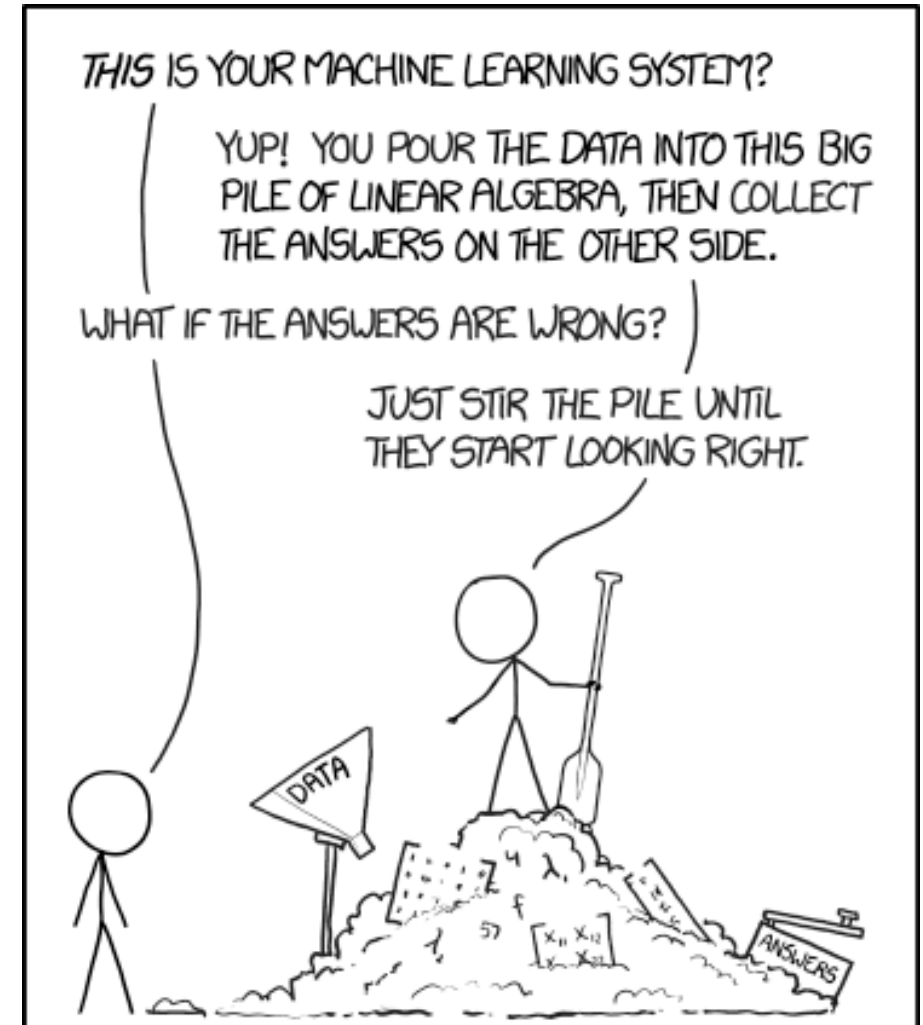


Evaluating Explainable AI

Are We Asking the Wrong Questions?

PROF. NAVA TINTAREV
DEPARTMENT OF ADVANCED COMPUTING SCIENCES
UNIVERSITY OF MAASTRICHT, THE NETHERLANDS



Comic from xkcd: <https://xkcd.com/1838/>

Disclaimers:

- No ChatGPT was used (or hurt) during the course of developing this presentation.
- All errors herein are my own.
- Focus on artificial advice givers.



Worlds apart?

SEPT 2000 – MARCH 2005
COMPUTER SCIENCE: UPPSALA UNIVERSITY,
SWEDEN

FEB 2003 – NOV 2003
PSYCHOLOGY: UNI. OF WOLLONGONG,
AUSTRALIA



Prof. Nava Tintarev, Explainable AI



Human-computer interaction
researcher

Adaptation of presentation

Evaluation

UMAP, CHIIR, CHI

Editorial board: User Modeling and
User-Adapted Interaction (UMUAI)



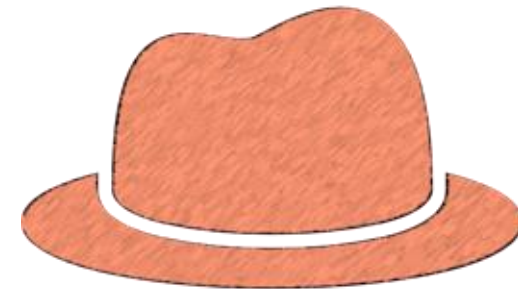
Intelligent systems researcher

Adaptation of content

Information retrieval and
recommendation algorithms

Recsys, ECIR, WSDM, ECAI

Editorial board: Transactions on
Recommender Systems (TORS)



(research) Policy maker

Director of Research

Chair of informatics tabel

Board member IPN

“AI Ambassador” Maastricht

Hat icons created by Good Ware - Flaticon

Why explain

Post Office scandal: The ordinary lives devastated by a faulty IT system

17 January 2024

Share

Save

Add as preferred on Google

Michael Race & Lora Jones

BBC News

17th January BBC, 2024

Welfare surveillance system violates human rights, Dutch court rules

Government told to halt use of AI to detect fraud in decision hailed by privacy campaigners

5th Feb 2020, The Guardian



Svenska Dagbladet

Försäkringskassans fuskjagande AI

AI-professor kritiserar Försäkringskassan



Professor Virginia Dignum anser att öppenhet är grundläggande för myndigheter som vill fånga fuskare med hjälp av AI-system. Foto: Erik Abel

27th Oct 2024, Svenska Dagbladet,

TINTAREV, D3A 4.0, OCTOBER 2026

Overview

What is an explanation?

Is current evaluation sufficient?

Do our assumptions hold?

- Primacy of explanation fidelity
- Model Complexity is necessary for accuracy and bad for human understanding
- Evaluation challenges are restricted to (messy humans)

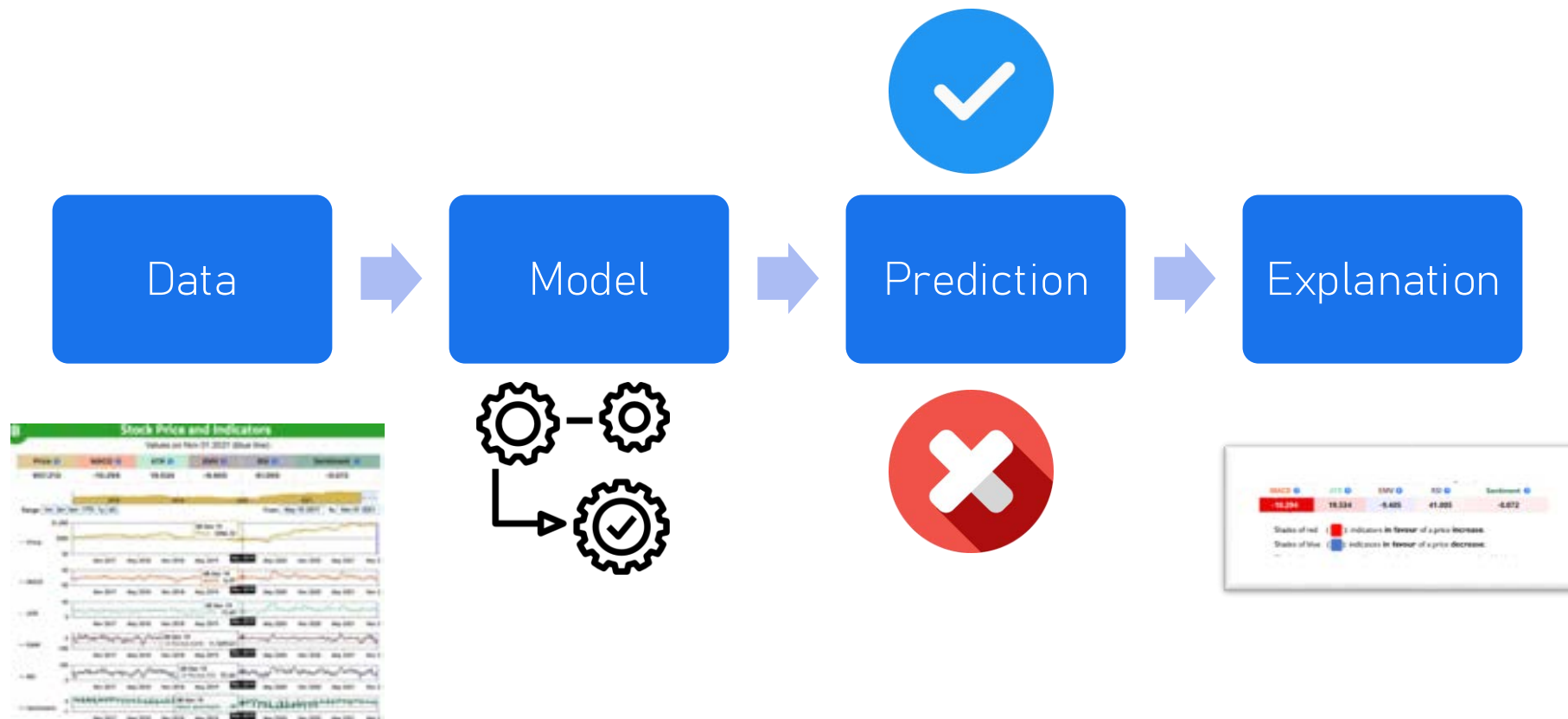
What does explanation effectiveness depend on?

Research agenda: more standardized evaluation protocols

What is an explanation?



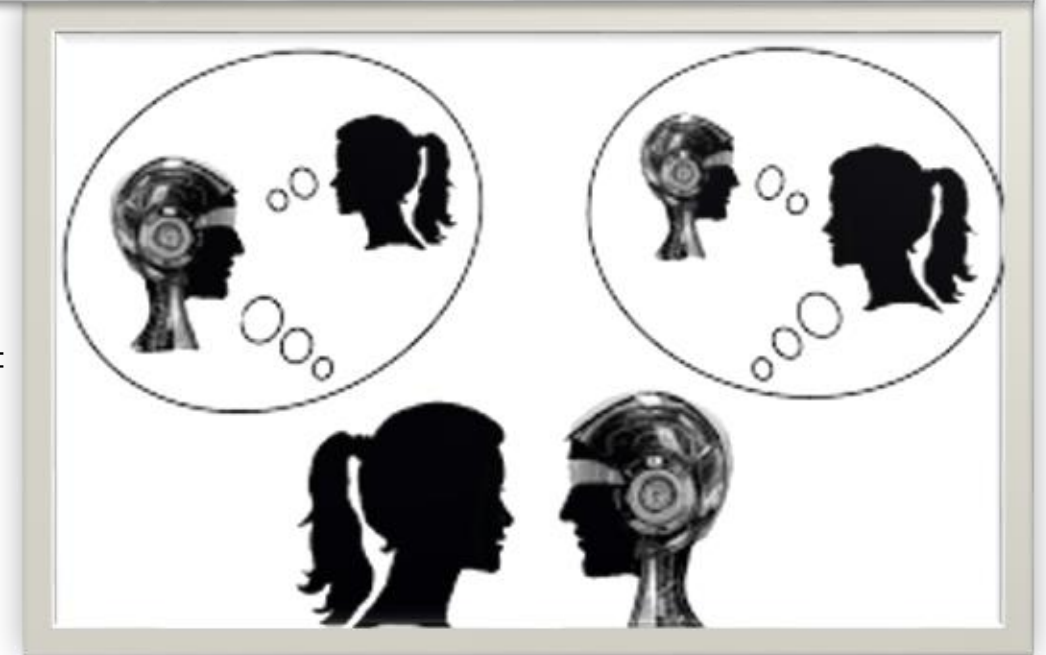
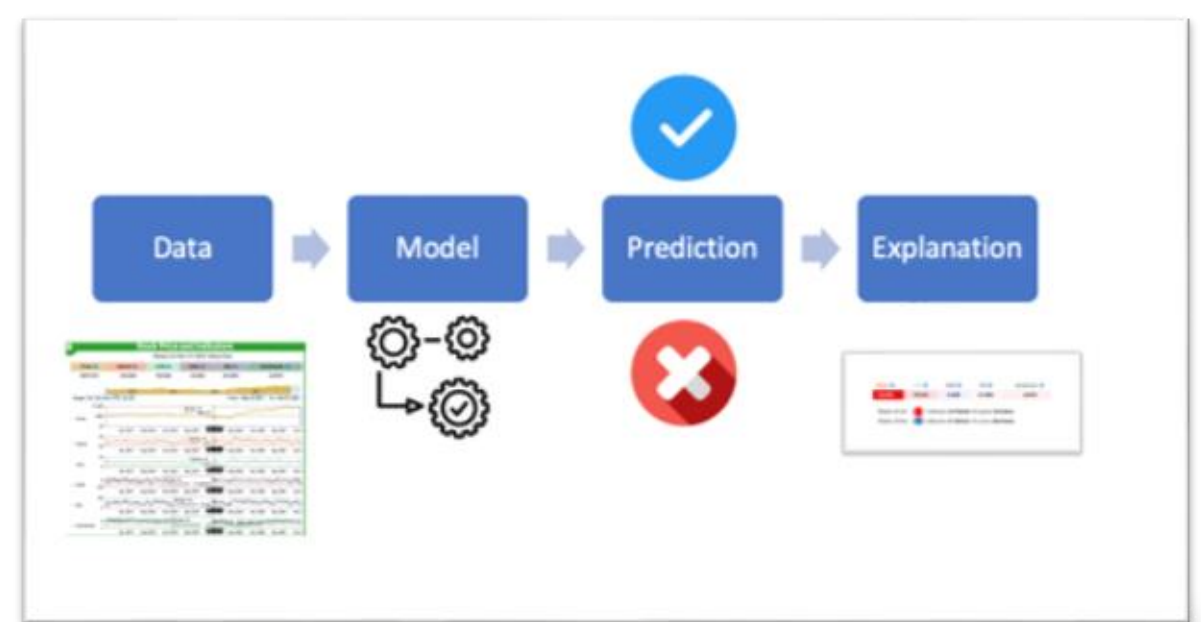
A common decision-making task...



Explain **verb**

- a. To make known
- b. To make plain or understandable
- c. To give the reason or cause for
- d. To show the logical development or relationships of

-- Merriam-Webster, 2026



Current approaches

By-design (also ante-hoc). e.g., linear/logistic regression, decision rules/trees

Model-agnostic. Only analyze how the model output changes with respect to changes in the feature inputs.

Local. E.g., ceteris paribus, LIME, Anchors, SHAP

Global/group-wise. E.g., partial dependence plot, functional decomposition, surrogate models

Model-specific. Analyze parts of the model to better understand it. E.g., saliency maps, concepts, sparse autoencoders, induction heads

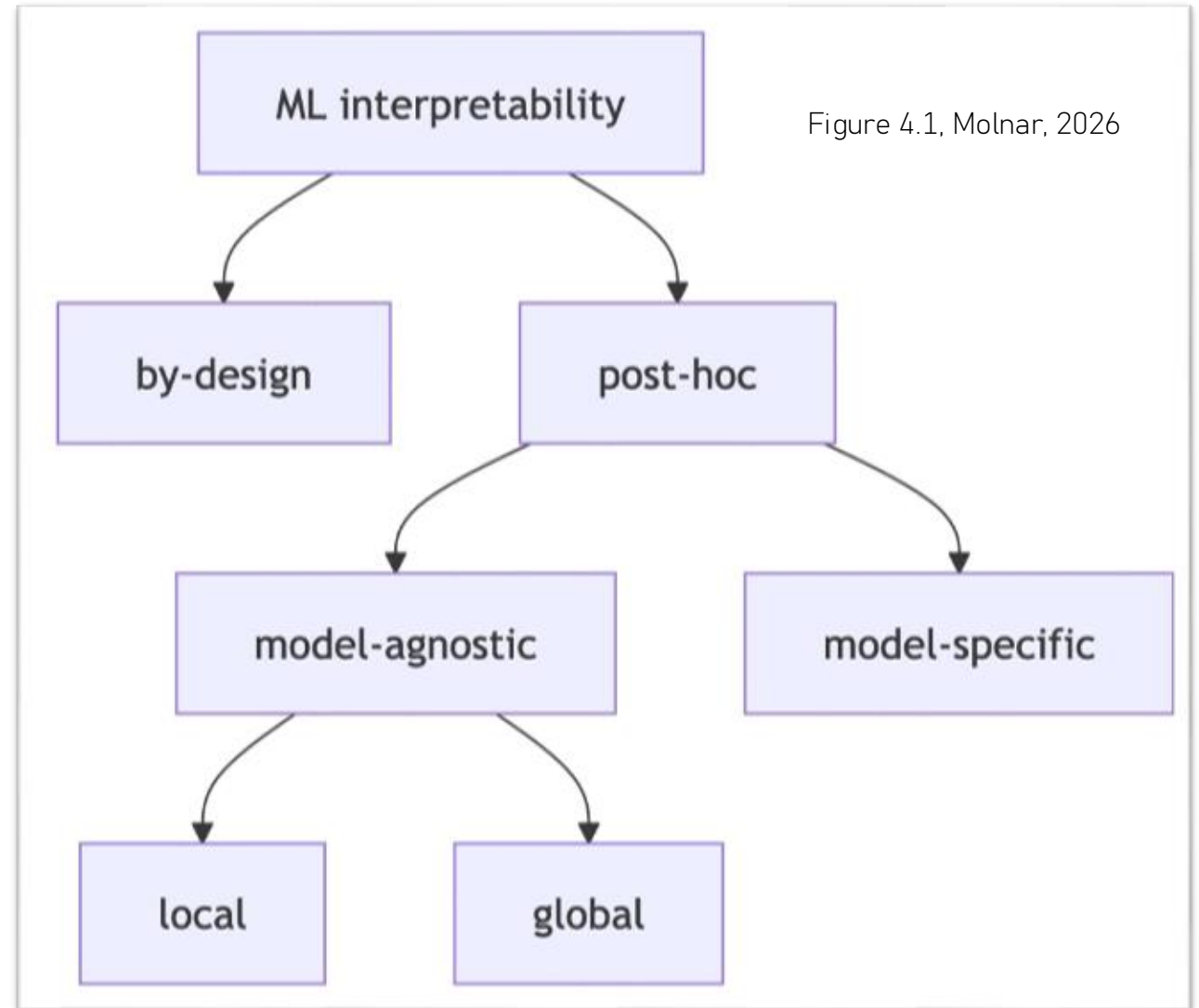
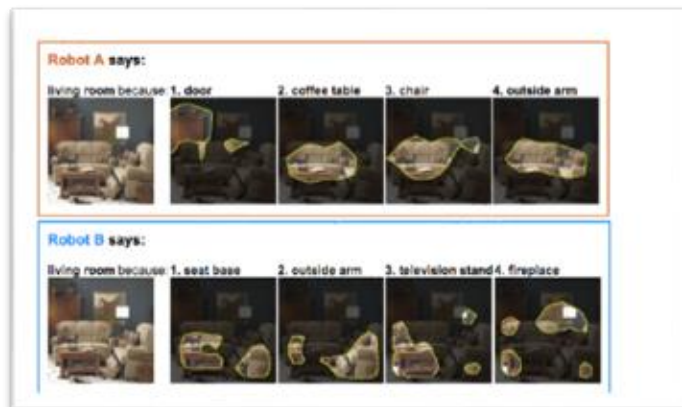
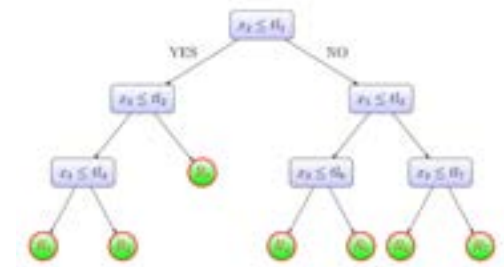


Figure 4.1, Molnar, 2026

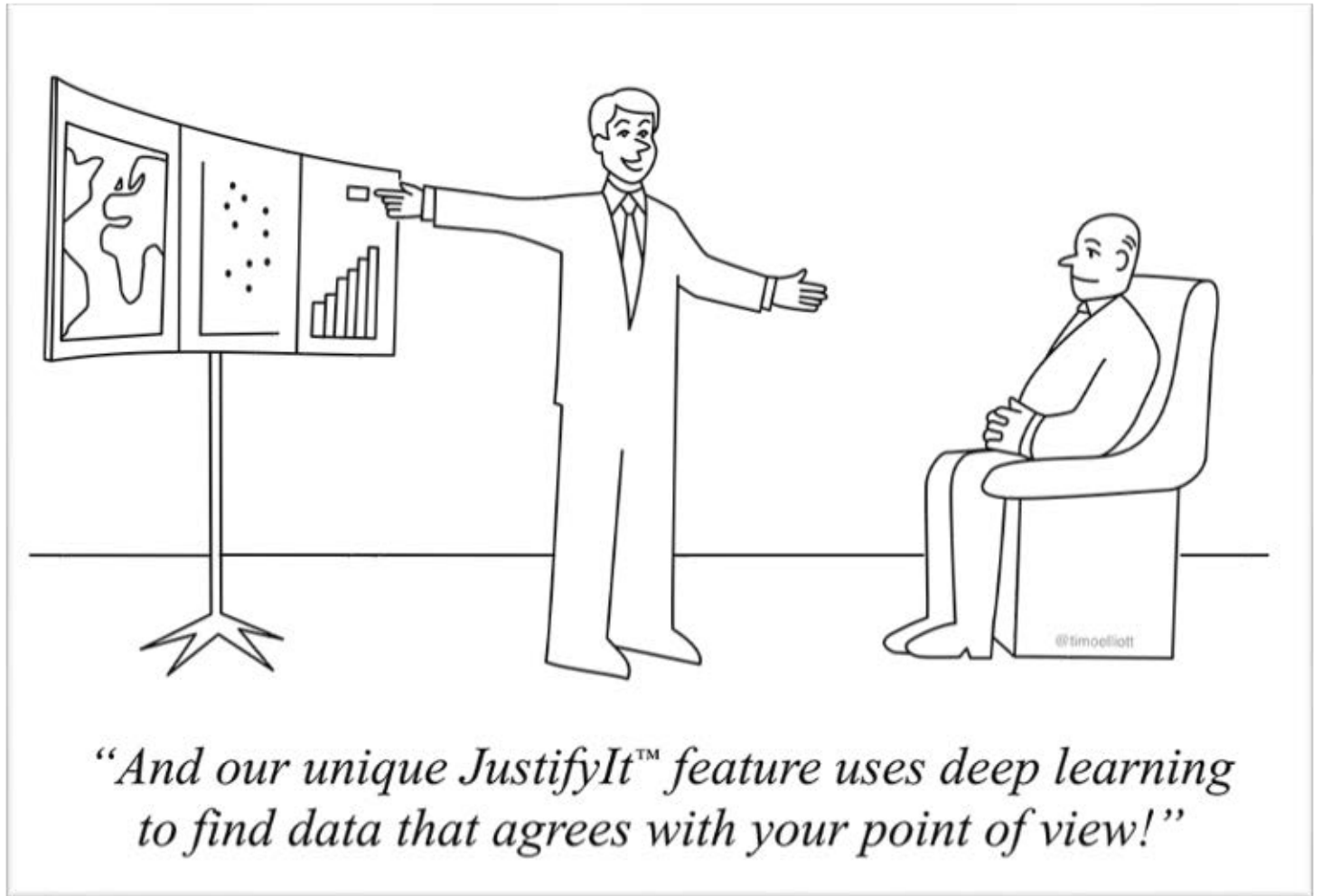
Explanation styles

- Model-based – model itself (e.g., decision tree) or approximation
- Instance-based
 - Attribution-based – explanatory power of input features. E.g., feature importance or influence.
 - Example-based – selecting or creating instances that are particularly well (or poorly) predicted. Includes counterfactuals.

Zhou et al. 2021.
Electronics. Evaluating
the quality of machine
learning explanations: A
survey on methods and
metrics



Is current evaluation sufficient?

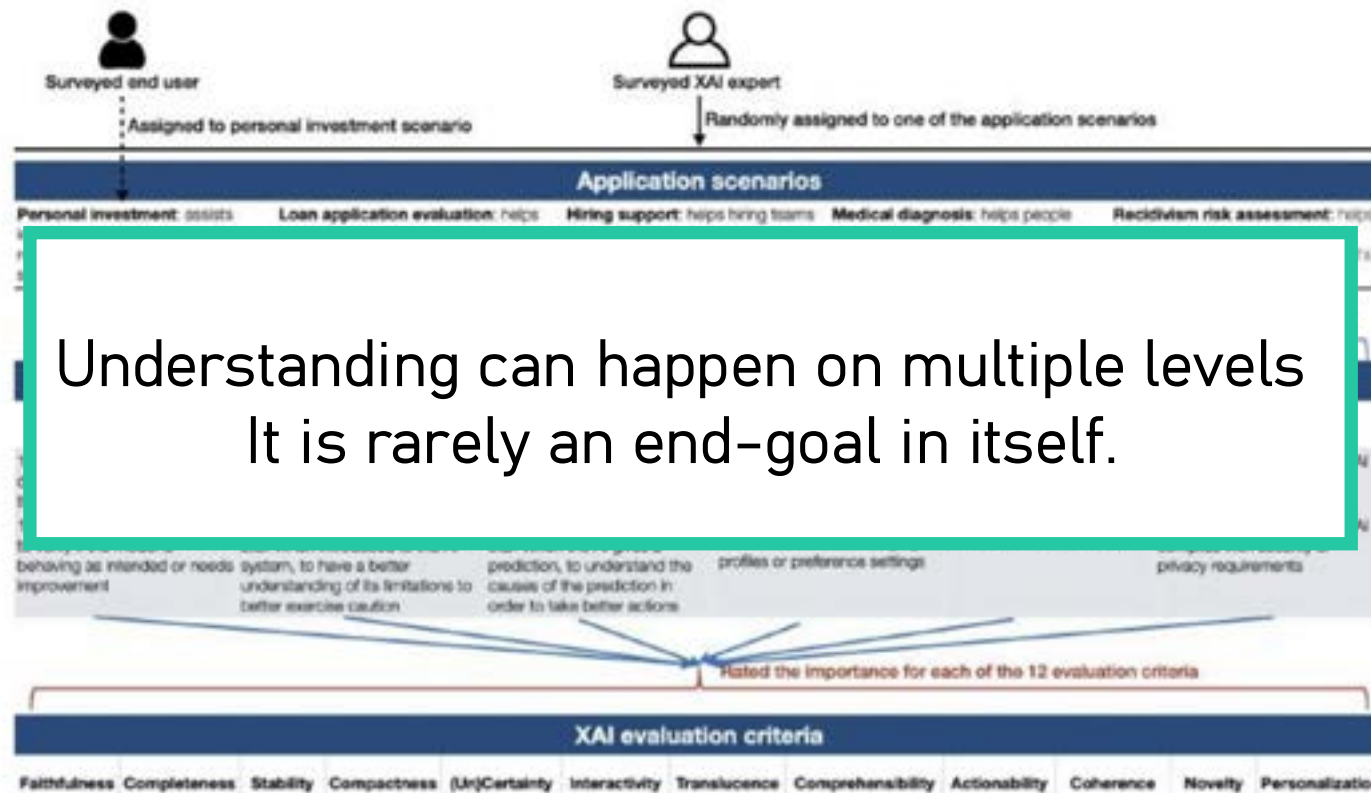


“Good” Explainable Artificial Intelligence (XAI)?

- Topic at ECAI and IJCAI, Dagstuhl, many literature surveys.
 - Mostly (supervised) machine learning
- E.g., Interpretability [Guidotti et al. 2019] or comprehensibility [Gleicher 2016], is the ability of stakeholders to **understand** relevant aspects of the modeling process.



Different explanatory goals and criteria [Liao et al, 2023]



Decision support Ultra-strong AI?

Most XAI papers
are still here!

Weak
criterion

- machine learner produces improved predictive performance with increasing data

Strong
criterion

- system provides its hypotheses in symbolic form

Ultra-strong
criterion

- the AI teaches a (learned) symbolic hypothesis to a human, **whose performance is consequently increased** to a level beyond human alone

Do our assumptions hold?



Credit: Assumption icons created by HAJICON – Flaticon

Predictive performance improves with model complexity

“Explanation fidelity is not a desirable property—it is THE property.”

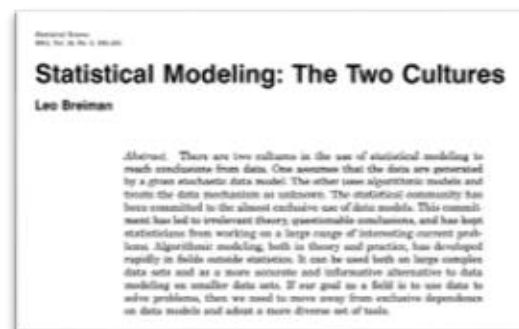
Evaluation challenges are restricted to (messy humans)



Icon: Leremy - Flaticon

“Predictive
performance
improves with model
complexity”

Rashomon effect



- Originally coined by Leo Breiman 2001 ("Two Cultures") – parametric statistical modeling vs algorithmic machine learning.
- Multiple models with similar performance but different interpretations
- Refers to the unreliability of eye-witness accounts
- Is named after Akira Kurosawa's 1950 Japanese movie Rashomon which tells four different versions of a murder story.
 - While each version can explain the events equally well, they are incompatible with each other.



Occam Dilemma

“Simple and interpretable functions do not make the most accurate predictors”

Other things being equal, simpler explanations are generally better than more complex ones



Commonly attributed to
William of Ockham
14th century priest and
philosopher

When is it a dilemma?

- **Rashomon Sets** is a set of almost-optimal models. Lots of work on sparse decision trees a.o.
- if there are many different-but-good models, some of them are likely to be interpretable (Cynthia Rudin et al)
- Semenova et al. (2022) formalise the Rashomon ratio, defined as the volume of accurate models relative to the overall hypothesis space, showing that when this ratio is large, simpler or more interpretable models are more likely to exist.
- Xin et al. (2022) explicitly enumerate the Rashomon set for sparse decision trees, enabling analysis of structural variability among equally accurate models



Reducing the set – user modeling

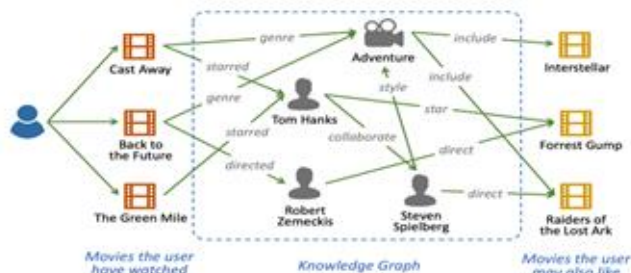


Adarsa Sivaprasad



Prof Emeritus
Ehud Reiter

- A lot of explanatory models: feature reduction --> reduced complexity.
- Reduction can also align models to user's explanation needs [Sivaprasad2024]
 - **Clinician:** Emphasis on Modifiable Physiological Parameters
 - **Patient:** Emphasis on lifestyle factors
- Knowledge-graph path reasoning [Xian et al 2019; Balluccio2022]





“(Explanatory)
Models can be both
accurate and less
complex”

“Explanation fidelity
is not a desirable
property—it is the
property.”



Icon: Leremy - Flaticon

Fidelity, soundness, completeness

Explanation Types	Quantitative Metrics	Properties of Explainability				
		Interpretability			Fidelity	
		Clarity	Breadth	Parsimony/Simplicity	Completeness	Soundness
Attribution-based explanations	Monotonicity [36]					✓
	Non-sensitivity [36,37]					✓
	Sensitivity [34,35]					✓
	Effective complexity [36]		✓	✓		
	Remove and retain [36]					✓
	Recall of important features [37]					✓
	Implementation invariance [35]					✓
	Selectivity [37]					✓
	Continuity [37]	✓				
	Sensitivity-a [36]					✓
	Mutual information [36]		✓	✓		✓

- **Mathematical fidelity** refers to the faithfulness of technology-based behavior and properties of virtual objects to the **mathematical** behavior and properties of the objects they are intended to represent.
- **Soundness** means correct and truthful. You *cannot* state/prove anything that's wrong.
 - E.g., the explanation does not indicate features that are not important.
- Also called Correctness [Nauta et al. 2023]
- **Completeness** means entire dynamics of the model. You *can* state/prove anything that's right.
 - E.g., it includes all the (important) features
 - Reasoning-completeness (global) versus output completeness (local) [Nauta et al. 2023]



Intuitions for these kinds of methods

- The explanation helps make a good prediction
- Removing aspects of the explanation will decrease quality of prediction
- Different protocols:
 - Addition or removal
 - Single or incremental
- **Non-sensitivity** ensures that a method assigns **zero-importance** only to features to which the model f is not functionally dependent on.
- **Sensitivity** assigns **non-zero-importance** to features which the model f is functionally dependent on.

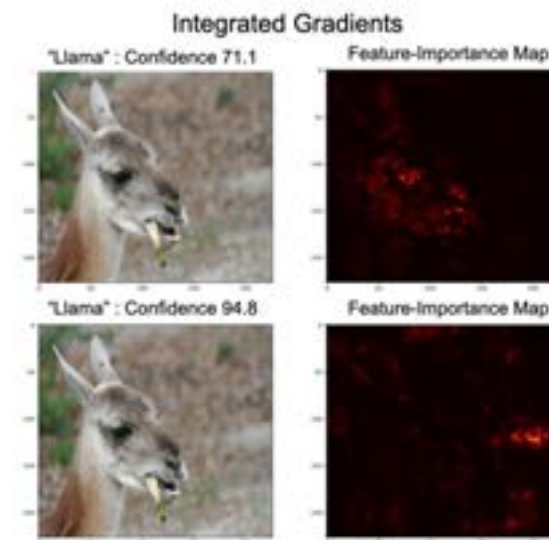
Content	
Correctness	Describes how faithful the explanation is wrt the black box
Output/Reasoning Completeness	Describes how much of the black box behavior is described in the explanation
Consistency	How deterministic and implementation-invariant the explanation
Continuity	How continuous and generalizable
Contrastivity	How discriminative wrt other targets
Co-variety complexity	How complex the (interaction of) features in the explanation are.
Presentation	
Compactness	The size.
Composition	Presentation format and organization.
Confidence	Presence and accuracy of probability information.
User	
Context	Relevance to user and their needs.
Coherence	How accordance with prior knowledge and beliefs.
Controllability	How interactive or controllable an explanation is for a user

What do these kinds of fidelity mean?

- Is your ground truth correct? Is it still correct (drift)?
- Complete? Out of distribution data, under-represented data/groups/languages
- Does your evaluation protocol capture the kind of fidelity that makes sense in practice?



Credit: Chaofen Chen, Duke University



[Ghorbani, Abid, Zhou; 2019]



Aash Ganesh

What if the predictive model itself is wrong?

- **Summarization** – temporally invariant models achieve competitive correlation close to the human baselines [Ganesh2024]
- **“Feature not known”** [Sivaprasad2025]
 - **Data shows the feature does not matter:** The feature is ignored because the data shows that the feature has minimal impact on the model's prediction.
 - **Insufficient data:** The feature may matter, but the model builders did not have sufficient high-quality training data to reliably model the feature's impact.
 - **Domain shift:** The world has changed since the model was built. This includes societal changes (e.g., legalisation of same-sex marriage) and changes in scientific knowledge and interventions.



Explanation fidelity is
not a desirable
property—it is the
property.

Evaluation
challenges are
restricted to (messy)
humans



Icon: Leremy - Flaticon

Why no consensus?



- Reflects diversity of explanatory goals, different user characteristics or even roles.
- The same explanations can do well on one explanatory goal, but poorly on another [Tintarev 2012, Cau2023, van der Waa 2021 AIJ Prominent Paper Award]
- Explanations' effectiveness is (presentation and content) particularly sensitive to: user, task, and context (c.f., Frontiers in AI [Tintarev 2025], XAI Manifesto 2.0 [Longo et al 2024])

Measuring Explanation Quality – a Path Forward

Prof. Nava Tintarev



Some issues are shared

- **Reliance on Ground truth** – with exceptions:
 - AXE measures how well a given explanation can help emulate model [Rawal et al, 2025]
 - Or instance-based explanations (e.g., diversity)
- **Reliance on underlying model**
 - Accuracy
 - And also proximity of reality
- **LLM-as-judge?**
 - More opacity
 - Historical biases and are prone to context shifts. Suffer from issues of errors, ambiguity, and too much homogeneity (model collapse).
 - Can be prompt-tuned and tools are being designed to support joint annotation



Users are hard.

- Let's do offline evaluation.
- BUT! The user, task, and context are assumed to be represented in the data
- We are making assumptions that we cannot check in using data-driven approaches.

Difference > measurement uncertainty?

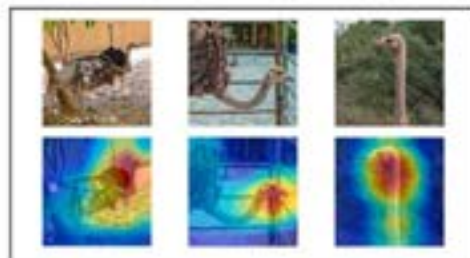
- “a common but incomplete understanding of empirical research in machine learning that leads to non-replicable results, makes findings unreliable, and threatens to undermine progress in the field” [Herrman et al, ICML'24]
- Source of reported improvements in performance [Shehzad and Jannach, Recsys'23]
- Unsupervised learning? Similar challenges have been identified for reinforcement learning [Henderson et al, AAAI'18]
- If changing the user, task, or context 'changes' explanation performance by 10%, it may not be meaningful to report a 2-3% performance improvement that does not control for these variables.



/imagesbykim@pixabay

Human / functional alignment?

- no differences in trust or satisfaction b/w XAI methods
- Grad-CAM improved user performance, while Guided Backpropagation did better according to mathematical metrics
- Mathematical metrics only partially related to user performance, sometimes counterintuitive



(a) Grad-CAM



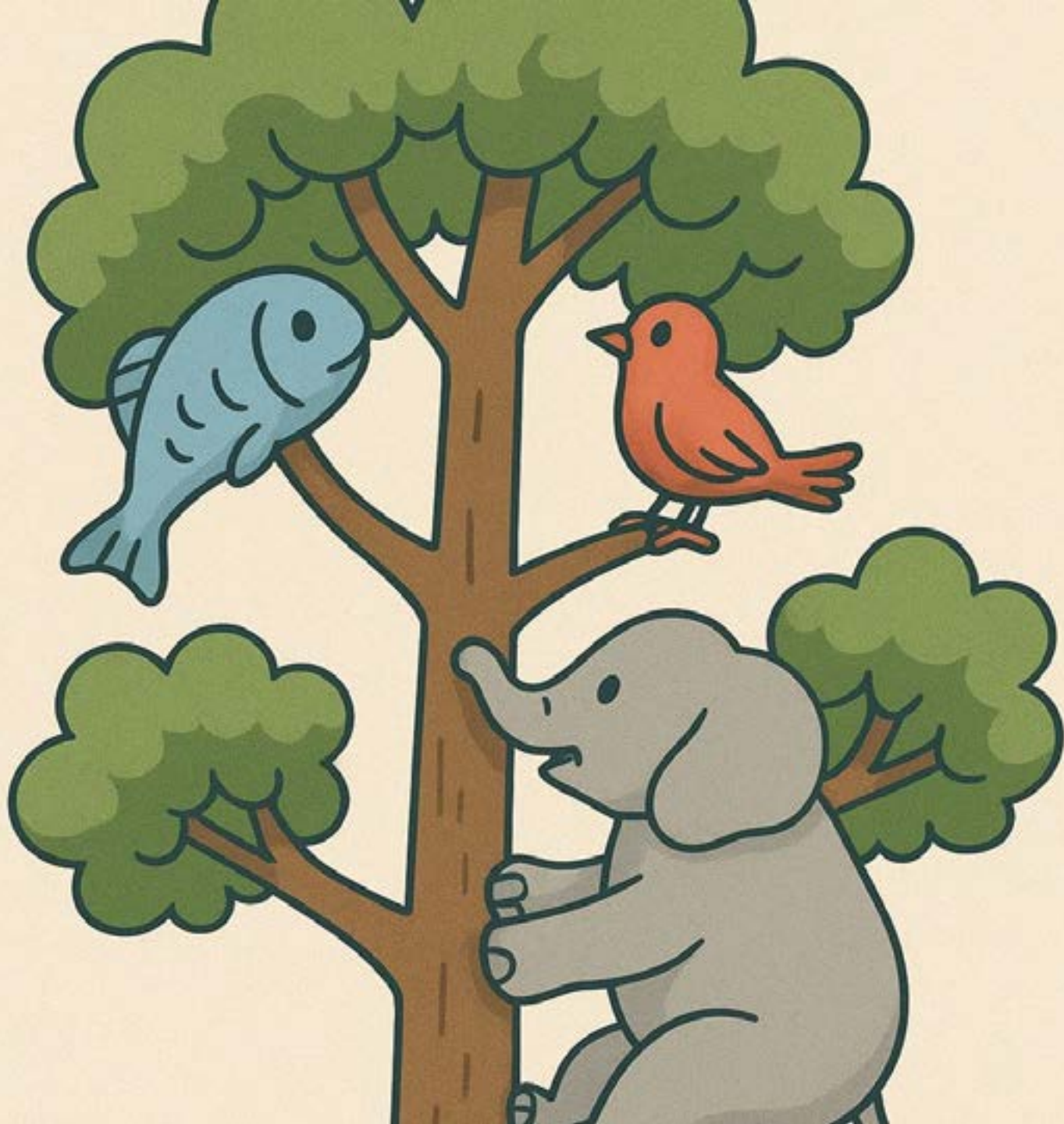
(b) Guided Backpropagation



(c) SHAP

Evaluation
challenges are
restricted to (messy)
humans





What do we need to adapt to? (Examples)

User – who are you explaining to?

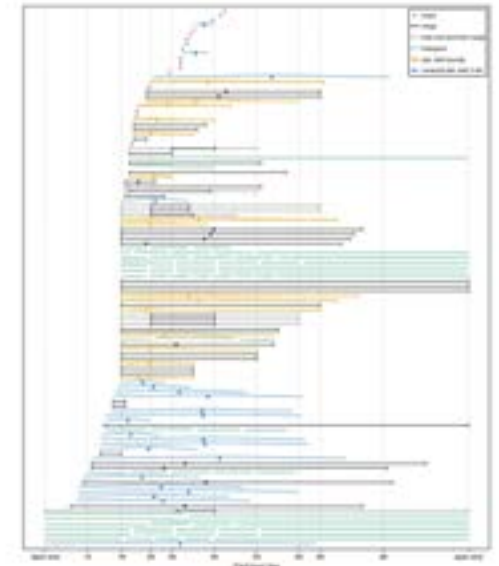
Whom do Explanations Serve? A Systematic Literature Survey of User Characteristics in Explainable Recommender Systems Evaluation

Authors:  Kathrin Wardatzky,  Oana Inel,  Luca Rossetto,  Abraham Bernstein

[Authors Info & Claims](#)

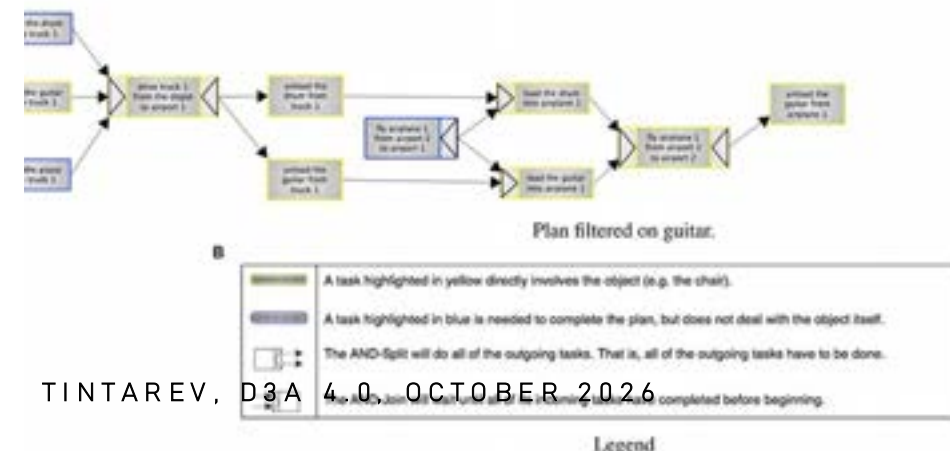
ACM Transactions on Recommender Systems, Volume 3, Issue 4 • Article No.: 49, Pages 1 - 35 • <https://doi.org/10.1145/3540000>

- Failed replication attempts
- Insufficient data on participants reported



The quality of an explanation is determined by the fit to the recipient(s)

- Visual working memory [Tintarev2016]
- Prior opinions [Cau2023]
- Expertise [Jin et al, 2018: UMAP and Recsys, UMUI 2019]
- Role [Szymanski2025]
- Privacy concern [Najafian2021]
- Overview of personalization recommender systems [Wardatzky2025]
- Sometimes, if you do not control for User, you will not see the effect of the Explanation!

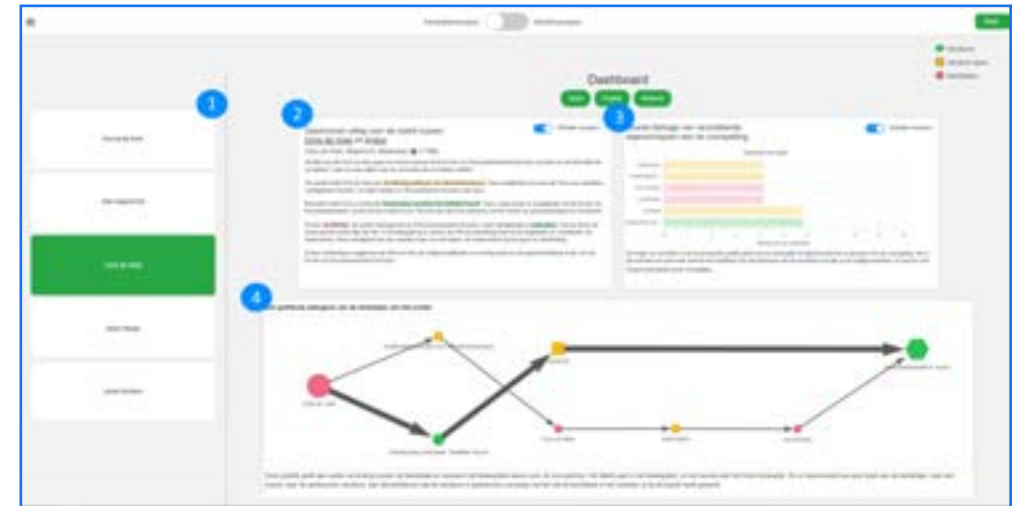


Explanations for different stakeholders

(Schellingerhout, Barile, Tintarev, WSDM under review)



- **Person-Job fit.** Collaboration with JobIndex + ITU (DDSA grant)
- Added value of GNN to unstructured data? How to link?
- LLMs to create bridge tuples (Qwen, Gemma, Llama)
 - CVs and vacancies (connectivity 20% - 80%) + bridge triples >95% for most model-prompt combinations
- **Rec accuracy** –
 - nDCG @10; 0.70-0.89 (c.f., company baseline 0.54)
- **User- (gender) and provider (rural) - fairness**
 - graph-based approaches match or outperform the text-only models



Task – what is the explanation doing?

Just Accepted

The Impact of Explanation Characteristics on the Effects of Recommender Systems Explanations

Authors: Kathrin Wardatzky, Oana Inel, Luca Rossetto, Athina Kyriakou, Abraham Bernstein [Authors Info & Claims](#)

ACM Computing Surveys • Accepted on 24 September 2026 • <https://doi.org/10.1145/3856207>

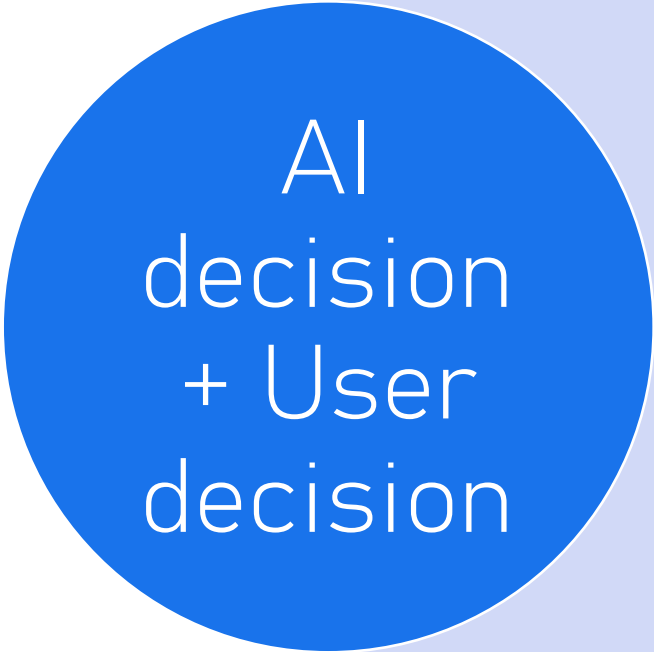
When Do Contrastive Explanations Really Matter in Recommender Systems?

Thi Ngoc Trang Tran* Institute of Software Engineering and Artificial Intelligence Graz University of Technology Graz, Steiermark, Austria trang.tran@tugraz.at	Sebastian Lubos Institute of Software Engineering and Artificial Intelligence Graz University of Technology Graz, Steiermark, Austria sebastian.lubos@tugraz.at	Alexander Felfernig Institute of Software Engineering and Artificial Intelligence Graz University of Technology Graz, Steiermark, Austria alexander.felfernig@tugraz.at
Mehrdad Rostami Faculty of Information Technology and Electrical Engineering University of Oulu Oulu, Finland Mehrdad.Rostami@oulu.fi	Viet-Man Le Institute of Software Engineering and Artificial Intelligence Graz University of Technology Graz, Steiermark, Austria v.m.le@tugraz.at	Damian Garber Institute of Software Engineering and Artificial Intelligence Graz University of Technology Graz, Steiermark, Austria damian.garber@tugraz.at

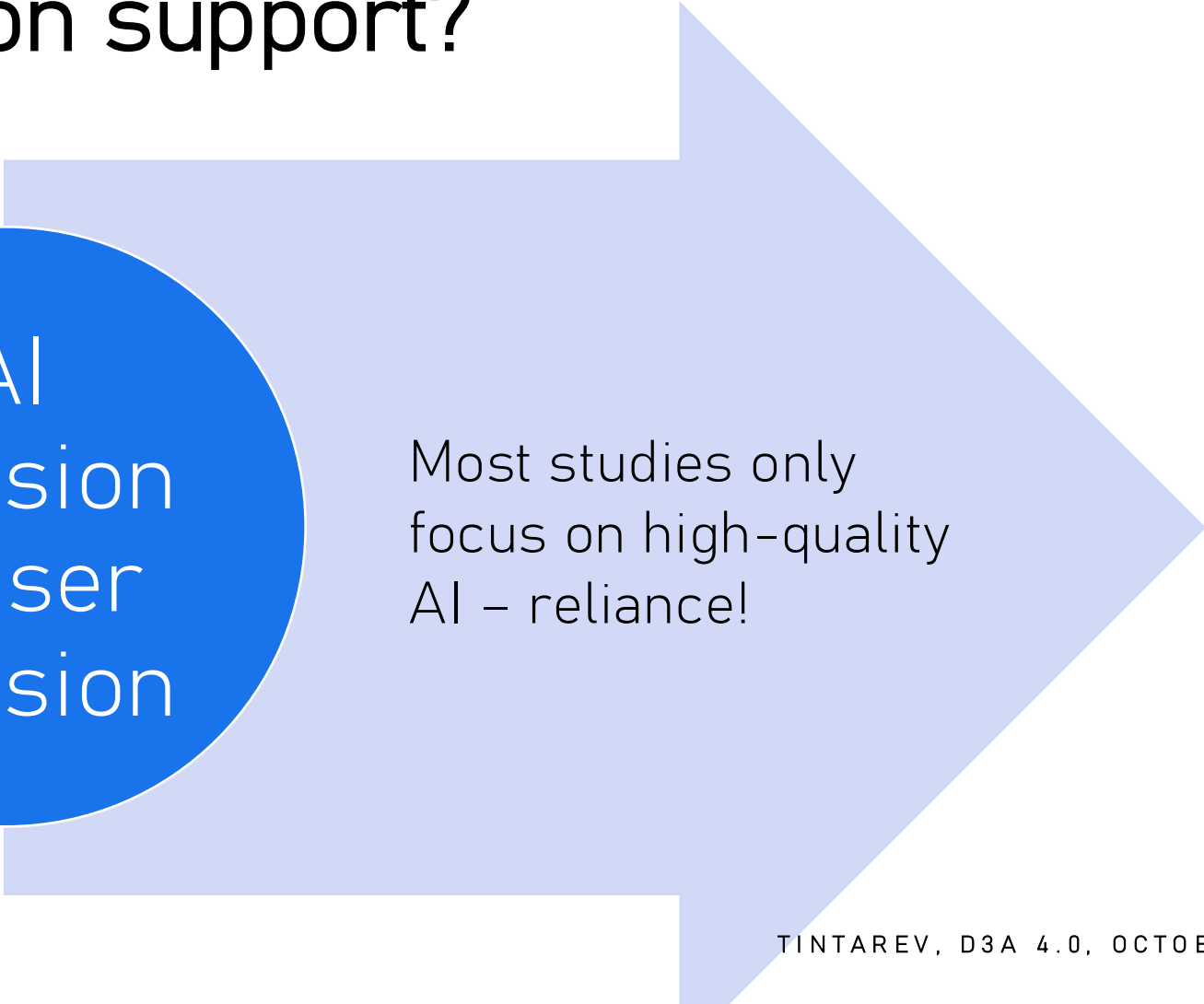
The **task** is designed to support certain kinds of quality measurements. This directly influences decisions about what stimuli are chosen and presented.

- Compared to another explanation or a system without explanation?
- Baseline explanations – efficient decision-making; contrastive explanations comparison and exploration, more cognitive processing [Tran et al, 2026]

Task – decision support?



AI
decision
+ User
decision



Most studies only
focus on high-quality
AI – reliance!

You can explain in many different ways!

Cau et al. IUI'23; TiiS'23



Federico Cau
(Cagliari/UM)



Davide Spano
(Cagliari)



Hanna Hauptmann
(Utrecht)

“Too much” trust

Agree more with the system

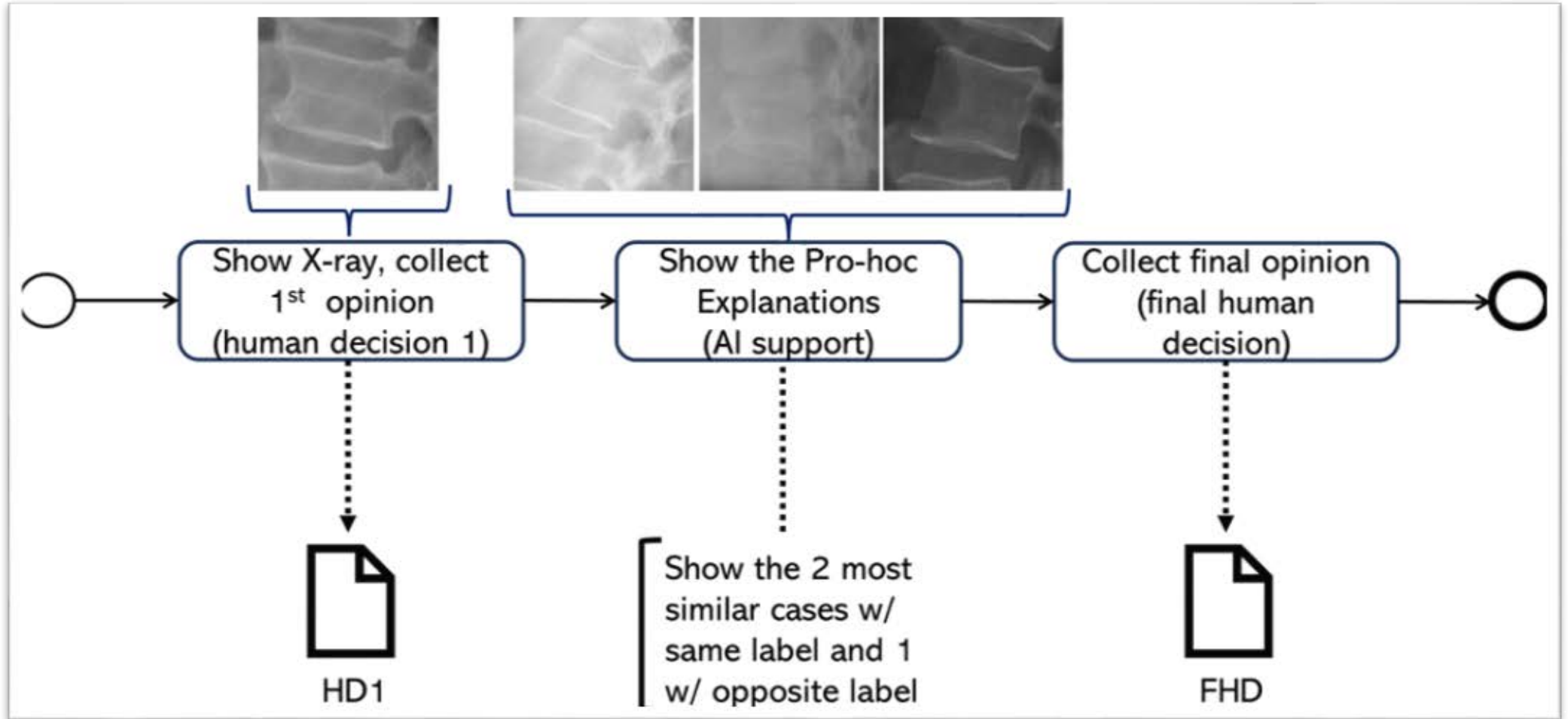
- System right – human right 94%.
- System wrong – human wrong 75%.
- Also agree when system is wrong
→ more mistakes for the person!



Inductive Explanation								
The AI generated this prediction learning from these examples:								
Example	Date	Price ?	MACD ?	ATR ?	EMV ?	RSI ?	Sentiment ?	Next week price ?
1	2018-06-07	539.507	-11.286	17.761	-7.281	42.927	-0.164	increase
2	2018-10-29	625.385	-8.183	18.369	-8.525	43.288	0.009	decrease
3	2019-10-14	438.583	-13.277	17.910	-8.187	44.041	-0.201	increase

Frictional and hypothesis driven XAI

Systems embedding cognitive friction for better decisions.



[Cabitza et al. 2024]

[Alshehri2025]



In which setting is the explanation being used?

The **context** influences how explanations are processed and which users and tasks are (most) relevant to study.



Context

- Time constraints – domain specific, and interacts with task (dual-task paradigm)
- Repetition/longitudinal effects – learning effects desirable?
- On-demand or never used?
- Other people. Contagion, privacy etc
- Relative performance (AI vs Human)

Hybrid team performance – relative accuracy?

WHEN AI OUTPERFORMED HUMANS ALONE, THEY FOUND LOSSES.

Article | [Open access](#) | Published: 28 October 2024

When combinations of humans and AI are useful: A systematic review and meta-analysis

[Michelle Vaccaro](#), [Abdullah Almaatoug](#) & [Thomas Malone](#) 

Nature Human Behaviour **8**, 2293–2303 (2024) | [Cite this article](#)

102k Accesses | 79 Citations | 456 Altmetric | [Metrics](#)

Investigating human–AI collaboration protocols in medical diagnosis

[Federico Cabitza](#)^{a b 1}  , [Andrea Campagner](#)^{b 1}, [Luca Ronzio](#)^c, [Matteo Cameli](#)^d, [Giulia Elena Mandoli](#)^d, [Maria Concetta Pastore](#)^e, [Luca Maria Sconfienza](#)^{b f}, [Duarte Folgado](#)^g, [Marília Barandas](#)^g, [Hugo Gamboa](#)^{g h}

WHEN AI OUTPERFORMED HUMANS ALONE, THEY FOUND GAINS.



Artificial Intelligence in Medicine

Volume 138, April 2023, 102506

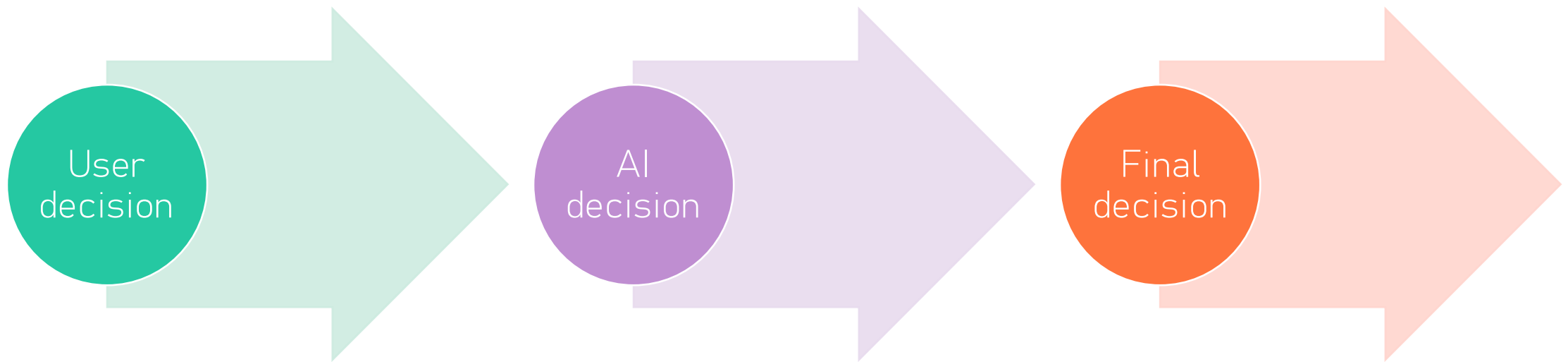


Research paper

Rams, hounds and white boxes: Investigating human–AI collaboration protocols in medical diagnosis

[Federico Cabitza](#)^{a b 1}  , [Andrea Campagner](#)^{b 1}, [Luca Ronzio](#)^c, [Matteo Cameli](#)^d, [Giulia Elena Mandoli](#)^d, [Maria Concetta Pastore](#)^e, [Luca Maria Sconfienza](#)^{b f}, [Duarte Folgado](#)^g, [Marília Barandas](#)^g, [Hugo Gamboa](#)^{g h}

Reliance patterns



Human judgment (HD1)	AI support (AI)	Final decision (FHD)	Reliance pattern	Other formulations	Facilitating Biases
wrong (0)	wrong (0)	wrong (0)	detrimental reliance	/	automation complacency
wrong (0)	wrong (0)	right (1)	beneficial under-reliance	/	extreme algorithmic aversion
wrong (0)	right (1)	wrong (0)	detrimental self-reliance	Detrimental overriding	self-anchoring bias
wrong (0)	right (1)	right (1)	beneficial over-reliance	Correct adherence	algorithm appreciation
right (1)	wrong (0)	wrong (0)	detrimental over-reliance	Detrimental adherence	automation bias
right (1)	wrong (0)	right (1)	beneficial self-reliance	Corrective overriding	algorithmic aversion
right (1)	right (1)	wrong (0)	detrimental under-reliance	/	extreme algorithmic aversion
right (1)	right (1)	right (1)	beneficial reliance	/	confirmation bias (in later cases)

[Campagner et al 2026]



Comparing apples with oranges. Now what?

- Systematic reporting, e.g., the human evaluation datasheet (Belz et al.)
- Replicability – model documentation form, benchmarkcards
- Meta-analysis

Thank you to these collaborators and others!



Francesco Barile



Mirela Popa



Roan Schellingerhout



Cedric Waterschoot



Adarsa Sivaprasad



Shabnam Najafian



Bulat Khaertdinov



Judith Masthoff



Ehud Reiter



Federico Cau



Aash Ganesh



Oana Inel



Yucheng Jin



Dina Zilbershtein



Katrien Verbert

Questions?

What is an explanation?

Is current evaluation sufficient?

Do our assumptions hold?

- Primacy of explanation fidelity
- Model Complexity is necessary for accuracy and bad for human understanding
- Evaluation challenges are restricted to (messy humans)

What does explanation effectiveness depend on?

Research agenda: more standardized evaluation protocols

Web references

- OXrec ethical concerns (2017): e.g., proxy variables and open tool without guidelines.
<https://pmc.ncbi.nlm.nih.gov/articles/PMC5493312/>
- Calibration from Sweden to NL (2019): “Our methods provide an approach to do so by using an incremental strategy of adjusting the baseline risk, and then recalibrating the linear predictor component of the prediction tool.” “...Revised thresholds of risk in the Dutch cohort.”
<https://www.nature.com/articles/s41598-018-37539-x>
- Documents from Dutch inspectorate (2026).
 - <https://www.inspectie-jenv.nl/actueel/nieuws/2026/02/12/risicovol-algoritmegebruik-door-reclassering>



References (1)

General-Purpose AI Code of Conduct. Model Documentation Form: <https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpa>

BenchmarkCards: <https://github.com/SokolAnn/BenchmarkCards>. Paper link: <https://arxiv.org/abs/2410.12974>

A. Alshehri, H. Alotaibi, T. Miller, and M. Vered. A hypothesis-driven approach to explainable goal recognition. In Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems, pages 105–114, 2025

Amparore, E., Perotti, A., & Bajardi, P. (2021). To trust or not to trust an explanation: using LEAF to evaluate local linear XAI methods. *PeerJ Computer Science*, 7, e479.

Balloccu, G., Boratto, L., Fenu, G., & Marras, M. (2022, July). Post processing recommender systems with knowledge graphs for recency, popularity, and diversity of explanations. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 646–656).

Barile, F., Najafian, S., Draws, T., Inel, O., Rieger, A., Hada, R., & Tintarev, N. (2021). Toward Benchmarking Group Explanations: Evaluating the Effect of Aggregation Strategies versus Explanation. *Perspectives@ RecSys*.

Been, K., Khanna, R. & Koyejo, O.O. "Examples are not enough, learn to criticize! Criticism for interpretability." Advances in Neural Information Processing Systems (2016)

Biran, O. and Courtenay, C. . "Explanation and justification in machine learning: A survey." *IJCAI-17 workshop on explainable AI (XAI)*. Vol. 8. No. 1. 2017.

Breiman, Leo. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)." *Statistical science* 16.3 (2001): 199–231.

Cabitza, F. et al. Never tell me the odds: Investigating pro-hoc explanations in medical decision making, *Artificial Intelligence in Medicine*, Volume 150, 2024



References (2)

- Cabitza, Federico, et al. "Rams, hounds and white boxes: Investigating human–AI collaboration protocols in medical diagnosis." *Artificial Intelligence in Medicine* 138 (2023): 102506.
- Guidotti, R. et al. "A survey of methods for explaining black box models." *ACM computing surveys (CSUR)* 51.5 (2018): 1–42.
- Campagner, Andrea, et al. "Under what influence: Measuring AI influence to fit user profiles in decision-making." *International Journal of Human-Computer Studies* (2026): 103763.
- Cau, F. M., Hauptmann, H., Spano, L. D., & Tintarev, N. (2023). Supporting High-Uncertainty Decisions through AI and Logic-Style Explanations. *IUI*.
- Cau, F., Hauptmann, H., Spano, D., & Tintarev, N. (2023). Effects of AI and Logic-Style Explanations on Users' Decisions under Different Levels of Uncertainty. *TiiS*.
- Draws, T., Liu, J., & Tintarev, N. (2020). Helping users discover perspectives: Enhancing opinion mining with joint topic models. *2020 International Conference on Data Mining Workshops (ICDMW)*, 23–30.
- Draws, T., Inel, O., Tintarev, N., Baden, C., & Timmermans, B. (2022). Comprehensive viewpoint representations for a deeper understanding of user interactions with debated topics. *ACM SIGIR Conference on Human Information Interaction and Retrieval*, 135–145.
- Draws, T., Ramamurthy, K. N., Soares, I. B., Dhurandhar, A., Padhi, I., Timmermans, B., & Tintarev, N. (2023). Explainable Cross-Topic Stance Detection for Search Results. *CHIIR*.
- Ghorbani, A., Abid, A., & Zou, J. (2019, July). Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 3681–3688).
- Ganesh, A., Popa, M., Odijk, D., & Tintarev, N. (2024, October). Does spatio-temporal information benefit the video summarization task?. In *CEUR Workshop Proceedings* (Vol. 3808). CEUR-WS.



References (3)

Gleicher M (2016) A framework for considering comprehensibility in modeling. *Big Data* 4(2):75–88

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5), 1–42.

Herrmann, M., Lange, F. J. D., Eggensperger, K., Casalicchio, G., Wever, M., Feurer, M., ... & Bischl, B. Position: *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235:18228–18247, 2024.

Henderson, P., et al. "Deep reinforcement learning that matters." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. No. 1. 2018.

Jin, Y., Tintarev, N., & Verbert, K. (2018). Effects of personal characteristics on music recommender systems with different levels of controllability. *Proceedings of the 12th ACM Conference on Recommender Systems*, 13–21

Jin, Y., Tintarev, N. & Verbert, K. Effects of individual traits on diversity-aware music recommender user interfaces." *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*. 2018.

Jin, Y., Tintarev, N., & Verbert, K.. Effects of personal characteristics on music recommender systems with different levels of controllability. *Proceedings of the 12th ACM Conference on Recommender Systems*, 13–21.

Jin, Y., Tintarev, N., Htun, N. N., & Verbert, K. (2020). Effects of personal characteristics in control-oriented user interfaces for music recommender systems. *User Modeling and User-Adapted Interaction*, 30, 199–249.

Kares, F., Speith, T., Zhang, H., & Langer, M. (2026). Are explainable AI (XAI) evaluation strategies aligned? Comparing subjective, objective, and mathematical evaluation measures using saliency maps. *Computers in Human Behavior*, 109163.

References (4)

Knijnenburg, Bart P., et al. "Explaining the user experience of recommender systems." *User modeling and user-adapted interaction* 22 (2012): 441–504.

Martin, F.J. 2009. Top 10 Lessons Learned Developing Deploying and Operating Real-world Recommender Systems. Proceedings of the Third ACM Conference on Recommender Systems (New York, NY, USA, 2009), 1–2

Najafian, S., Draws, T., Barile, F., Tkalcic, M., Yang, J., & Tintarev, N. (2021). Exploring User Concerns about Disclosing Location and Emotion Information in Group Recommendations. *ACM Hypertext*.

Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial intelligence* 267 (2019): 1–38.

Molnar, C. "Interpretable machine learning: A guide for making black box models explainable." *3rd Edition*, 2026.

Mulder, M., Inel, O., Oosterman, J., & Tintarev, N. (2021). Operationalizing framing to support multiperspective recommendations of opinion pieces. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 478–488.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Comput. Surv.*

Rawal, K., Fu, Z., Delaney, E., & Russell, C. (2025, June). Evaluating model explanations without ground truth. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (pp. 3400–3411).

🏆 Rieger, A., Draws, T., Theune, M., & Tintarev, N. (2021, August). This Item Might Reinforce Your Opinion: Obfuscation and Labeling of Search Results to Mitigate Confirmation Bias. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media* (pp. 189–199).



References (5)

Cynthia, R., Leo Breiman, the Rashomon Effect, and the Occam Dilemma, Proceedings of IWSM, 2025.

Schellingerhout, Roan, Volodymyr Medentsiy, & Maarten Marx. "Explainable Career Path Predictions using Neural Models." (2022). RecSys-in-HR Workshop. Recsys'22.

Shehzad, Faisal, and Dietmar Jannach. "Everyone's a winner! on hyperparameter tuning of recommendation models." *Proceedings of the 17th ACM Conference on Recommender Systems*. 2023.

Sivaprasad, A., Reiter, E., McLernon, D., Tintarev, N., Bhattacharya, S., & Oren, N. (2025, August). Patient-centred explainability in IVF outcome prediction. In *International Conference on AI in Healthcare* (pp. 87-99). Cham: Springer Nature Switzerland.

Sullivan E., et al. (2019). Reading news with a purpose: explaining user profiles for self-actualization. *Adjunct Proceedings of the Conference on User Modeling, Adaptation and Personalization (UMAP)*.

Lorenz-Spreen, Philipp, et al. "Boosting people's ability to detect microtargeted advertising." *Scientific Reports* 11.1 (2021): 1-9.

Semenova, L., Rudin, C., & Parr, R. (2022, June). On the existence of simpler machine learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1827-1858).

Tintarev, N., Rostami, S., & Smyth, B. (2018). Knowing the Unknown: Visualising Consumption Blind-Spots in Recommender System. *ACM Symposium on Applied Computing (SAC)*.

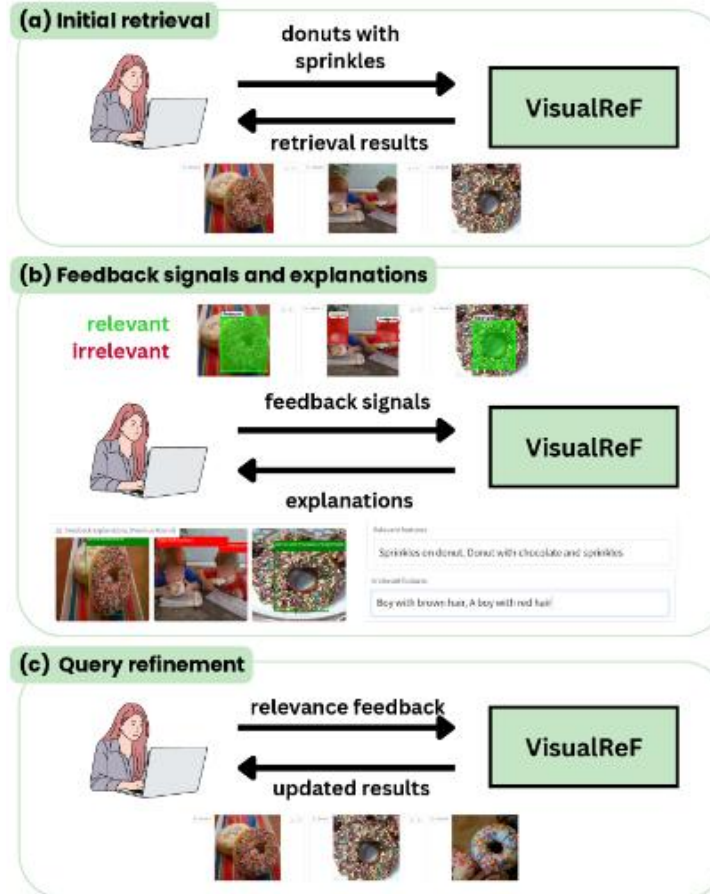
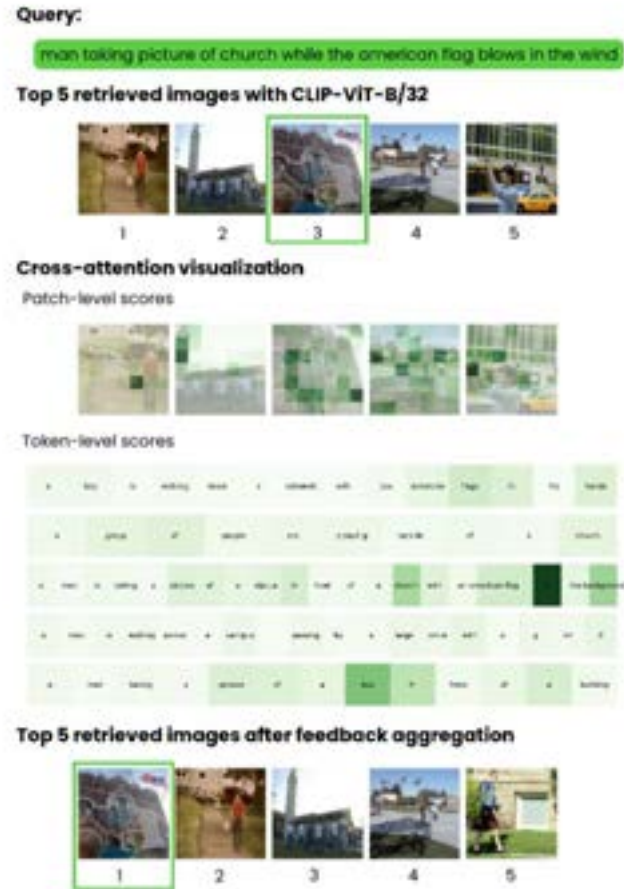
Tintarev, N., & Masthoff, J. (2022). Beyond Explaining Single Item Recommendations. In: Ricci, F., Rokach, L., Shapira, B. (eds) *Recommender Systems Handbook*. Springer, New York, NY.
https://doi.org/10.1007/978-1-0716-2197-4_1



References (6)

- Tintarev, N., & Masthoff, J. (2012). Evaluating the effectiveness of explanations for recommender systems: Methodological issues and empirical studies on the impact of personalization. *User Modeling and User-Adapted Interaction*, 22, 399–439.
- Tintarev, N., Kang, B., Höllerer, T., & O'Donovan, J. (2015). Inspection Mechanisms for Community-based Content Discovery in Microblogs. *Joint Workshop on Interfaces and Human Decision Making for Recommender Systems in Conjunction with Recsys.*, 1438, 21–28.
- Tintarev, N., O'Donovan, J., & Felfernig, A. (2016). Introduction to the special issue on human interaction with artificial advice givers. *ACM Transactions on Interactive Intelligent Systems (TiIS)*, 6(4), 1–12.
- Tran, T. N. T., Lubos, S., Felfernig, A., Rostami, M., Le, V. M., & Garber, D. (2026, September). When Do Contrastive Explanations Really Matter in Recommender Systems?. In *Proceedings of the 20th ACM Conference on Recommender Systems* (pp. 611–621).
- Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12), 2293–2303.
- Van Der Waa, Jasper, et al. "Evaluating XAI: A comparison of rule-based and example-based explanations." *Artificial intelligence* 291 (2021): 103404. Zipkin, D.A. et al. 2014. Evidence-based risk communication: a systematic review.
- Wardatzky, Kathrin, et al. "Whom do Explanations Serve? A Systematic Literature Survey of User Characteristics in Explainable Recommender Systems Evaluation." *ACM Transactions on Recommender Systems* 3.4 (2025): 1–35.
- Xin, Rui, et al. "Exploring the whole rashomon set of sparse decision trees." *Advances in neural information processing systems* 35 (2022): 14071–14084.
- Xian, Y., Fu, Z., Muthukrishnan, S., De Melo, G., & Zhang, Y. (2019, July). Reinforcement knowledge graph reasoning for explainable recommendation. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval* (pp. 285–294).

Interactive Image Search



- Fine-grained, region-level visual relevance feedback into VLMs,
- Iterative query refinement improving interpretability and user control.
- Published in [RecSys'25](#), [WACV 2026](#)
- Follow-up planned for **integration for in-video ads, image selection (products within RTL/DPG Media)**



Human vs functional

HUMAN-CENTERED

- Bias toward simpler explanations
- Closer to the user's expectations
- Hard to separate explanation and model correctness

FUNCTIONALITY GROUNDED

- Limited by system functionality
 - May not be understandable
 - May not be useful or actionable
- Closer to the system's behavior
- (Valid in early development)