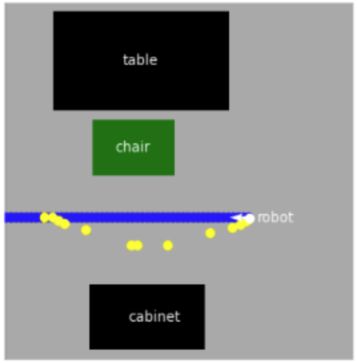
Visual-textual environment-centered explanations



Simulation scenario



Semantic map



Chair-movability explanation map (initial state: in the robot's neighborhood)

Chair-movability textual explanation: *"Because of the chair right-front of me, I deviate from the initial plan."*

Chair-movability textual suggestion: *"Dear human, please move the chair, so I proceed more smoothly."*

Halilovic & Lindner, Visuo-textual explanations of a robot's navigational choices, HRI LBR 2023.

Visual-textual environment-centered explanations

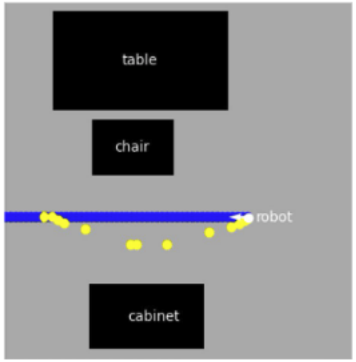
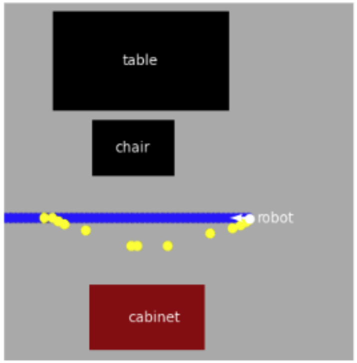
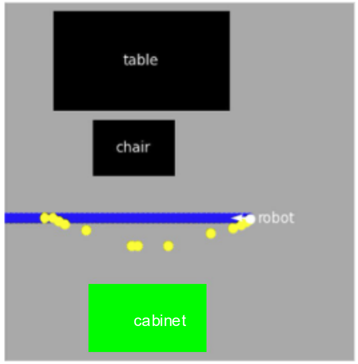


Table-movability explanation map
(initial state: in the robot's neighborhood)



Cabinet-movability explanation map
(initial state: in the robot's neighborhood)



Cabinet-openability explanation map
(initial state: closed)

Cabinet-movability textual explanation: "If the cabinet left-front of me was not there, I would deviate more from the initial plan."

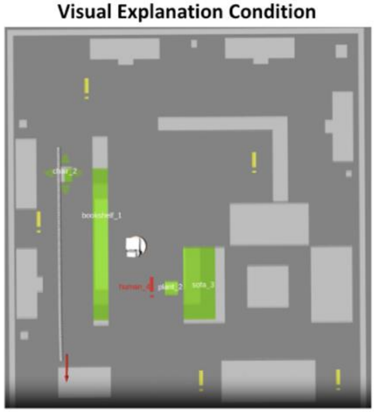
Cabinet-openability textual explanation: "If the cabinet left-front of me was open, I would deviate less from the initial plan."

Halilovic & Lindner, Visuo-textual explanations of a robot's navigational choices, HRI LBR 2023.

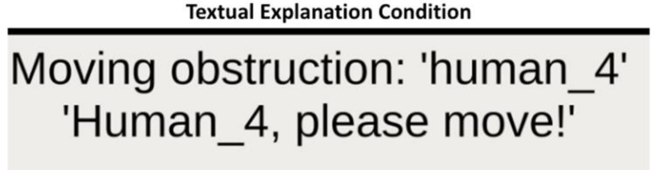
Explanation modality

- Visual and textual modalities explored
- First user study on most preferred color scheme for visual explanation
- Second user study on satisfaction with different explanation modalities (visual vs. textual vs. visual-textual)
- People are more satisfied multimodal (visual-textual) over unimodal (visual, textual) explanations

Visual Explanation Condition



Textual Explanation Condition

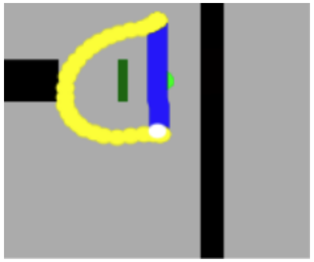


Hybrid (visual-textual) Explanation Condition

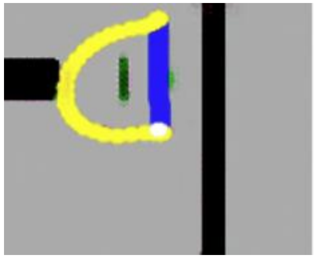
Halilovic et al., Exploring the impact of explanation representation on user satisfaction in robot navigation, TAHRI 2024.

Visual explanations with generative AI

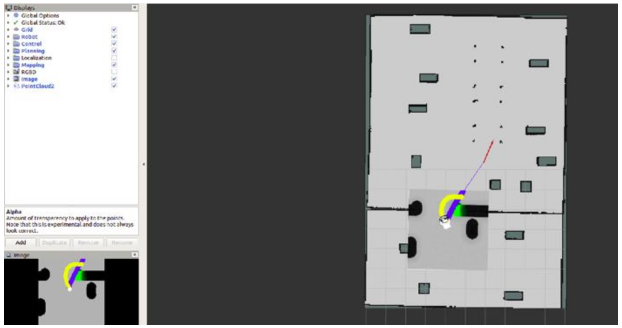
- Visual explanation generation with generative AI (GAN - Generative Adversarial Networks, image-to-image translation)
- Requires dataset created with replanning
- Faster than replanning (4 Hz), but the explanation quality is lower
- Not that scalable
- Real-time* visual explanation layer



Replanning



GAN



Visual explanation layer

Halilovic & Krivic, Towards Fast Visual Explanations of Local Path Planning with LIME and GAN. HI-AI@ KDD 2024.

Different textual explanations

- Descriptive, suggestive and counterfactual textual explanations
- Affordance-based verbalizer
- Robot librarian failure example

<u>Descriptive explanation</u>

"I failed to fetch a book, because the chair and the **closed** cabinet both **in front of** me blocked my path"

<u>Suggestive explanation</u>

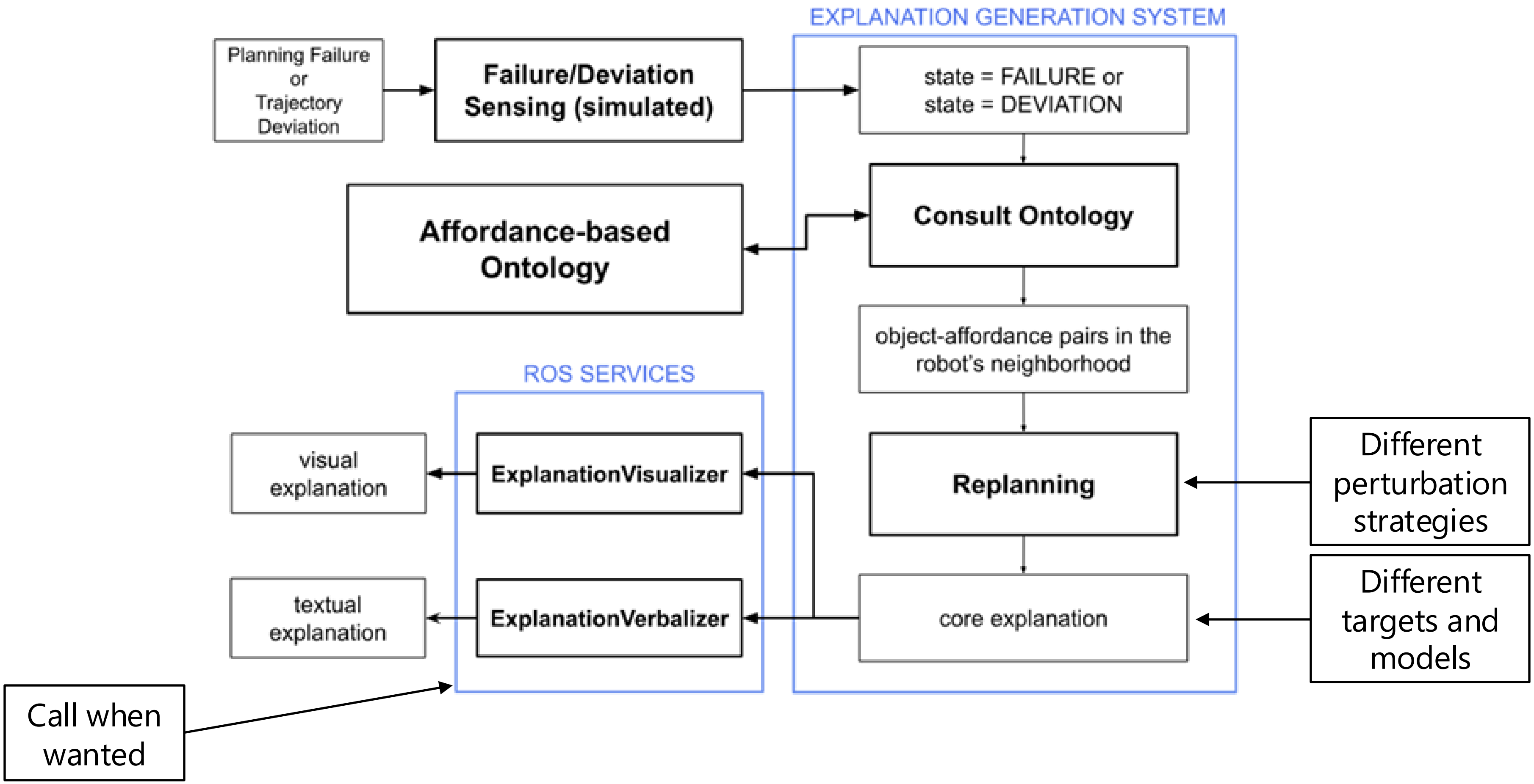
"Dear human, please **move** the chair and **open** the cabinet, both **in front of** me, so I can fetch a book."

<u>Counterfactual explanation</u>

"If the chair **in front of** me was **not there** and the cabinet **in front of** me was **open**, I would not fail to fetch the book."

Halilovic & Krivic, Affordance-Based Explanations of Robot Navigation, ICRA 2025.

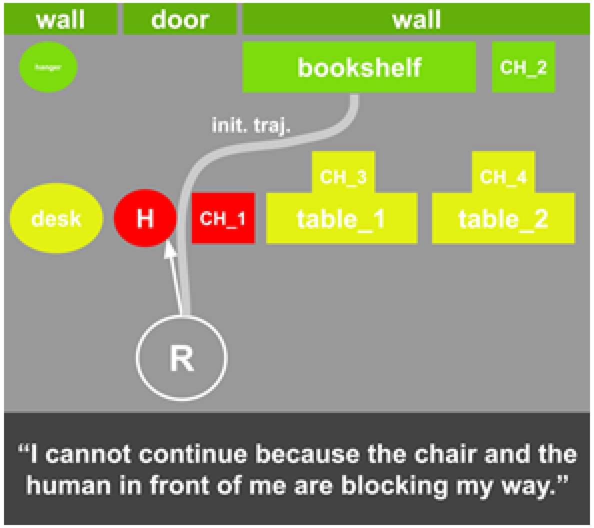
Explanation generation framework



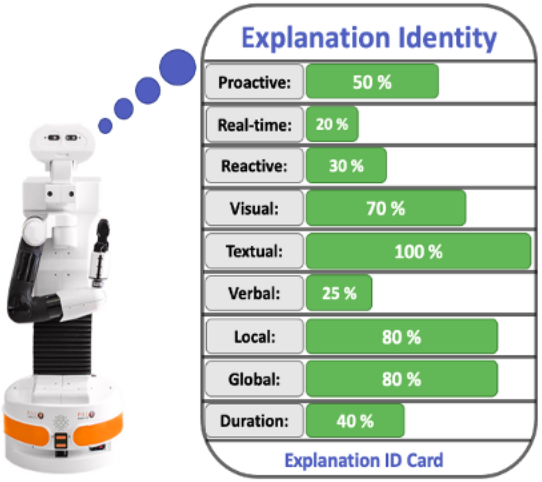
Halilovic & Krivic, Affordance-Based Explanations of Robot Navigation, ICRA 2025.

Research challenges

Environment-centered explanation generation for robot navigation

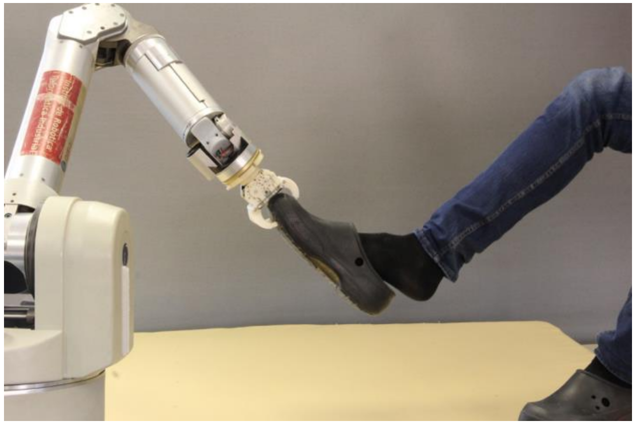


Human-centered explanation planning for robot navigation

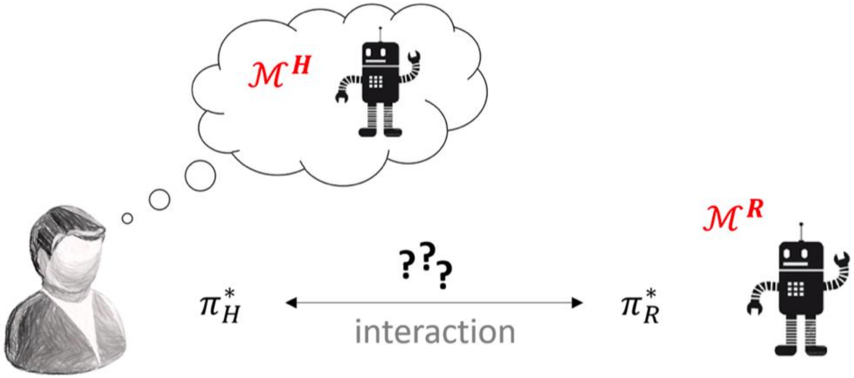


SOTA Overview

- Explainable AI Planning (XAIP) – explainable planning of robot actions, but not planning of robot explanations
- Preference-driven assistive and social robotics, but not explainable robotics
- Explanations through model reconciliation



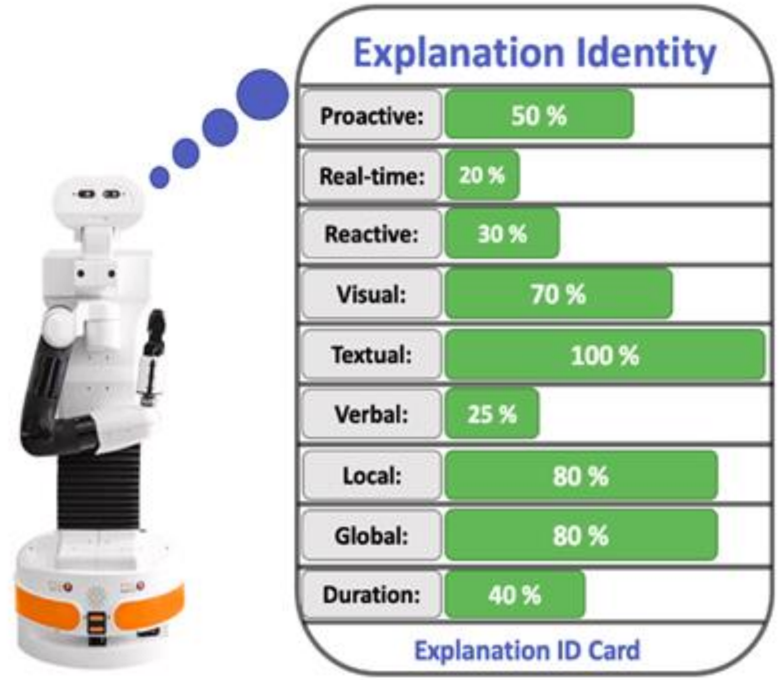
Canal, G., Alenyà, G., & Torras, C. (2019). Adapting robot task planning to user preferences: an assistive shoe dressing example. Autonomous Robots, 43(6), 1343-1356.



Chakraborti, T., Sreedharan, S., Zhang, Y., & Kambhampati, S. (2017). Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. arXiv preprint arXiv:1701.08317.

Robot explanation identity

- Unique identity characteristics
- Adaptive
- Contextual
- Personalized
- Multimodal
- Probabilistic
- “Good explanations require additional knowledge represented as preferences over explanations” [1]



Halilovic & Krivic, Robot Explanation Identity, 2024.

[1] Sohrabi, S., Baier, J., & McIlraith, S. (2011, August). Preferred explanations: Theory and generation via planning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 25, No. 1, pp. 261-267).

Key questions identified

What to explain?

- Failure
- Deviation
- Path optimality

When to explain?

- Every time step
- When human is nearby
- After a question
- Proactive, reactive, post-hoc

How to explain?

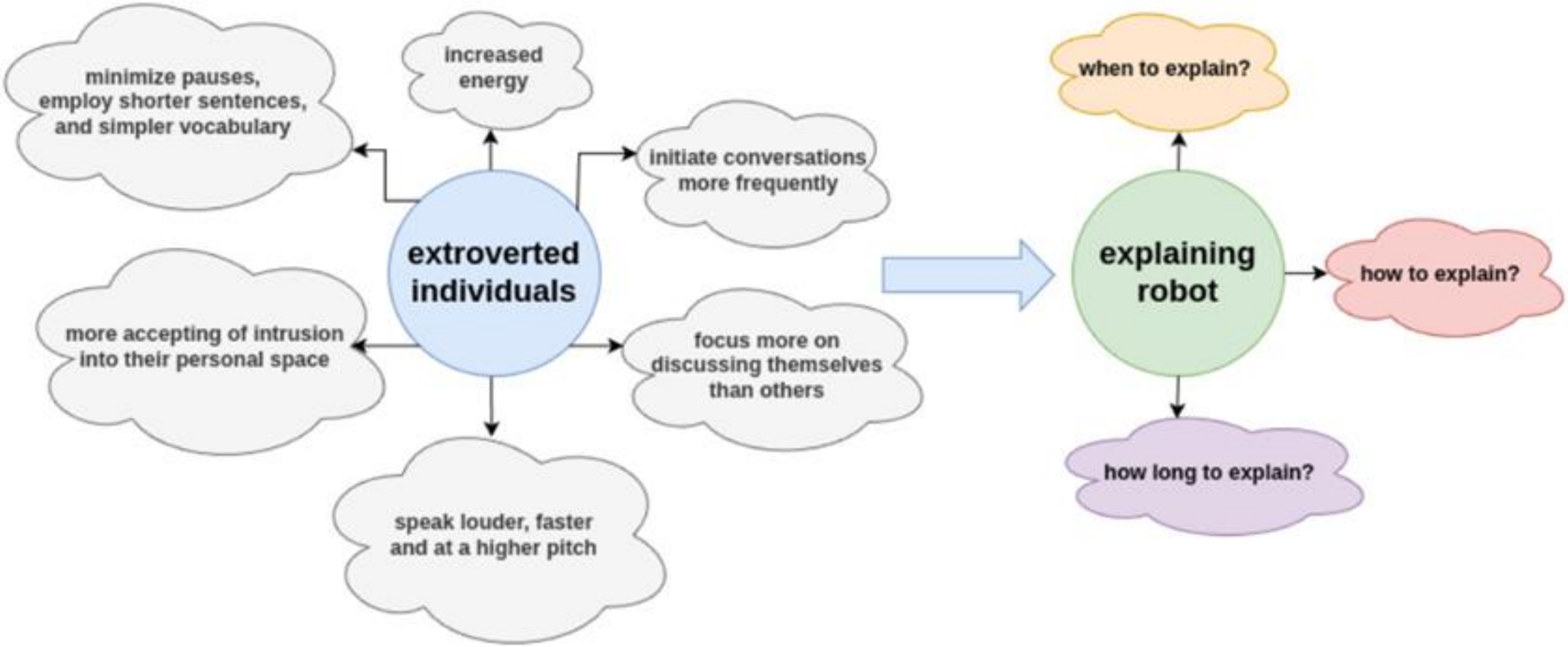
- Modality
- Scope

How long to explain?

- Until the action is finished
- Until human stops asking
- After a predefined interval

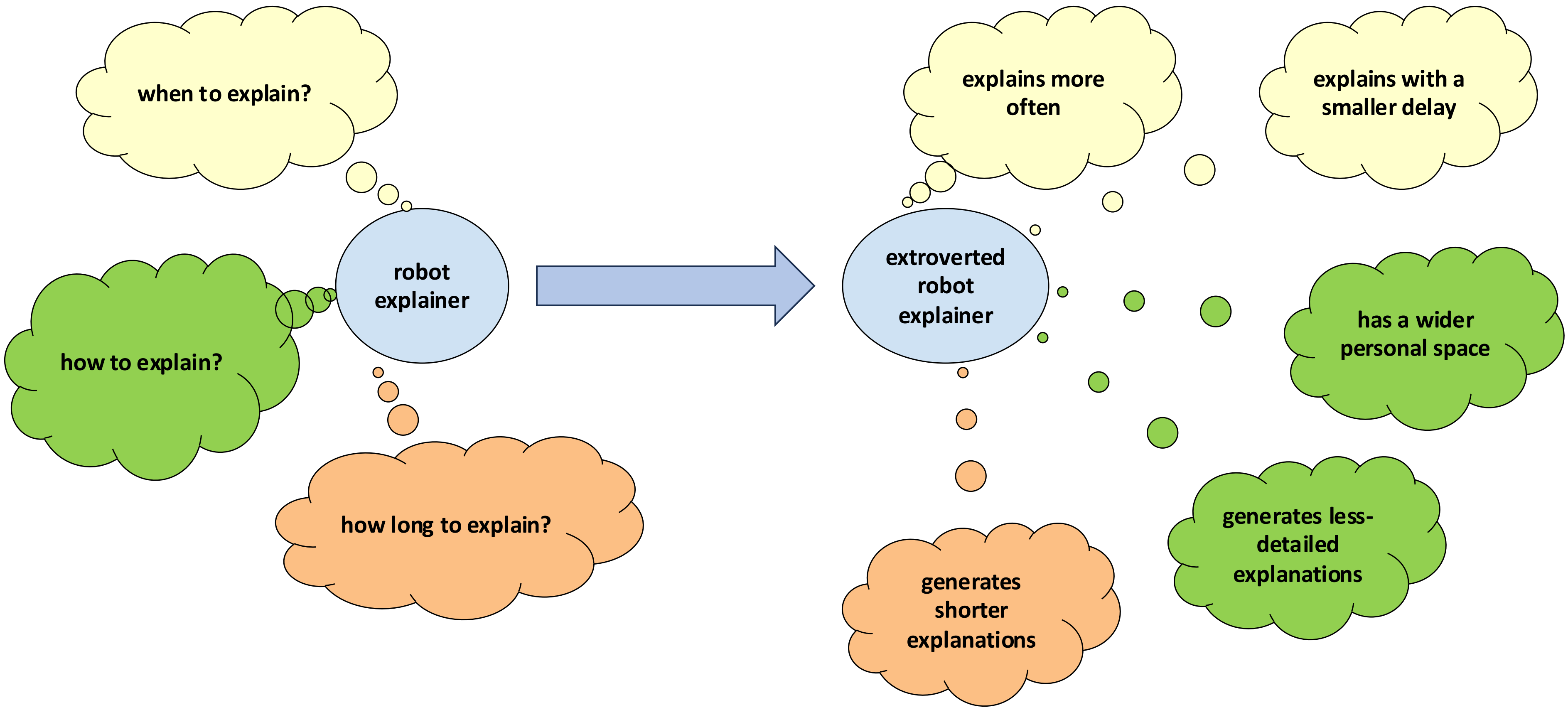
Halilovic & Krivic, Towards a Holistic Framework for Explainable Robot Navigation, HFR 2023.

From human to robot personality



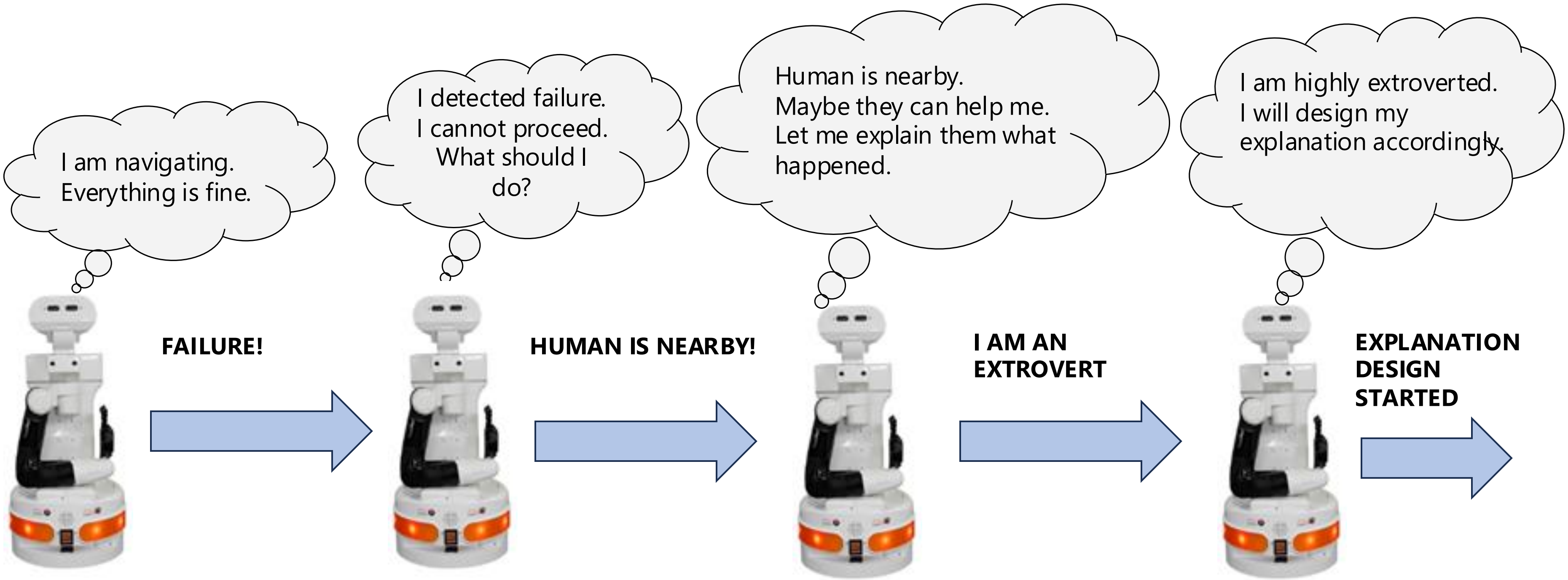
Halilovic & Krivic, The influence of a robot's personality on real-time explanations of its navigation, ICSR 2023.

From human to robot personality



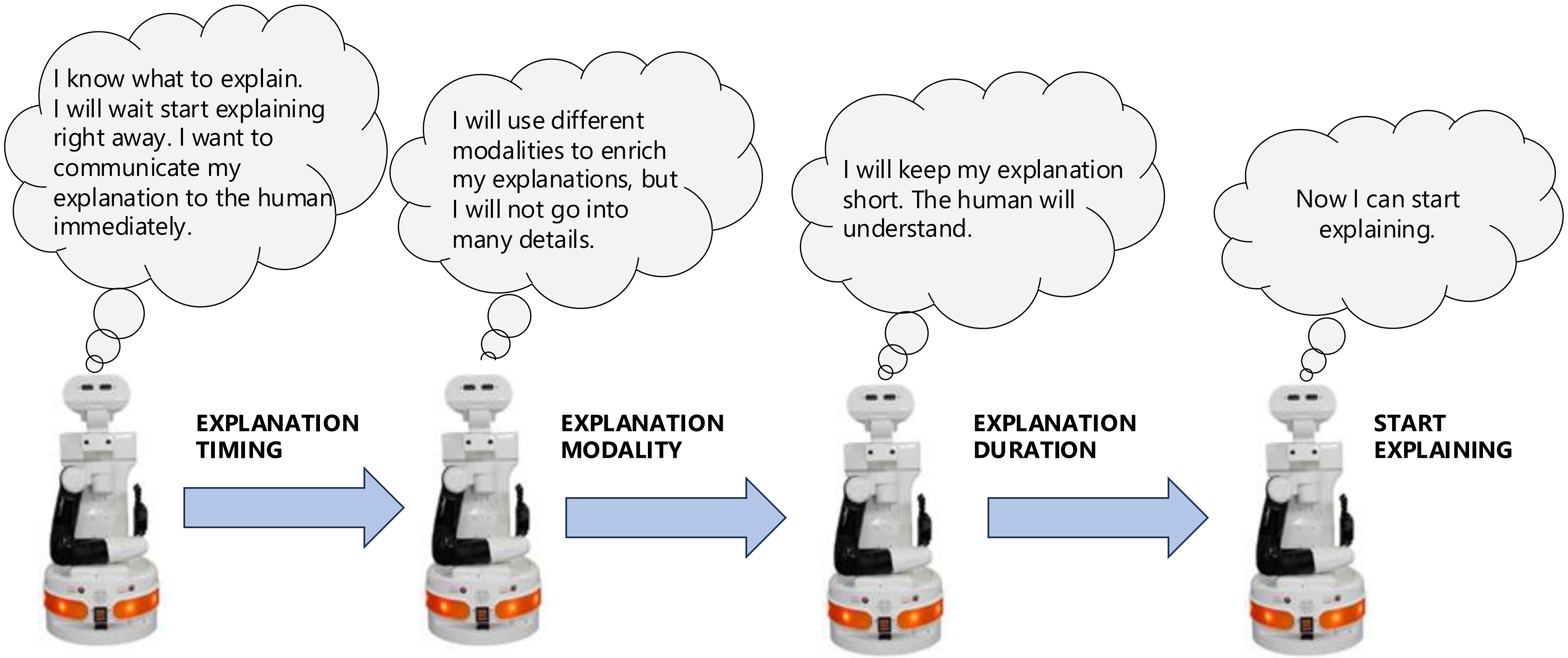
Halilovic & Krivic, The influence of a robot's personality on real-time explanations of its navigation, ICSR 2023.

Planning of explanations



Halilovic & Krivic, Planning of explanations for robot navigation, ICRA 2024.

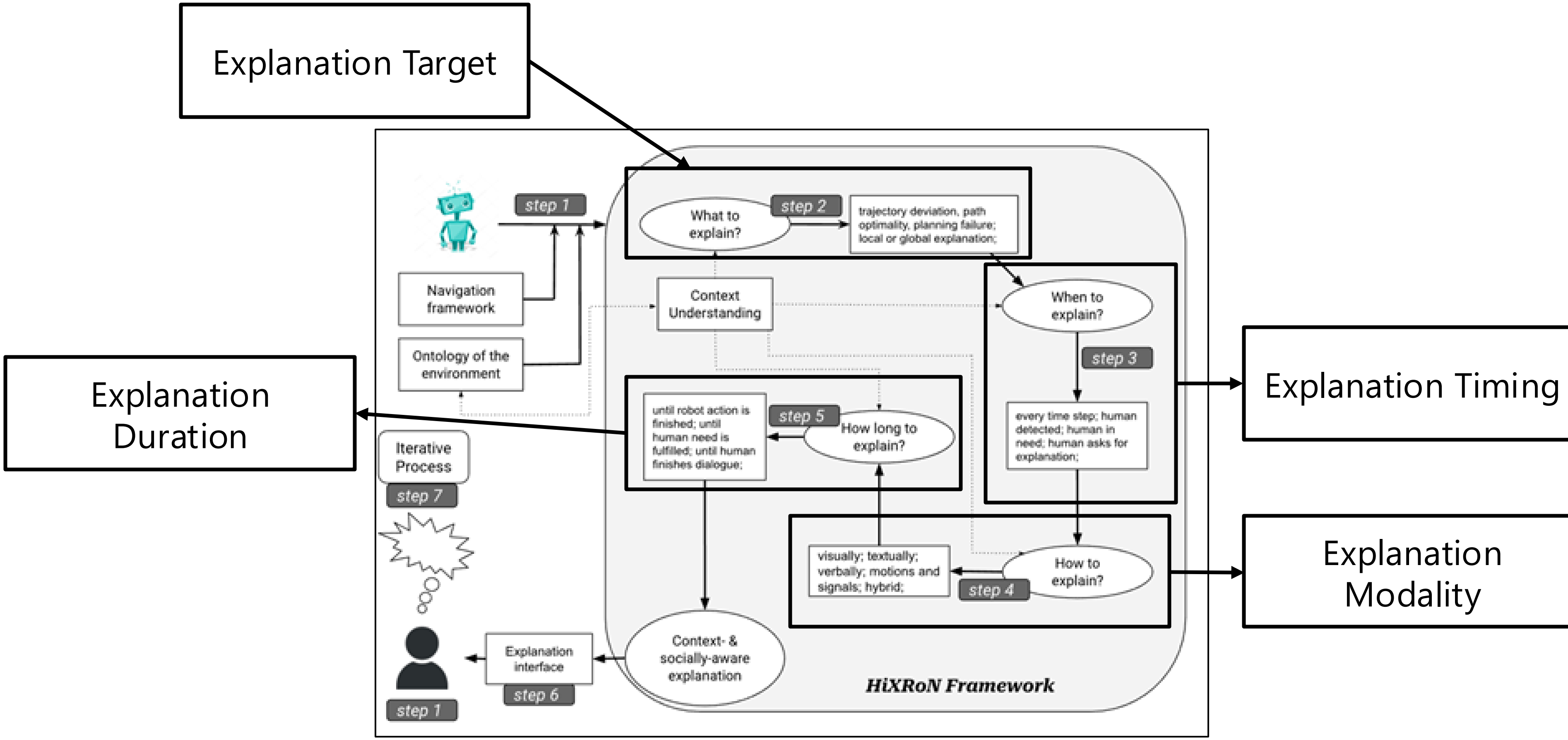
Planning of explanations



EXPLANATIONS SHOULD BE PLANNED!

Halilovic & Krivic, Planning of explanations for robot navigation, ICRA 2024.

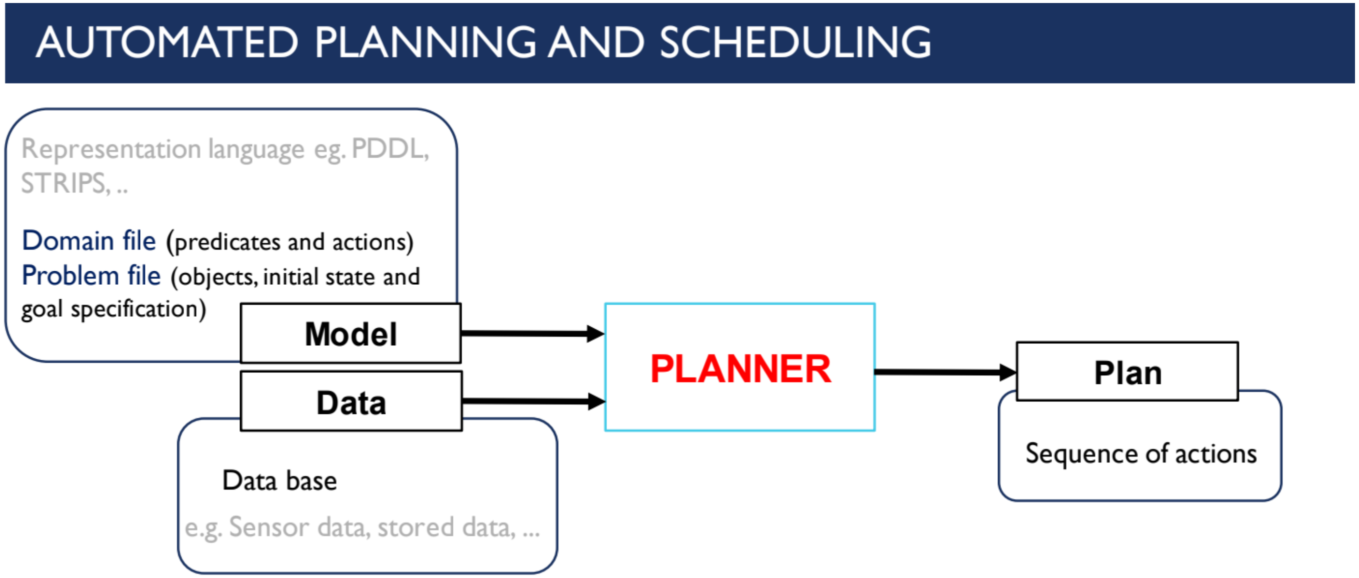
Framework for explanation planning



Halilovic & Krivic, Towards a Holistic Framework for Explainable Robot Navigation, HFR 2023.

Planning of explanations

- Planning explanations along with other robot actions.
- The explanation occurrence can vary based on parameters such as user preference or other task priorities.

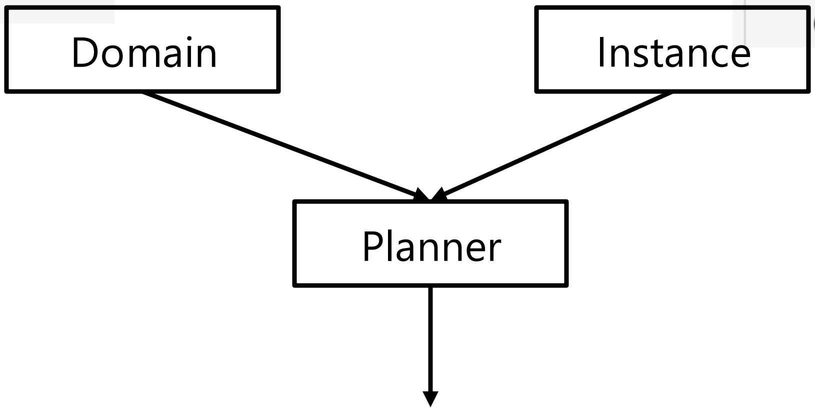


Halilovic & Krivic, Planning of explanations for robot navigation, ICRA 2024.

Deterministic Planning

```
(:durative-action explain_failure
    :parameters (?r - robot, ?h - human, ?f - failure)
    :duration (= ?duration (expl_duration ?r))
    :condition (and
        (at start (failure_detected ?r ?f))
        (at start (human_detected ?h))
        )
    :effect (and
        (at end (explained_failure ?r ?f))
    )
)
```

```
(:init
    (is_extrovert tiago)
    (= (expl_duration tiago) 1)
    (navigating tiago)
    (at_place chair)
    (is_not_detected amar)
```




```
0.000: (detect_failure tiago)   [1.000]
1.000: (detect_human tiago explainee)   [1.000]
2.000: (explain_failure_start tiago explainee) [1.000]
3.000: (calculate_explanation_timing tiago) [1.000]
4.000: (pick_visual_textual_representation tiago explainee) [1.000]
5.000: (calculate_explanation_duration tiago) [1.000]
6.000: (wait tiago)   [1.000]
7.000: (explain_failure tiago explainee) [1.000]
8.000: (explain_failure_ended tiago explainee)   [1.000]
```

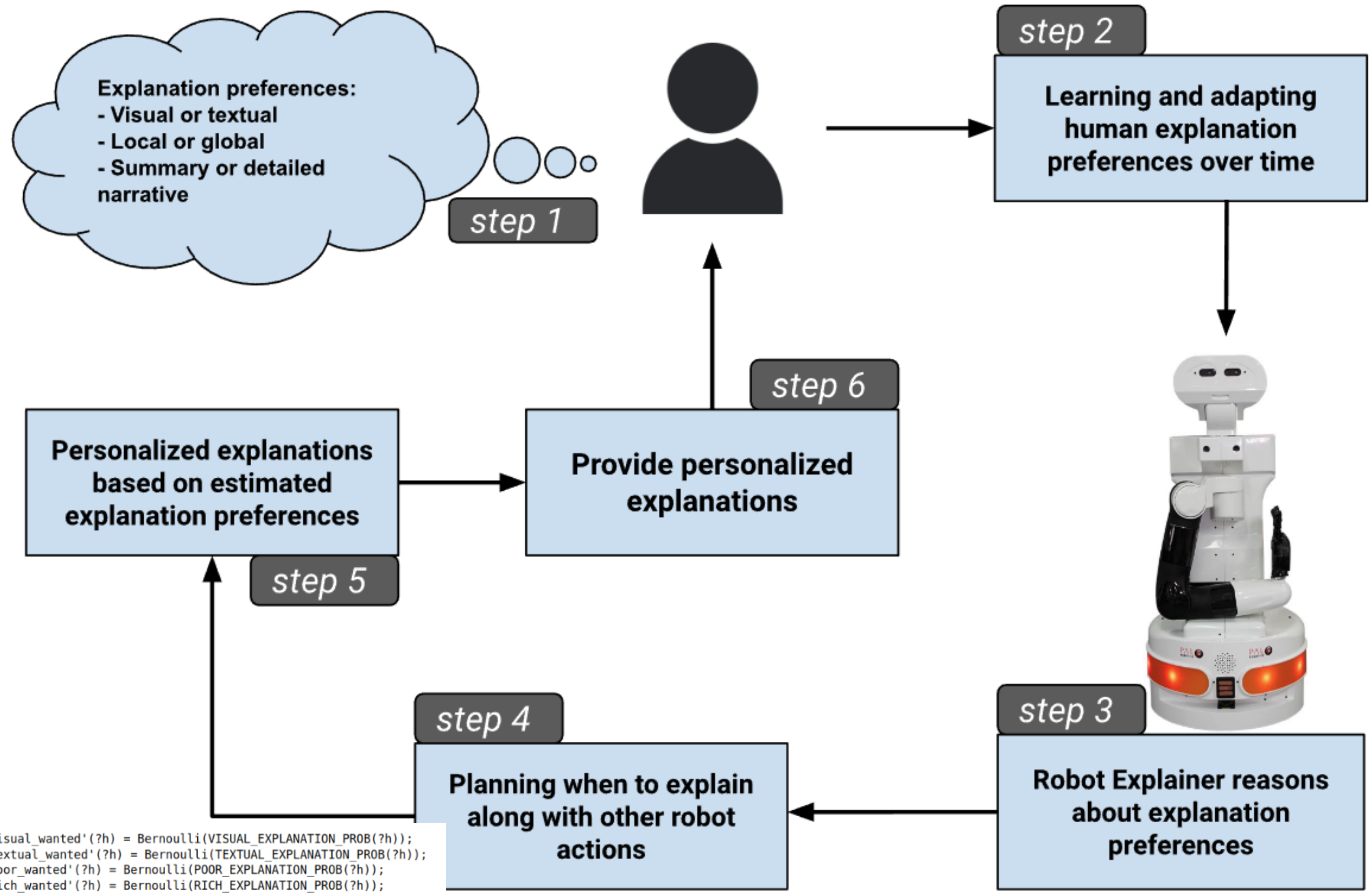
(a) Explanation plan of a totally extroverted robot

```
0.000: (detect_failure tiago)   [1.000]
1.000: (detect_human tiago explainee)   [1.000]
2.000: (explain_failure_start tiago explainee) [1.000]
3.000: (calculate_explanation_timing tiago) [1.000]
4.000: (pick_visual_representation tiago explainee) [1.000]
5.000: (calculate_explanation_duration tiago) [1.000]
6.000: (wait tiago)   [11.000]
17.000: (explain_failure tiago explainee) [11.000]
28.000: (explain_failure_ended tiago explainee)   [1.000]
```

(b) Explanation plan of a totally introverted robot

Halilovic & Krivic, Planning of explanations for robot navigation, ICRA 2024.

Probabilistic Planning

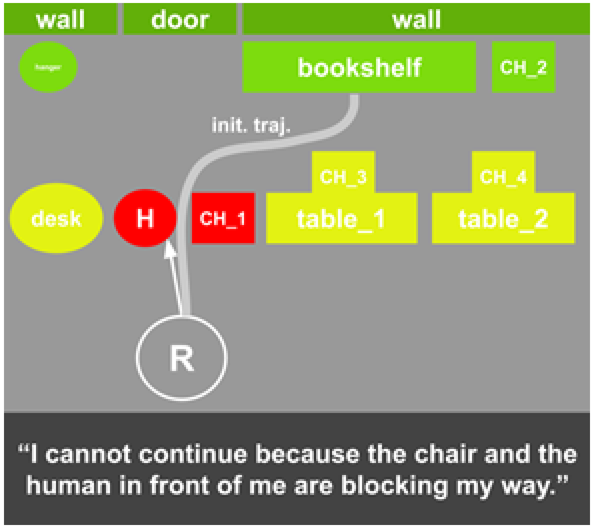



```
explanation_visual_wanted'(?h) = Bernoulli(VISUAL_EXPLANATION_PROB(?h));
explanation_textual_wanted'(?h) = Bernoulli(TEXTUAL_EXPLANATION_PROB(?h));
explanation_poor_wanted'(?h) = Bernoulli(POOR_EXPLANATION_PROB(?h));
explanation_rich_wanted'(?h) = Bernoulli(RICH_EXPLANATION_PROB(?h));
explanation_short_wanted'(?h) = Bernoulli(SHORT_EXPLANATION_PROB(?h));
explanation_long_wanted'(?h) = Bernoulli(LONG_EXPLANATION_PROB(?h));
explanation_local_wanted'(?h) = Bernoulli(LOCAL_EXPLANATION_PROB(?h));
explanation_global_wanted'(?h) = Bernoulli(GLOBAL_EXPLANATION_PROB(?h));
```

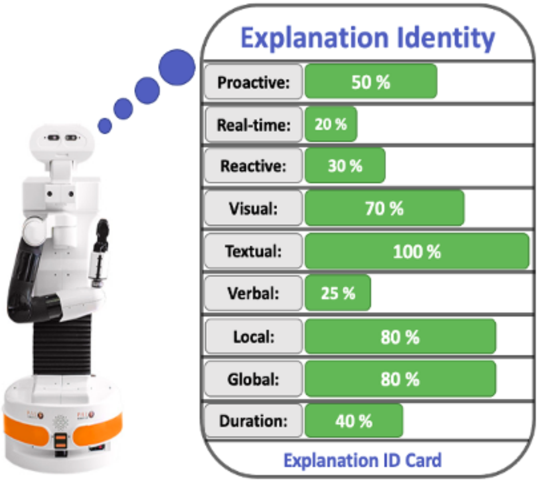
Halilovic et al., Towards Probabilistic Planning of Explanations for Robot Navigation, 2024.

Research Contributions

Environment-centered explanation generation for robot navigation



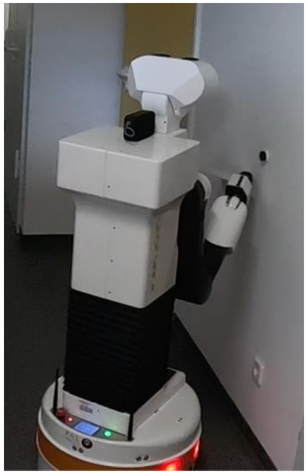
Human-centered explanation planning for robot navigation



Conclusions – Research Contributions

- Developed a framework for environment-centered explanations
- Showed that environment context (obstacles, affordances, spatial relations) and explanation representation is important for satisfiable explanations.
- Have developed a framework human-centered explanation planning
- Incorporated human explanation preferences into explanation generation to improve relevance, clarity and satisfaction.

“When robots stumble, explanations can help keep humans on their side.”



Past Wins – Current Battles – Future Conquests

- Environment-centered explanations:
  - Replanning- and affordance-based framework for explanation generation
  - The impact of explanation modality and representation on user satisfaction
  - Generalizability to other domains and models (50%)
  - LLMs for explanation verbalization (20%)
  - Richer set of affordances; VLMs for scene understanding and ontology creation
- Human-centered explanations:
  - Framework for explanation planning
  - Preference-based deterministic explanation planning
  - The role of robot personality (extroversion) on explanations
  - Preference-based probabilistic explanation planning with preference learning (40%)
  - Different explanation planning timing strategies (50%)
  - Development of robot explanation identity

Conference Publications

1. **Halilovic, A.**, & Lindner, F. (2022). Explaining local path plans using LIME. In International Conference on Robotics in Alpe-Adria Danube Region (pp. 106-113). Cham: Springer International Publishing.
2. **Halilovic, A.**, & Krivic, S. (2023). 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 Singapore.
3. **Halilovic, A.**, & Krivic, S. (2024). Planning of explanations for robot navigation. In 2024 IEEE International Conference on Robotics and Automation (ICRA) (pp. 5478-5484). IEEE.
4. **Halilovic, A.**, & Krivic, S. (2025). Affordance-Based Explanations of Robot Navigation. To be published in 2025 IEEE International Conference on Robotics and Automation (ICRA). IEEE.

Other Peer-Reviewed Publications

1. **Halilovic, A.**, & Lindner, F. (2023). Visuo-textual explanations of a robot's navigational choices. In Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (pp. 531-535).
2. **Halilovic, A.**, & Krivic, S. (2023). Towards a Holistic Framework for Explainable Robot Navigation. In International Workshop on Human-Friendly Robotics (pp. 213-228). Cham: Springer Nature Switzerland.
3. **Halilovic, A.**, Chandrayan, V., & Krivic, S. (2024). Exploring the impact of explanation representation on user satisfaction in robot navigation. In Proceedings of the 2024 International Symposium on Technological Advances in Human-Robot Interaction (pp. 1-9).
4. **Halilovic, A.**, & Krivic, S. (2024). Robot Explanation Identity. arXiv preprint arXiv:2405.13841.
5. **Halilovic, A.**, Krivić, S., & Canal, G. (2024). Towards Probabilistic Planning of Explanations for Robot Navigation. In RSS 2024 Workshop on Unsolved Problems in Social Robot Navigation.
6. **Halilovic, A.**, & Krivic, S. (2024). Towards Fast Visual Explanations of Local Path Planning with LIME and GAN. In HI-AI@ KDD.