From Explanations to Trust

Human-centred approach AI research

Senka Krivić, PhD
Assistant Professor
Faculty of Electrical Engineering
University of Sarajevo

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

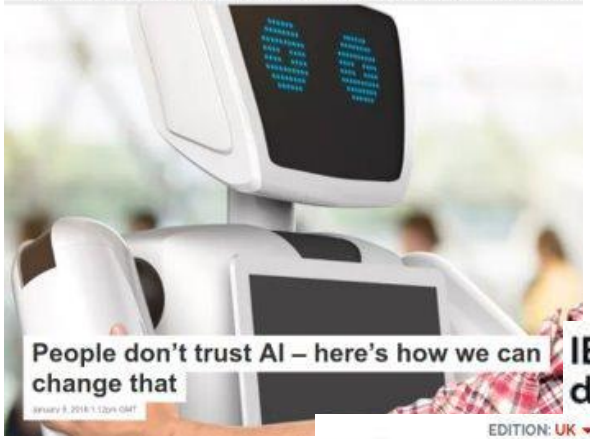
GDPR and Other Regulations Demand Explainable AI

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

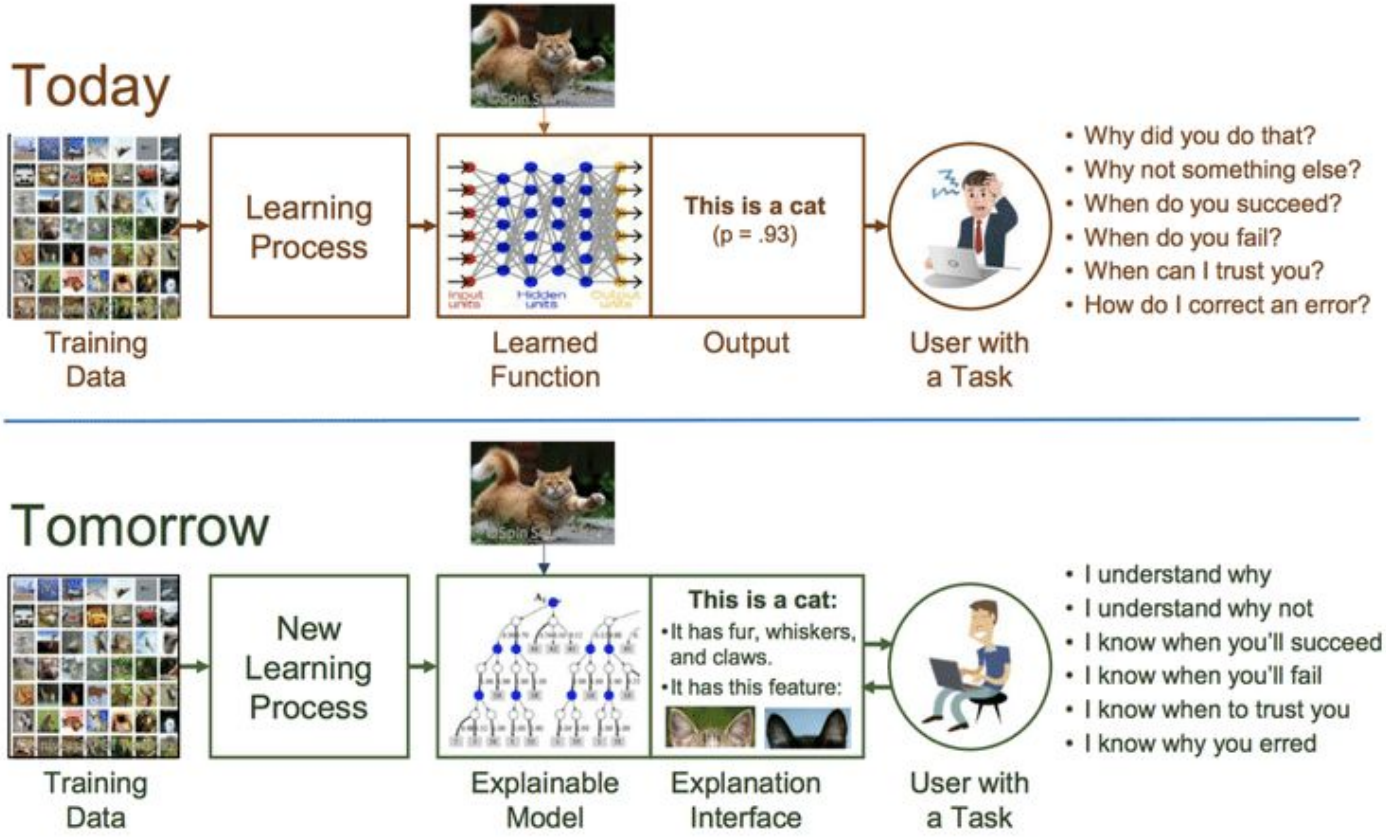
other technologies and industries, artificial intelligence will need to adopt supplier's declaration of conformity uments to build trust. How was that model built exactly?

EDITION: UK

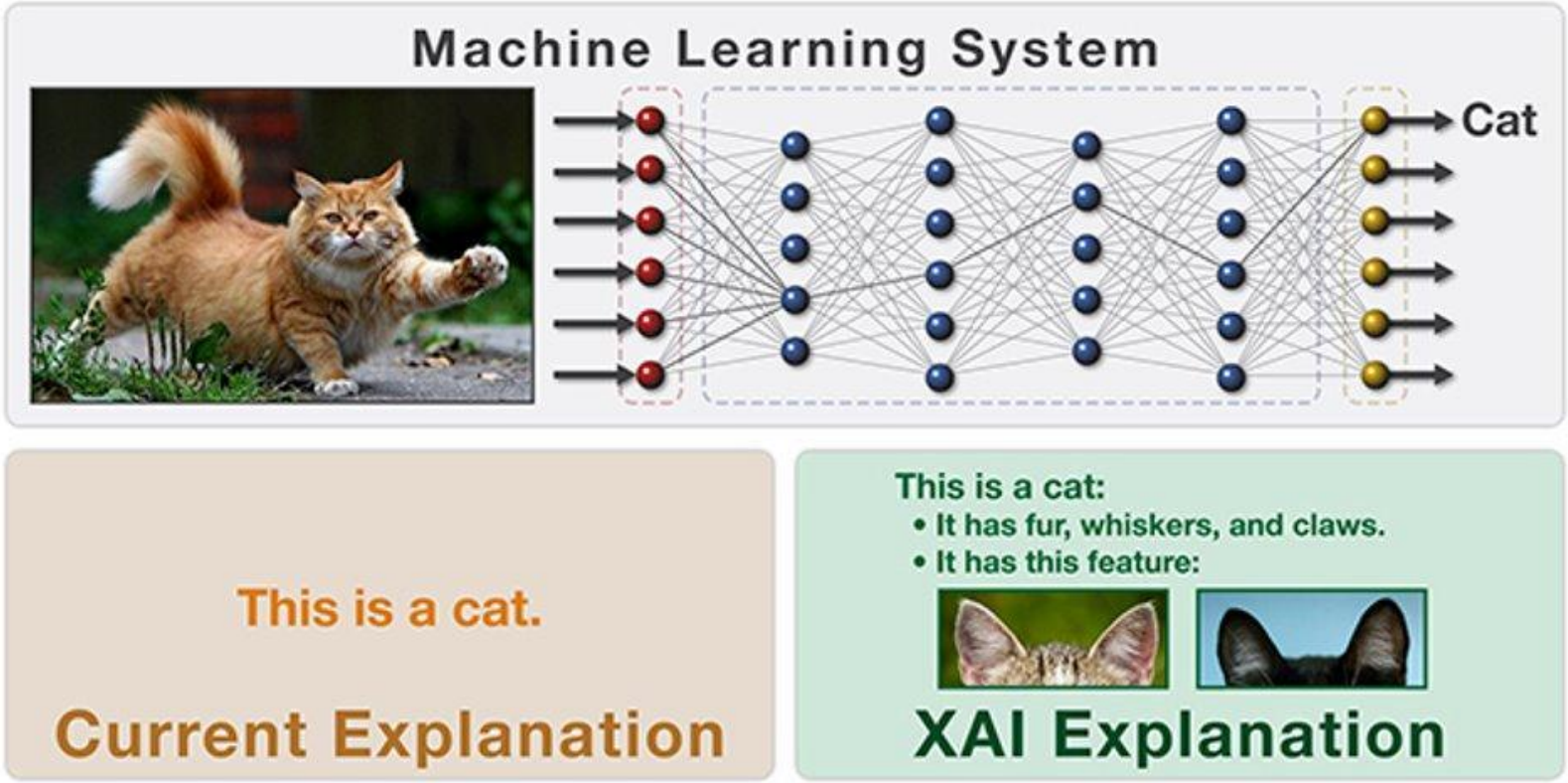
ZDNet

Explainable AI

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

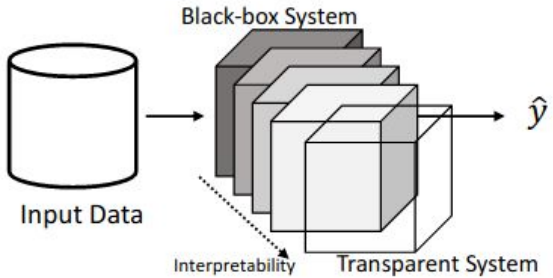


XAI Explanation

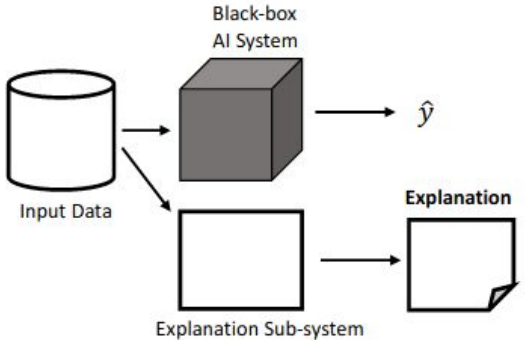


XAI Systems

Transparent-by-design systems



Post-hoc Explanation (black-box explanation) systems



[Mittelstadt et al. 2018]

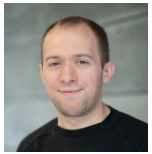
Our work in XAI

- Human-Centred AI Lab in Sarajevo, UNSA
- Explainable Robotics
- Explainable and trusted AI Industry - Semiconductor industry use case
- AI-assisted learning

Ajla Karajko

Mubina Kamberović

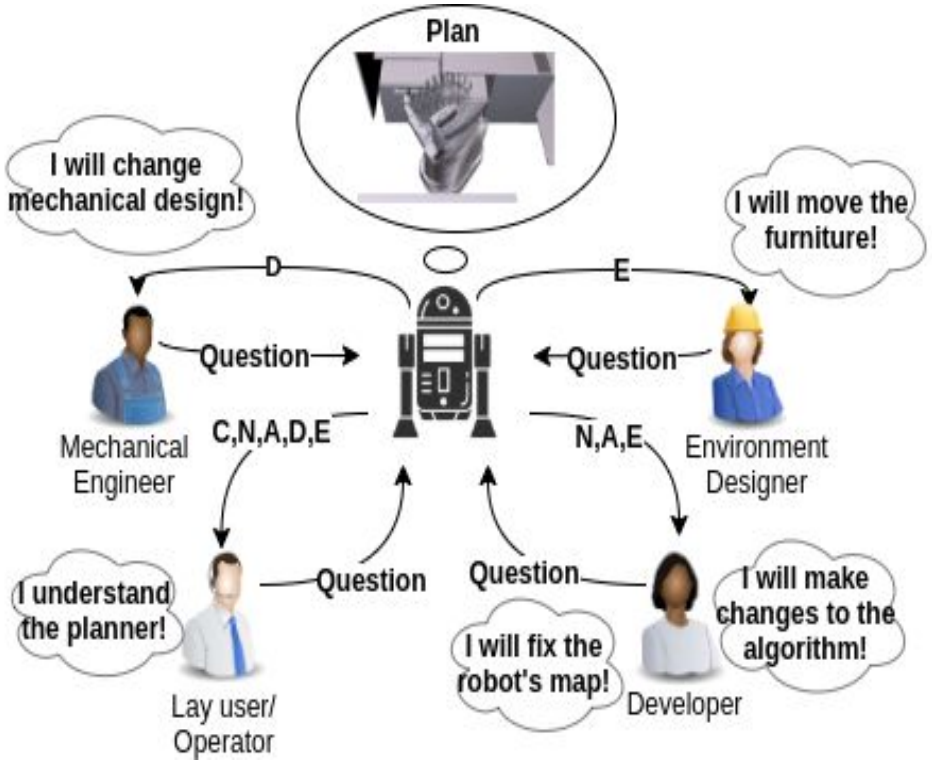
Vahidin Hasić



Amar Halilović

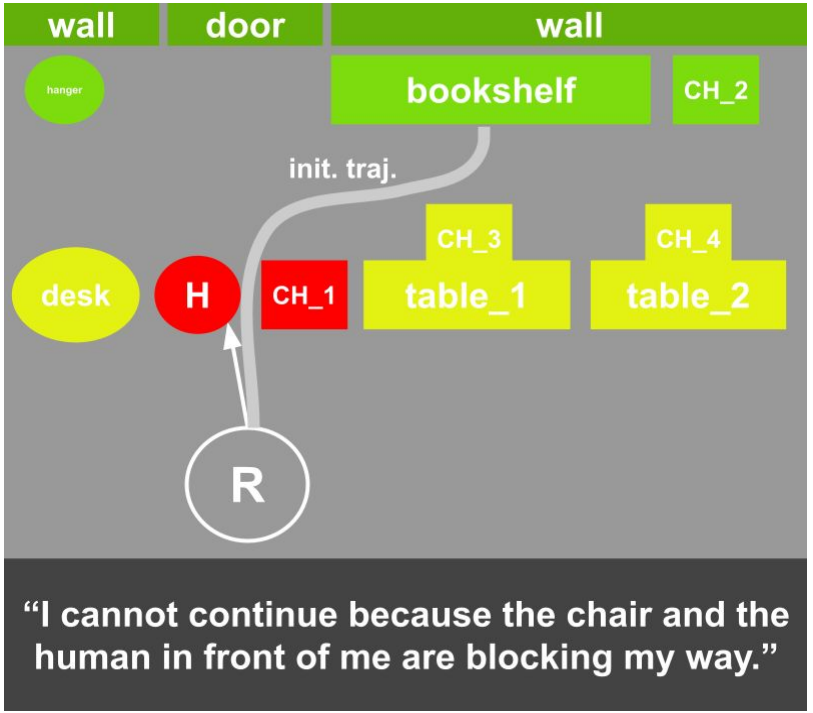
Amina Mević

Explainable AI and Robotics

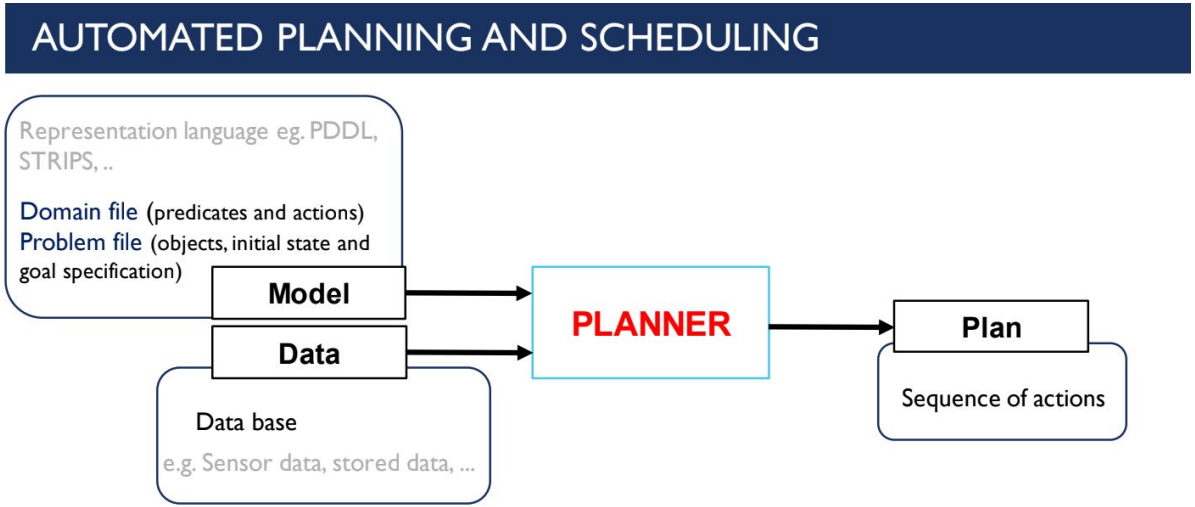


M Brandao, G Canal, S Krivić, D Magazzeni, ICRA 2021

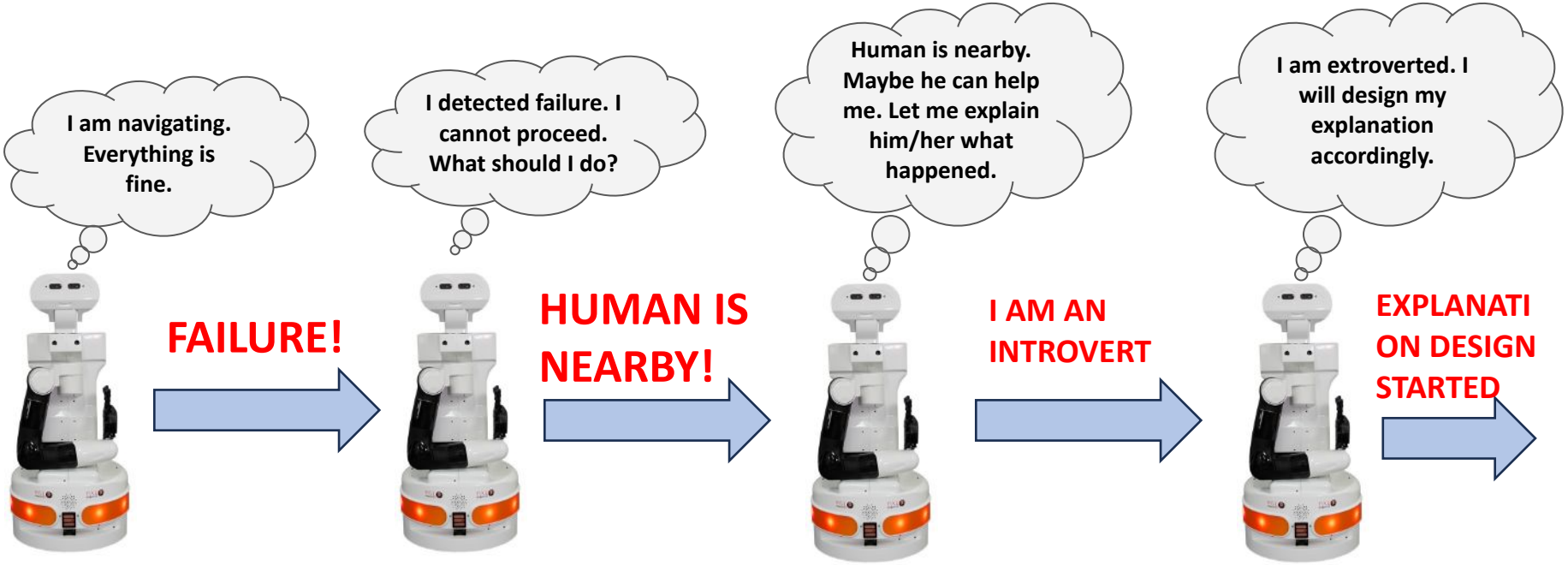
Explanation of Robot Navigation



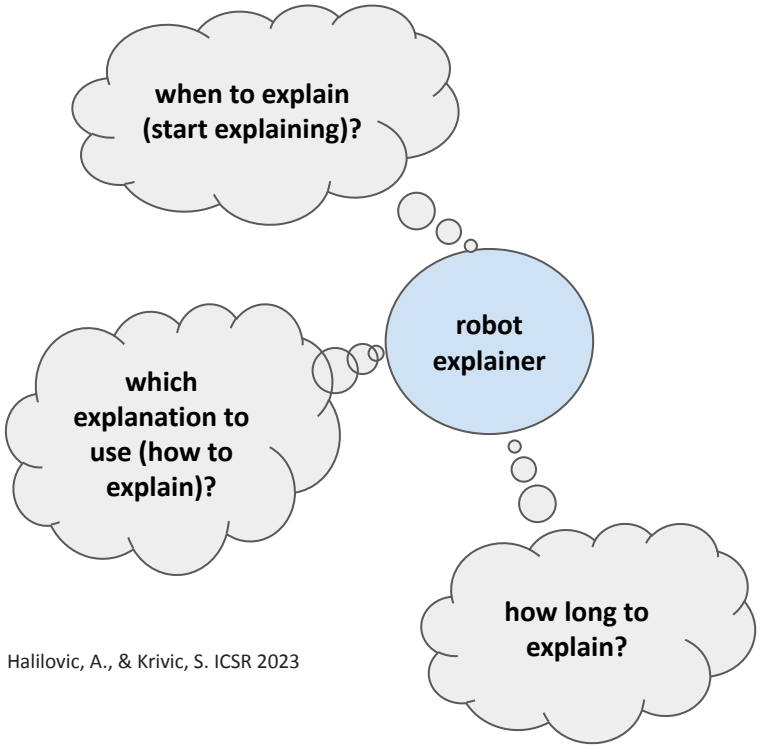
Planning of Explanations



- We propose a way of planning of explanations along with other robot actions.
- The explanation occurrence can vary based on parameters such as user preference or other task priorities.

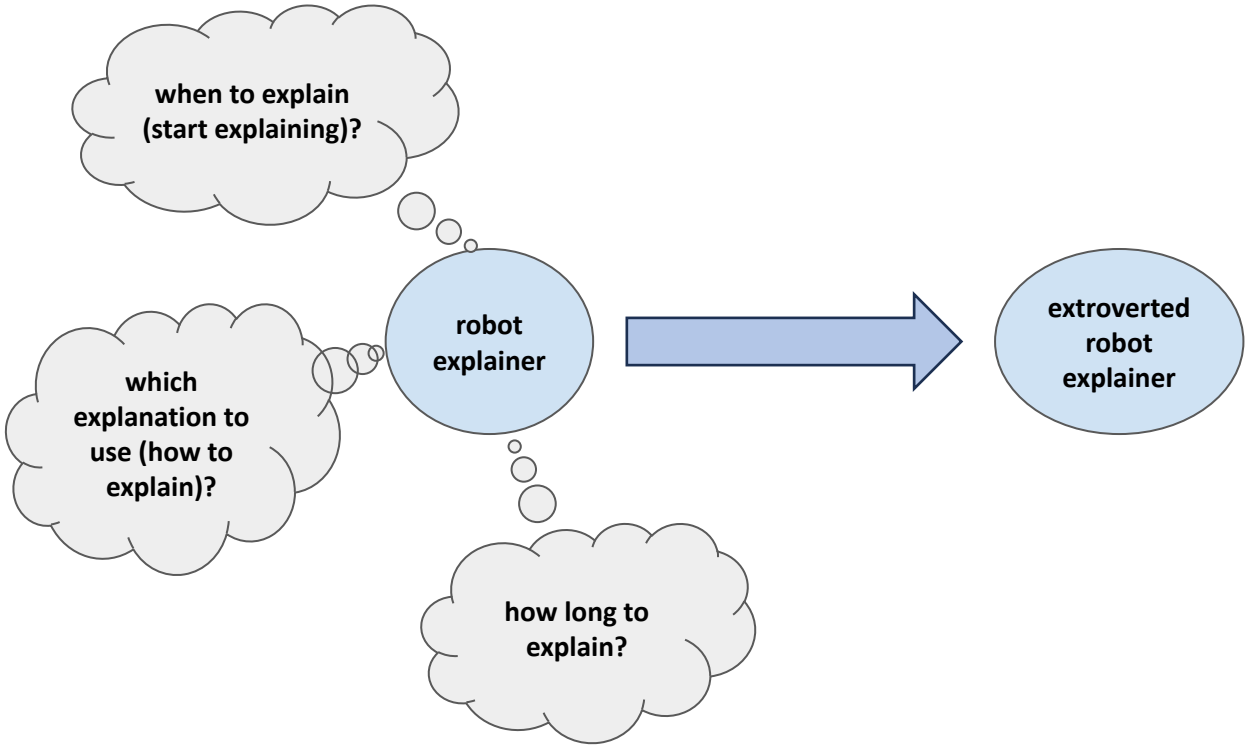

I am navigating. Everything is fine.
FAILURE!
I detected failure. I cannot proceed. What should I do?
HUMAN IS NEARBY!
Human is nearby. Maybe he can help me. Let me explain him/her what happened.
I AM AN INTROVERT
I am extroverted. I will design my explanation accordingly.
EXPLANATION DESIGN STARTED

Influence of Robot Personality on Explanations

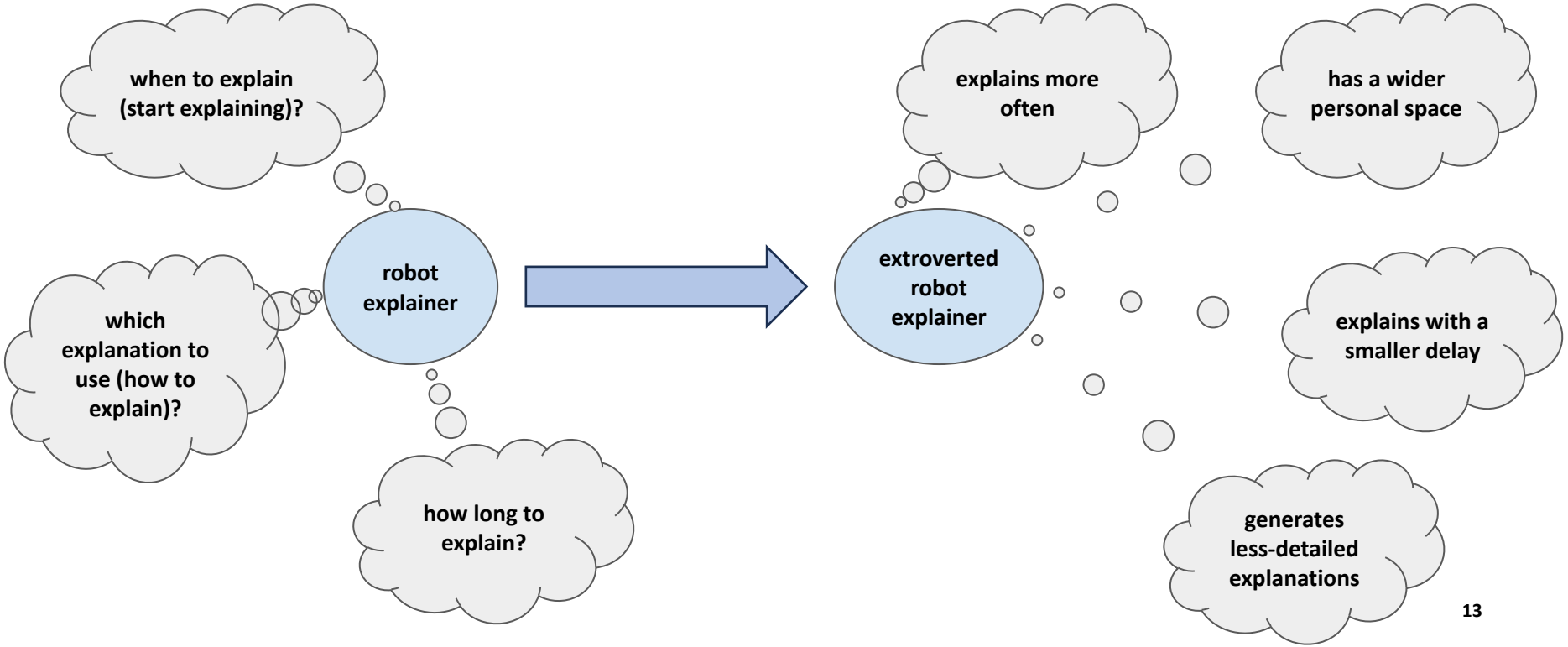


Halilovic, A., & Krivic, S. ICSR 2023

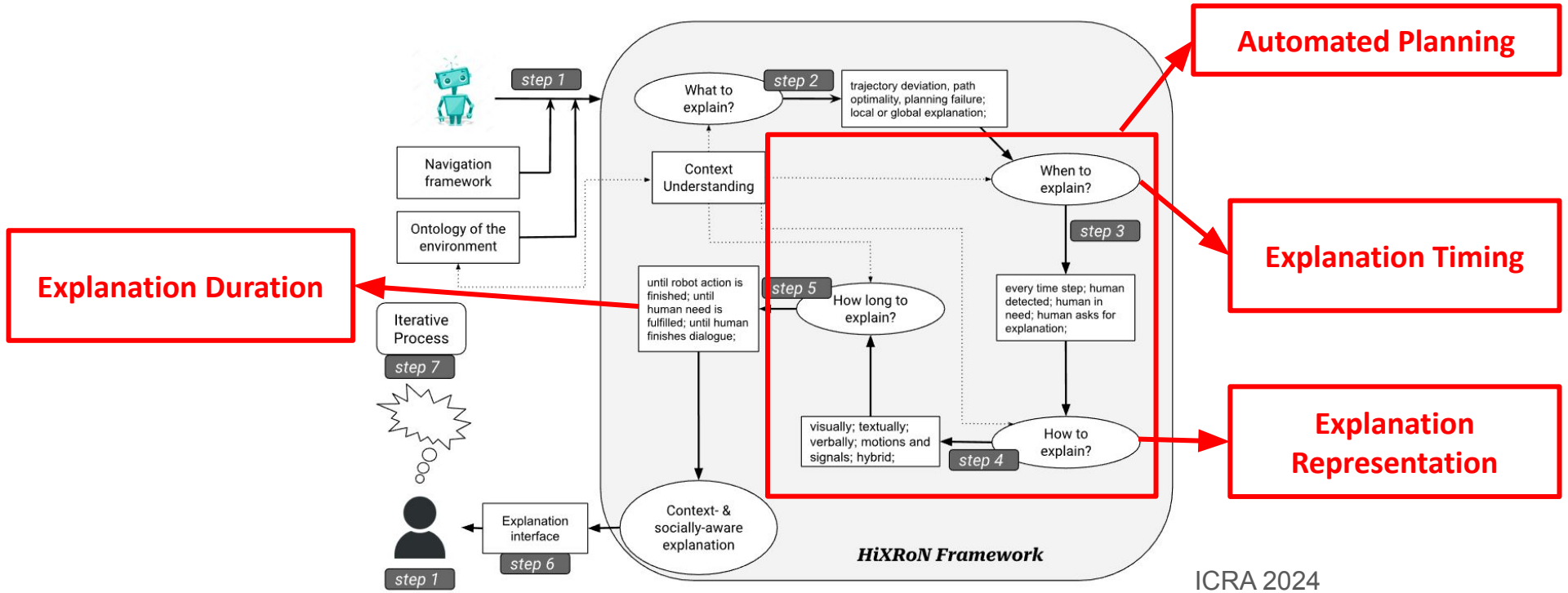
Influence of Robot Personality on Explanations



Influence of Robot Personality on Explanations



Explanation Generation



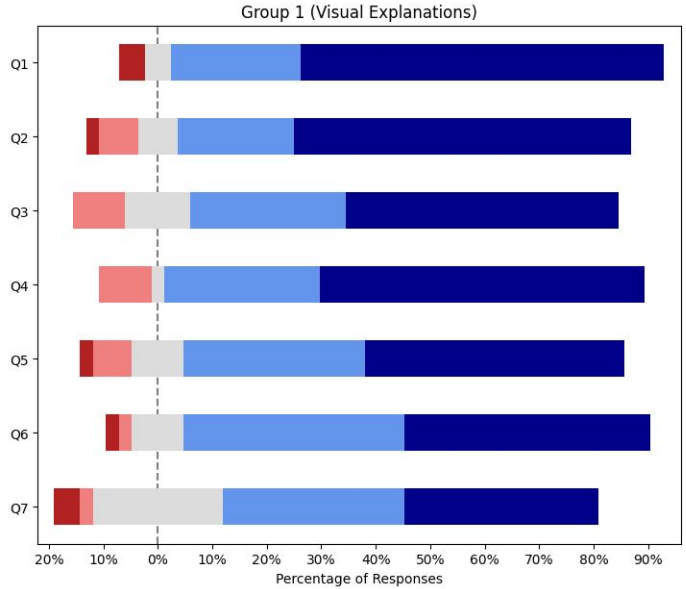
User Study on Explanation Representation

- Task: fetching a coffee from a coffee machine on a table
- Online study
- 84 participants (volunteers)
- Two groups:
  - Control group: visual explanations
  - Experimental group: visual-textual explanations
- User satisfaction scale used (Hoffman et al. 2018)*

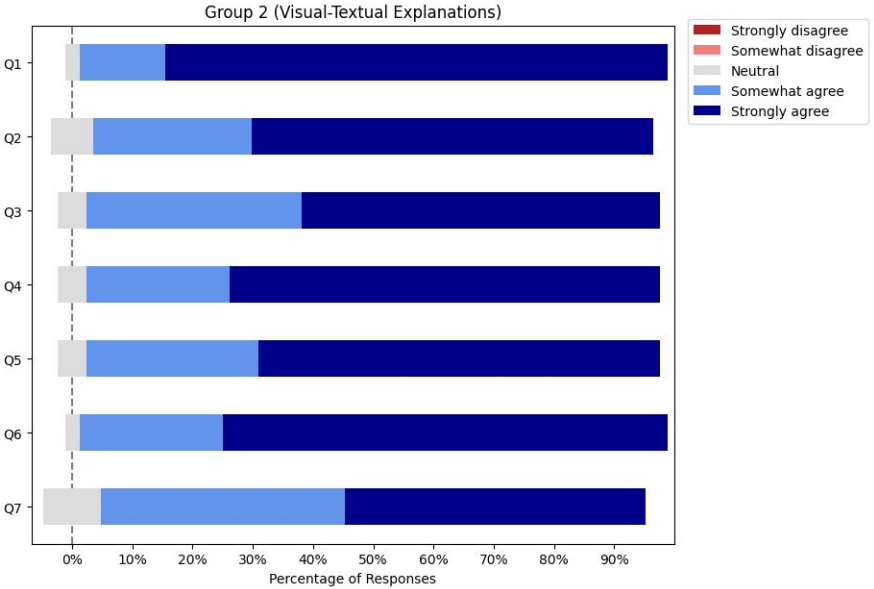
ICRA 2024

*Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608.

User study on explanation representation



The mean of all responses for Group 1 is 4.24 (SD = 0.99), corresponding to the overall attitude of "somewhat agree".

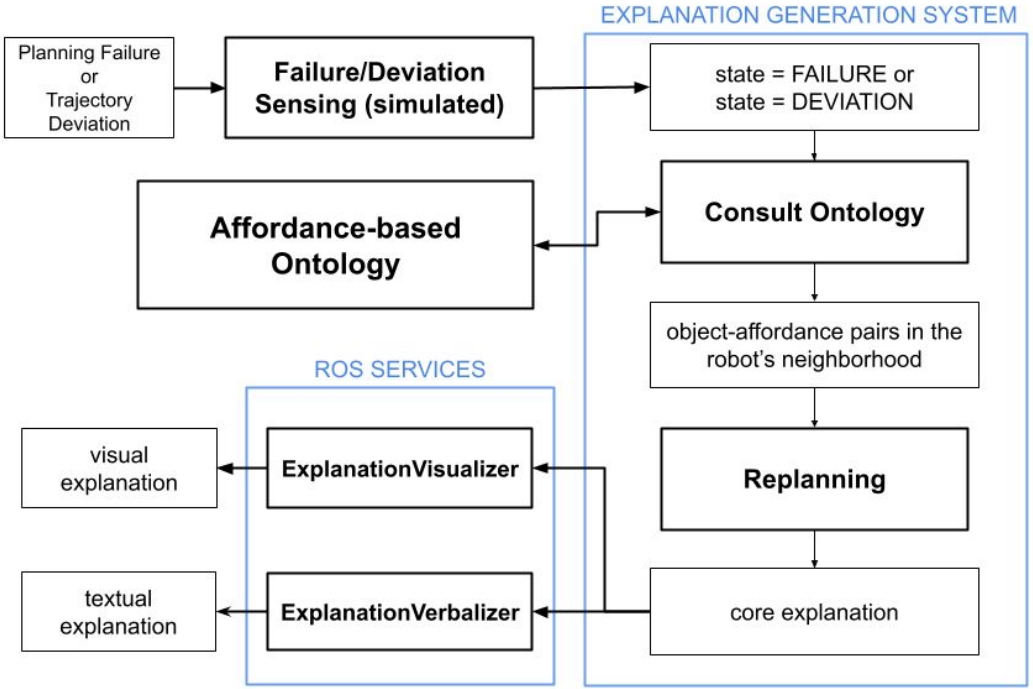


The mean of all responses for Group 2 is 4.62 (SD = 0.58), corresponding to the overall attitude of "strongly agree".

Motivation for Affordance-based Explanations

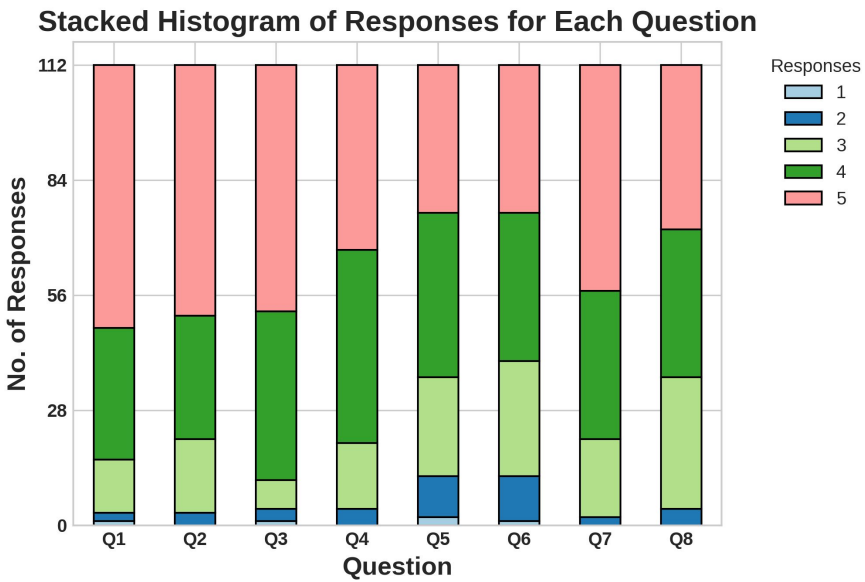
- Lack of environmental explanations
- Affordances for explanations
- Service robot navigation in indoor social settings
- Visual and textual explanation embodiments

Generation of Affordance-Based Explanations



User study on explanation satisfaction

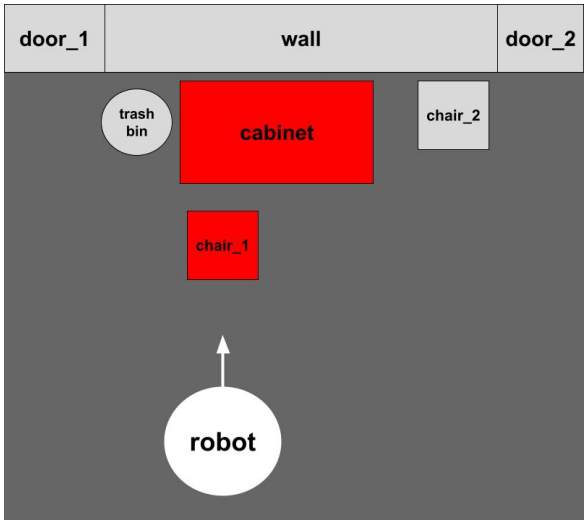
- Five different tasks (scenarios)
- Online study
- 112 participants (volunteers)
- One participant group
- User explanation satisfaction scale used (Hoffman et al. 2018)*



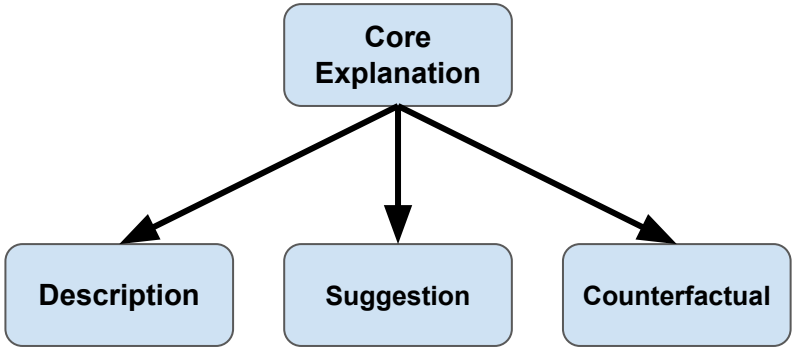
*Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608.

Explanation Visualization and Verbalization

Visualization

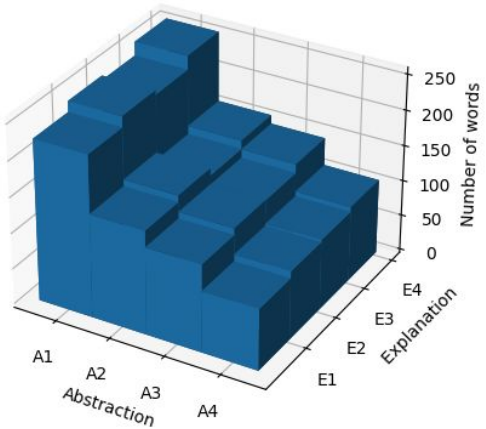


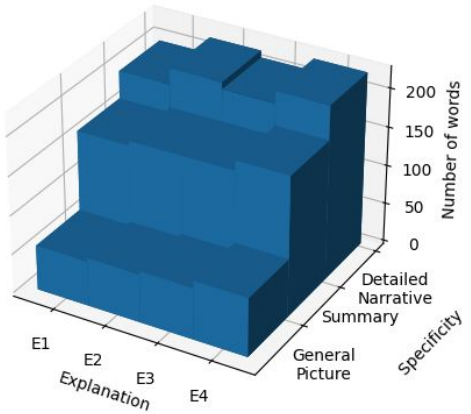
Verbalization



Verbalization experiments

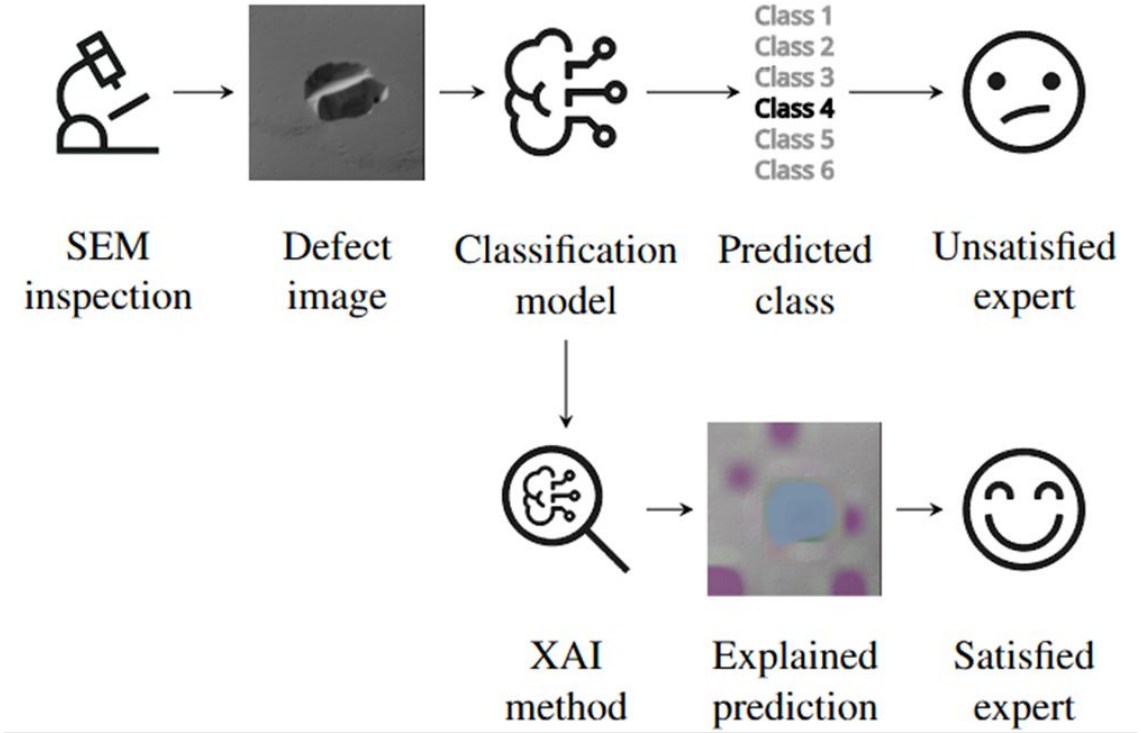
- Expanded the existing 3D verbalization framework* with affordance-based explanations
- Explanations do not increase the number of words in verbalizations





*Rosenthal, Stephanie, Sai P. Selvaraj, and Manuela M. Veloso. "Verbalization: Narration of Autonomous Robot Experience." IJCAI. Vol. 16. 2016.
*Canal, Gerard, et al. "PlanVerb: Domain-Independent Verbalization and Summary of Task Plans." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. No. 9. 2022.

Explaining Scanning Electron Microscope (SEM) Defect Image Classification



Submitted to 34th International Conference on Artificial Neural Networks.

ASE - Any Segment Explanation

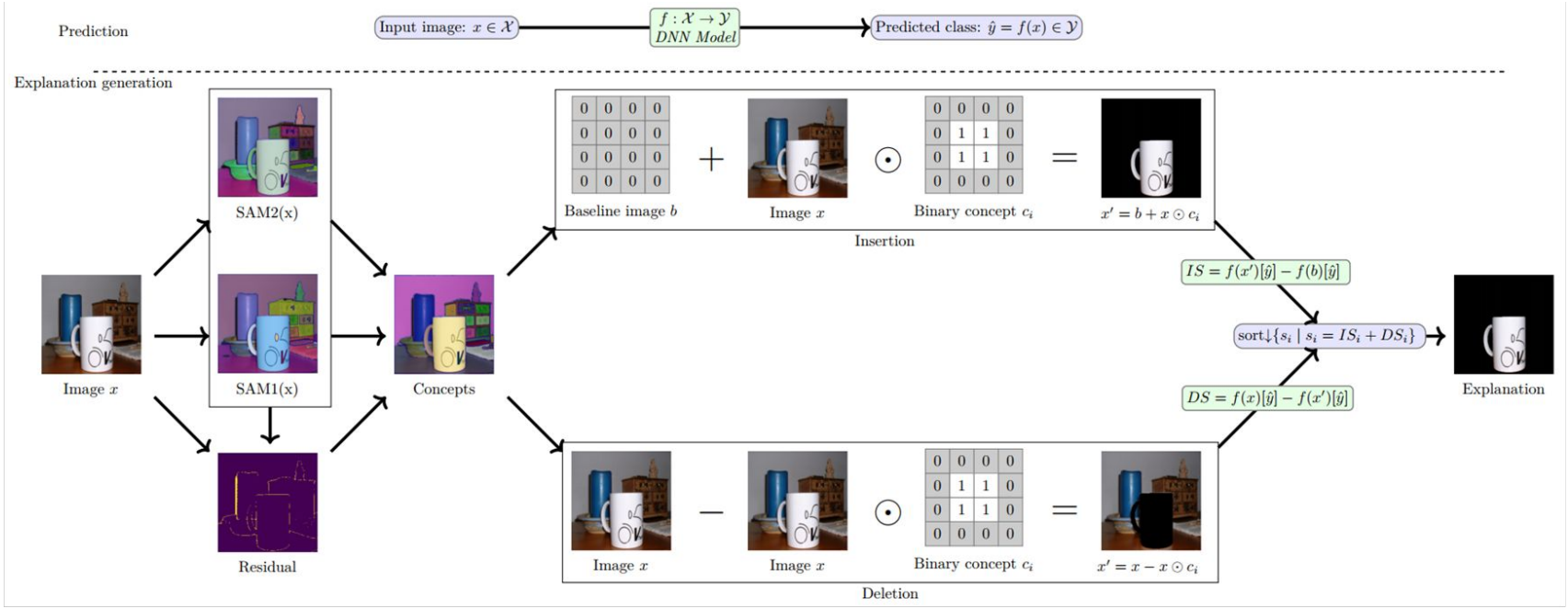
- Traditional XAI relies on individual pixel-based explanations, which highlight significant pixels but may not align with human intuition
- Concept-based explanations are more human-understandable but typically depend on human-annotated concepts
- **ASE - black-box method that automatically extracts concepts and assigns importances based on concept insertion and deletion**

Prediction output: goldenfish

Original image

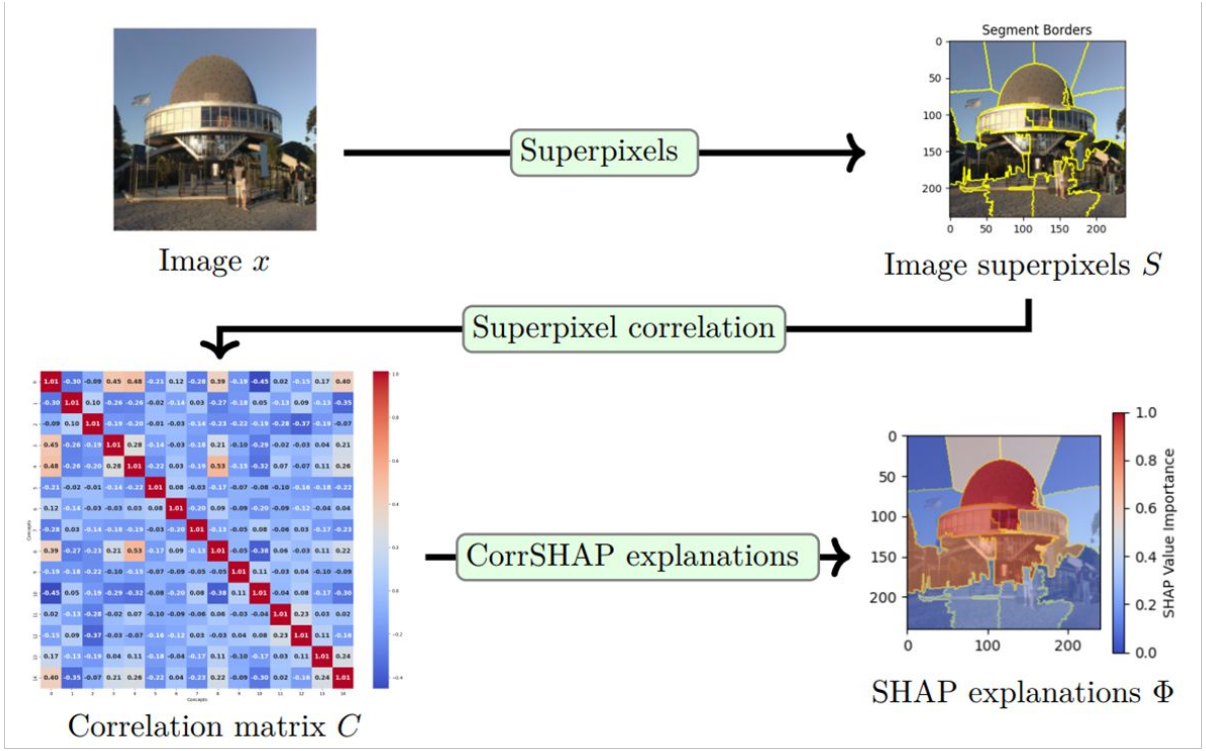
ASE explanation

ASE - Any Segment Explanation



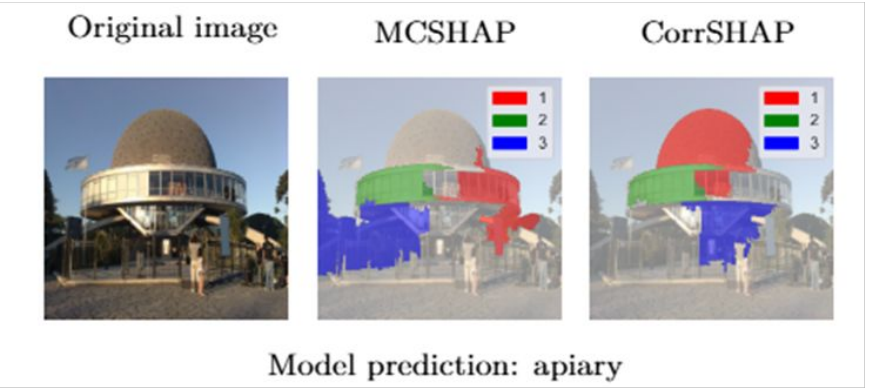
Submitted to the International Conference on Artificial Neural Networks. (ICANN 2025)

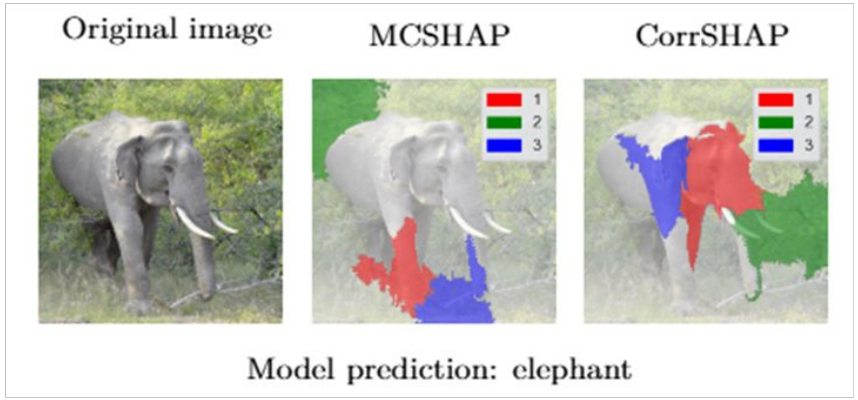
CorrSHAP



Accepted at World Conference on eXplainable Artificial Intelligence (XAI) 2025.

CorrSHAP - Results

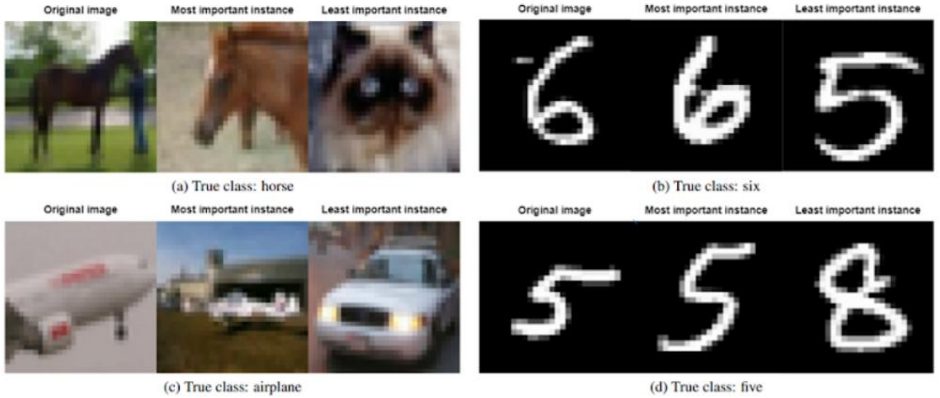




Model	Superpixels	CorrSHAP 1	CorrSHAP 2	CorrSHAP 3	MCSHAP
Area Under the Curve (AUC) Insertion ↑					
MobileNet-v2	Quickshift	80.4	80.37	80.29	**80.89**
	SLIC	**78.12**	78.13	78.15	77.79
ResNet-18	Quickshift	60.21	60.23	60.21	**61.60**
	SLIC	54.66	54.65	54.65	**55.41**
ResNet-50	Quickshift	**82.63**	82.61	82.63	82.27
	SLIC	**81.20**	81.20	81.20	80.66
ViT-b16	Quickshift	80.83	80.83	80.74	**81.84**
	SLIC	76.36	76.33	76.41	**76.82**
Area Under the Curve (AUC) Deletion ↓					
MobileNet-v2	Quickshift	**20.14**	20.14	20.16	20.93
	SLIC	**19.48**	19.48	19.48	21.40
ResNet-18	Quickshift	8.25	8.25	8.25	**8.03**
	SLIC	9.06	9.06	9.06	**9.01**
ResNet-50	Quickshift	**22.79**	22.79	22.79	24.16
	SLIC	**22.36**	22.36	22.39	23.79
ViT-b16	Quickshift	17.20	17.17	17.13	**16.86**
	SLIC	20.53	20.53	20.53	**20.48**
Execution Time (seconds) ↓					
MobileNet-v2	Quickshift	**0.42**	0.54	1.15	16.13
	SLIC	**0.74**	0.89	1.86	14.98
ResNet-18	Quickshift	**0.43**	0.50	1.09	7.85
	SLIC	0.66	**0.51**	1.91	13.82
ResNet-50	Quickshift	**0.46**	0.58	2.76	25.16
	SLIC	**0.78**	0.93	5.93	36.49
ViT-b16	Quickshift	**0.40**	0.52	5.41	7.26
	SLIC	**0.73**	1.01	10.53	18.00

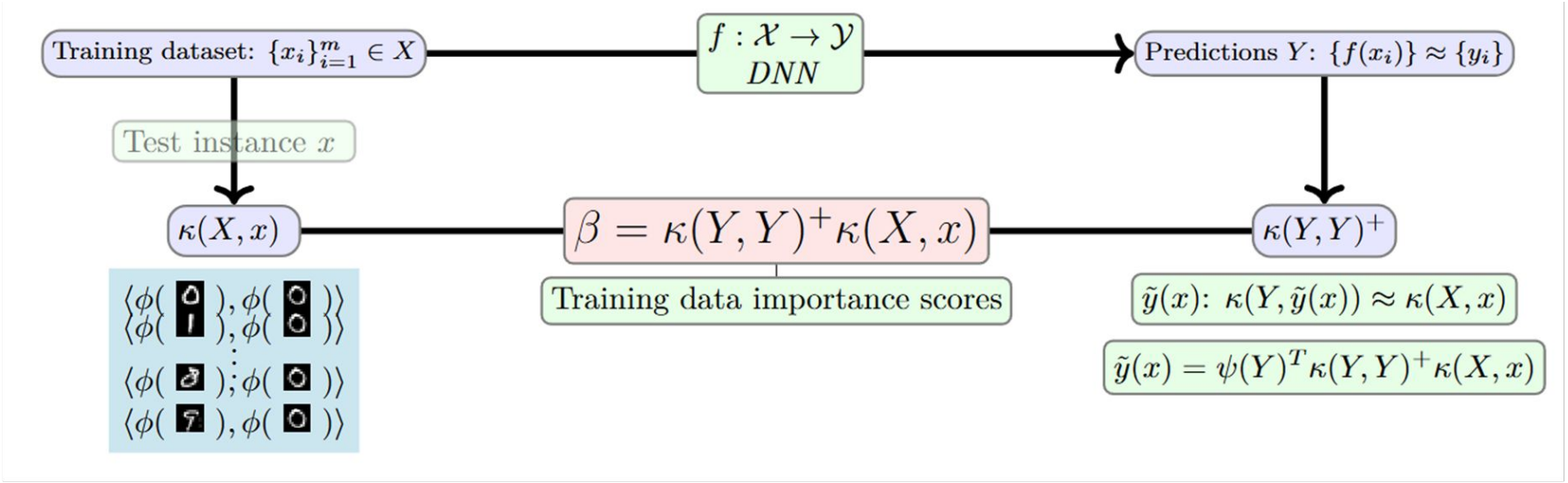
Kernel Sample-Based Explanations

- Knowing the influence of individual training instances on the predictions has a lot of applications:
  - detecting mislabeled data
  - identifying data leakage
  - analyzing memorization effects
  - optimizing training dataset
  - control theory
  - active learning
  - system identification



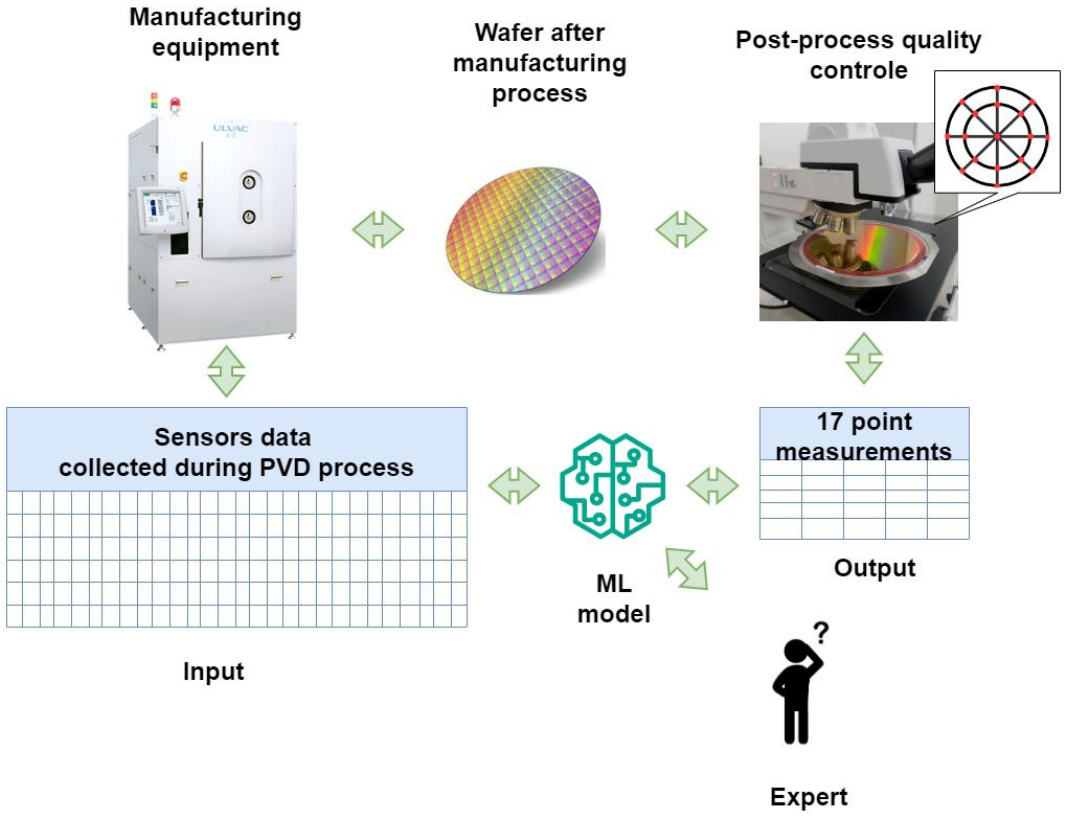
- **Kernel Sample Based Explanations (K-SBE) - novel model-agnostic approach built upon kernel functions**

Kernel Sample-Based Explanations

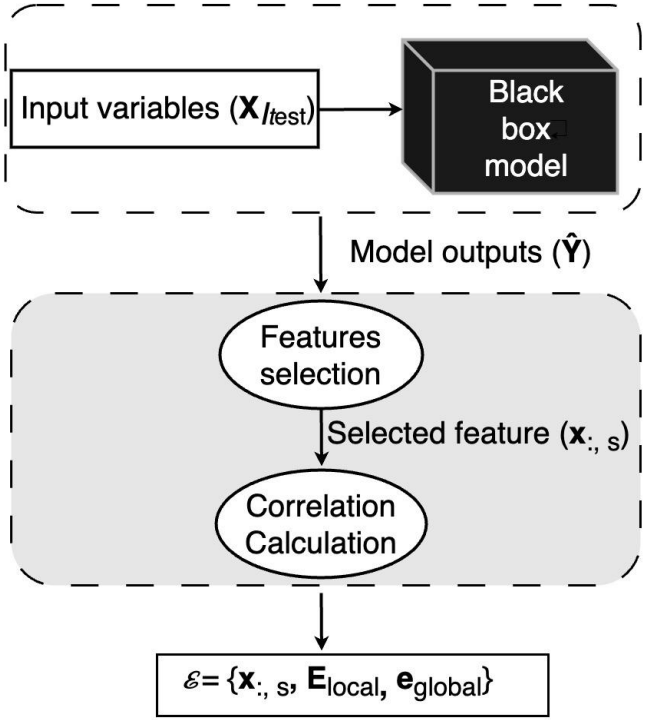


Submitted to International Conference on Computer Vision (ICCV 2025)

Virtual Metrology (VM) in Semiconductor Industry



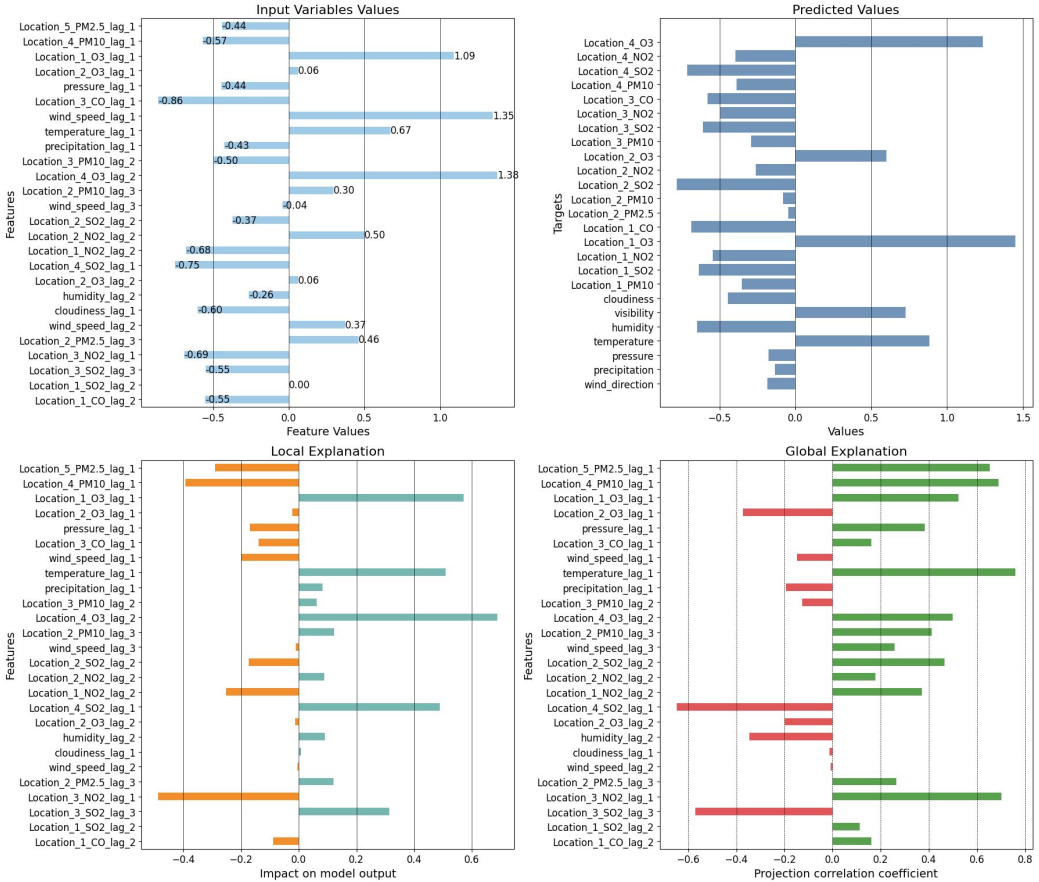
Explanations based on Projective Operator for Multi-Output Models



$\mathbf{E}_{local}$: Local explanation

$\mathbf{e}_{global}$: Global explanation

Example of MIMO Explanations



eXProj explanation for one input sample:

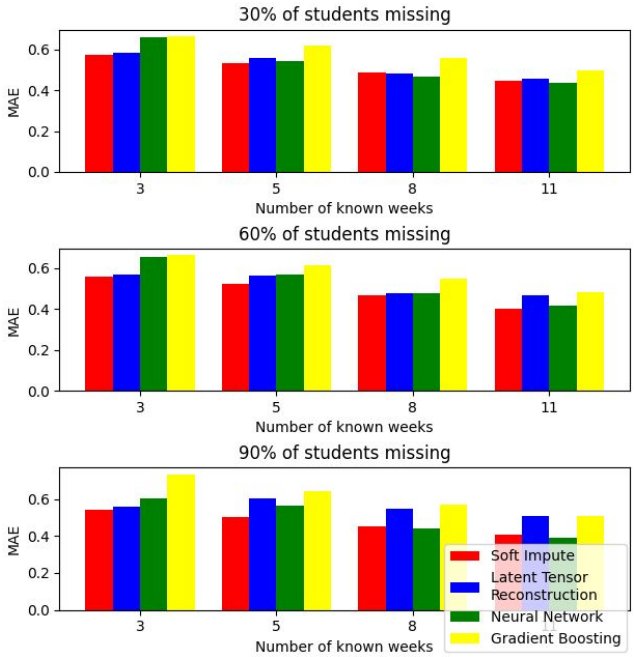
- Local explanation indicates the impact of the feature in on the prediction.
- Global explanation of the model indicate the influence of features on the predicted outputs.

AI-Assisted Learning

Objective

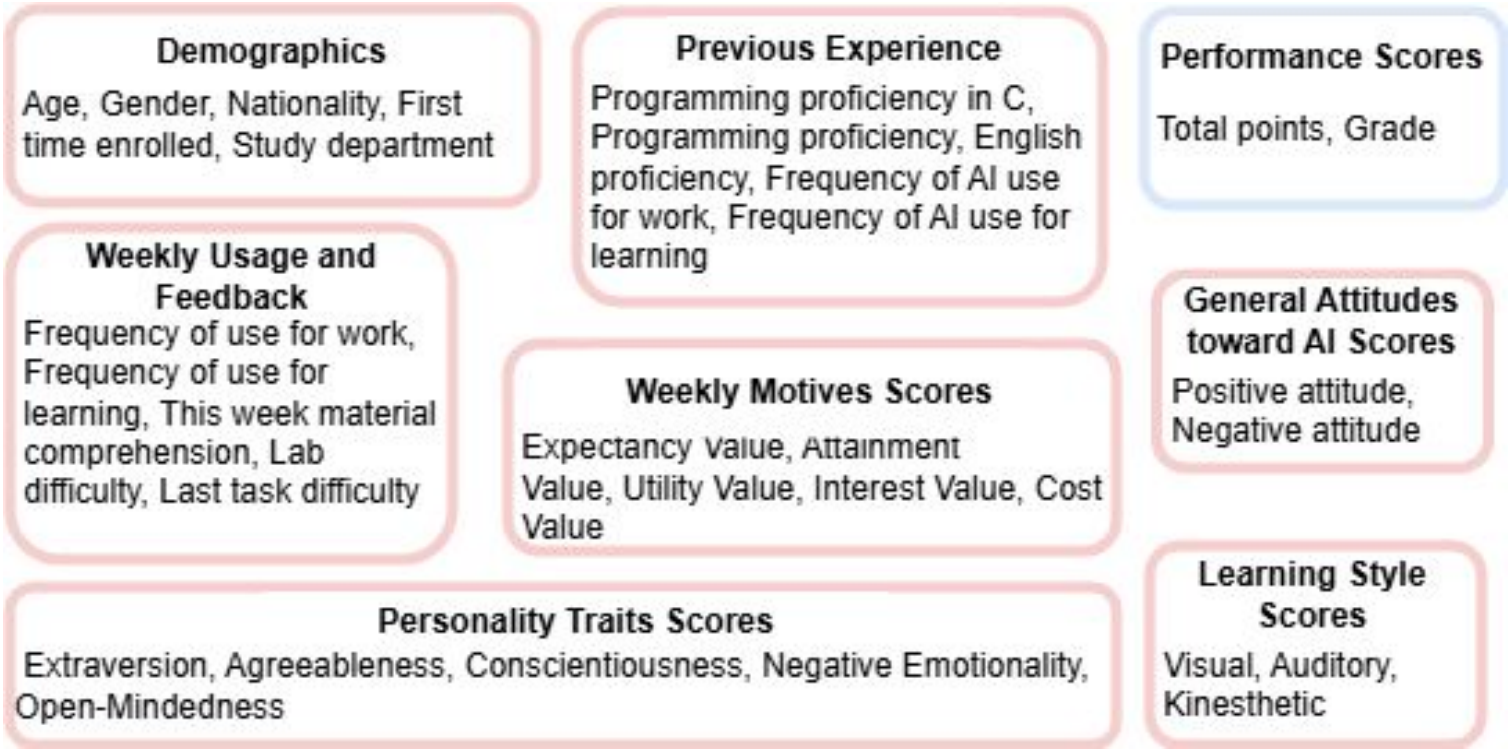
- Create AI systems that adapt to student's individual needs and knowledge state in order to maximize learning gains

Predicting mistakes that students make during learning to program using collaborative filtering techniques



UMAP 2023, UMAP 2025

Study on Investigating AI in Programming Education



Thank you for you attention!

Email: senka.krivic@etf.unsa.ba