XAI ROADBLOCKS, TRENDS, AND DIRECTIONS

Rosina Weber, PhD

College of Computing & Informatics

Drexel University

A BIT ABOUT MYSELF

Major in Finance, Economics

Engineering, Operations Research

Engineering, Artificial intelligent methods

Use-inspired research in AI

Case-based reasoning

Textual methods

Knowledge management

Explainable AI, Interpretable machine learning

Multiple research projects related to XAI

Chaired 4 wksp last 4 years

Information Science

Computer Science affiliation pending

WHAT WILL WE TALK ABOUT?

Intro	Roadblocks pose risks that can be suicidal to the field
Roadblock I	Lack of consensus on foundational term definitions
Roadblock II	Misleading motivations
Roadblock III	Dependencies on multiple disciplines
Trends	No representative trends to eliminate roadblocks
Summary directions	Clearing the fog so we can see the road

ROADBLOCKS POSE RISKS THAT CAN BE SUICIDAL TO THE FIELD

- Lack of consensus on foundational term definitions
- Misleading motivations
- Dependencies on multiple disciplines

Image by Al Seeger from Pixabay

LACK OF CONSENSUS ON FOUNDATIONAL TERM DEFINITIONS

- Scope of the field
- Interpretable model
- Explanation

Image by Herbert Bieser from Pixabay

LACK OF CONSENSUS ON FOUNDATIONAL TERM DEFINITIONS: SCOPE OF THE FIELD

EXPLAINABLE AI (XAI) AND INTERPRETABLE MACHINE LEARNING (IML)

W.R.T. USERS

IML

ML experts

XAI

All users

W.R.T. METHODS

IML

Neural networks

Decision trees

Deep learning

Reinforcement learning

Neural probabilistic language models

XAI

Planning

Bayesian nets

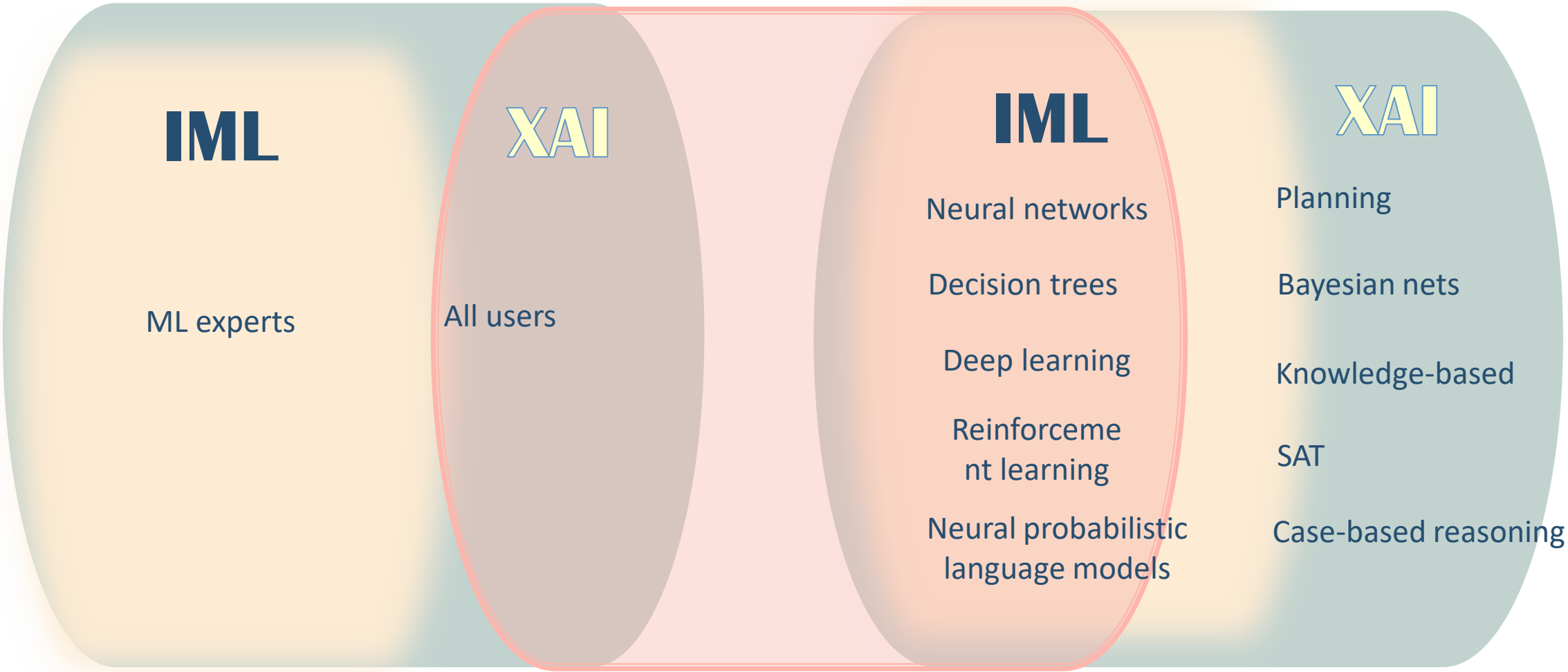
Knowledge-based

SAT

Case-based reasoning

EXPLAINABLE AI (XAI) AND INTERPRETABLE MACHINE LEARNING (IML)

W.R.T. USERS



W.R.T. METHODS

EXPLAINABLE AI (XAI) AND INTERPRETABLE MACHINE LEARNING (IML)

The prestigious author Rudin, recipient of the AAAI 2022 Squirrel Award, stated in 2022, respectively, p. 9, p. 61

Hence, this survey concerns the former. This is not a survey on Explainable AI (XAI, where one attempts to explain a black box using an approximation model, derivatives, variable importance measures, or other statistics), it is a survey on *Interpretable Machine Learning* (creating a predictive model that is not a black box). Unfortunately, these topics are much too often lumped together within the misleading term "explainable artificial intelligence" or "XAI" despite a chasm separating these two concepts [250]. Explainability and interpretability techniques are not alternative choices for many real problems, as the recent surveys often imply; one of them (XAI) can be dangerous for high-stakes decisions to a degree that the other is not.

Interpretable ML is not a subset of XAI. The term XAI dates from ∼2016, and grew out of work on function approximation; i.e., explaining a black box model by approximating its predictions by a simpler model [e.g., 70, 69], or explaining a black box using local approximations. Interpretable ML also has a

the reasoning processes of black box models. Explaining black boxes, rather than replacing them with interpretable models, can make the problem worse by providing misleading or false characterizations [250, 173, 171], or adding un-

EXPLAINABLE AI (XAI) AND INTERPRETABLE MACHINE LEARNING (IML)

The implication of Rudin's review is that XAI is unnecessary because its methods are dangerous and are not needed, as it is always better to use an interpretable model than try to explain a non-interpretable one (2022).

DIRECTION I: Engage the XAI community to describe and make explicit their broad view of the sub-field of XAI.

LACK OF CONSENSUS ON FOUNDATIONAL TERM DEFINITIONS: INTERPRETABLE MODEL

Authors gave definitions for interpretability but did not describe any scientific methodology in their support

(e.g., Schielzeth 2010, Lou et al. 2012, Doshi-Velez and Kim 2017, Drumond et al. 2017, Zhang & Zhu 2018, Gilpin et al. 2018, Lipton 2018, Chen et al. 2019, Lalor & Guo 2022, Rudin et al. 2022).

Some AI methods are referred to as interpretable (e.g., decision trees) but authors have argued in favor of explaining such 'interpretable' methods (e.g., Izza, Ignatiev, Marques-Silva 2020 and 2022)

Rudin et al. (2022) proposes that "an interpretable model is constrained, following a domain-specific set of constraints that make reasoning processes understandable (p.11)."

DIRECTION II: Investigate a precise means to describe and recognize interpretability aspects of a model both at the global and local levels so it can be determined when explanation methods for the model are needed.

LACK OF CONSENSUS ON FOUNDATIONAL TERM DEFINITIONS: EXPLANATION

LACK OF CONSENSUS ON FOUNDATIONAL TERM DEFINITIONS: EXPLANATION AND PROBLEM WITH EVALUATIONS

The literature reveals disagreement in the literature on multiple aspects of explanations (e.g., Buchholz 2022 states that authors disagree on what to explain, to whom, what methods to use, and why) .

Paper title: ""Explanation" is Not a Technical Term: The Problem of Ambiguity in XAI" Gilpin et al. 2022

A definition for explanation was given by social scientists (Mueller et al. 2019), particularly by psychologists who study trust.

The contributions from social science field have a place in evaluating the user aspects; they should not stop us from advancing the computing aspects of XAI methods.

For example, for evaluation, the claim is that we cannot use benchmark datasets because each user requires a different explanation Yang, Du, and Hu (2019).

Various authors agree that the lack of ground-truth for evaluating explanations is a limitation (Tomsett et al. 2019; Hooker et al. 2019; Yang, Du, and Hu 2019; Montavon 2019).

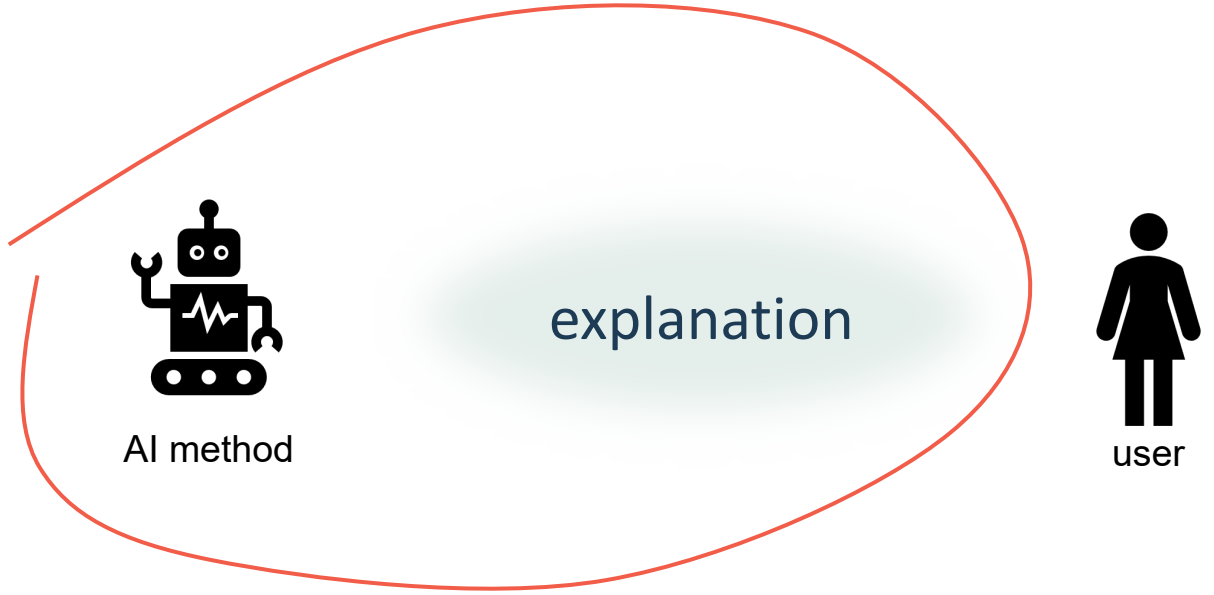
Many others have proposed datasets to evaluate explanations (Barr et al. 2020, Mahajan, Tan, Sharma 2019, Yang & Kim 2019, Amiri et al. 2020, Zhou, Booth, Ribeiro, Shah 2022).

DIRECTION III: Investigate approaches to evaluate the competence of XAI methods to produce each type of information content that can have explanatory value including benchmark datasets.

EXAMINE THE CONTEXT OF EXPLANATIONS

- Explanation context
- AI method context
- Human in a decision-making context

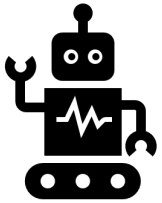
EXPLANATION CONTEXT



An AI method should be able to explain its decisions

AI METHOD

AI is a field of study dedicated to methods that produce rational decisions via computations of tasks such as planning, classification, and vision.

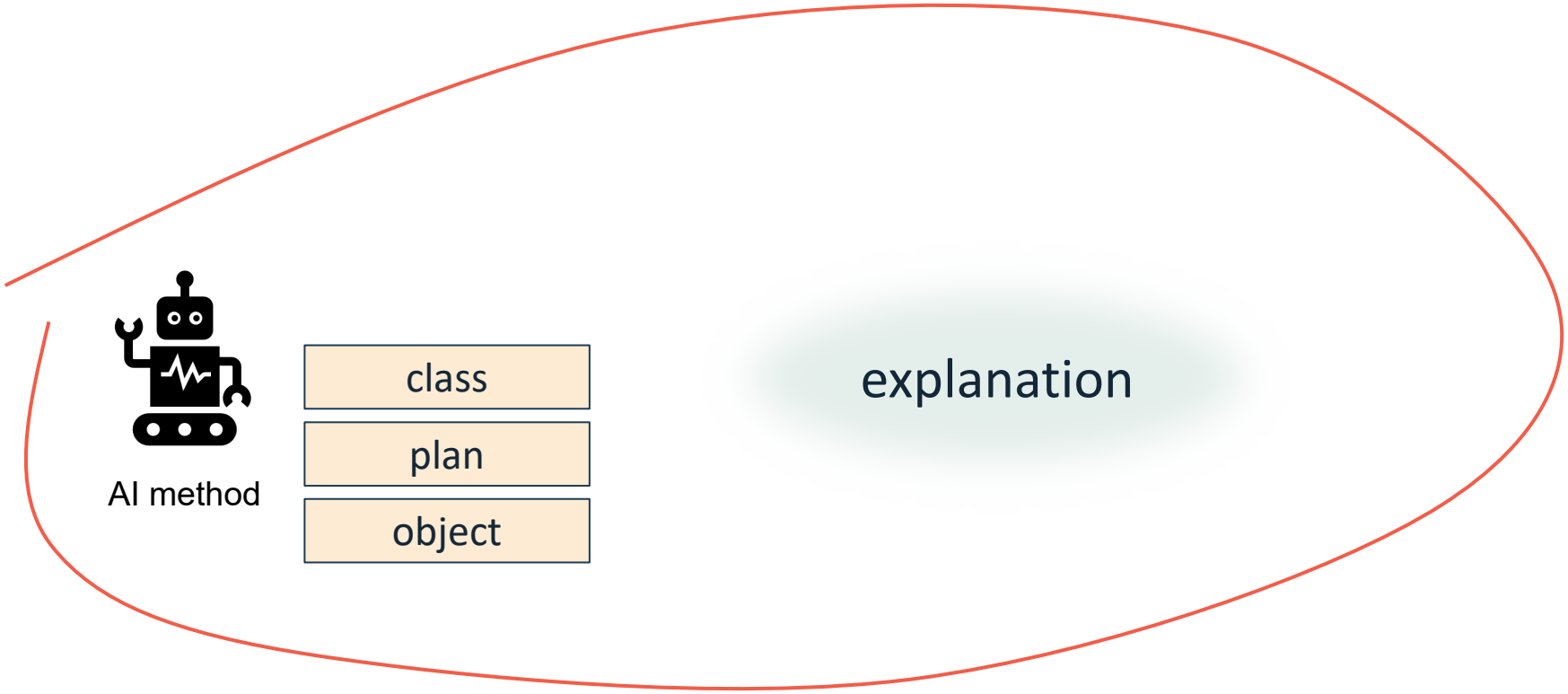


AI method

output

explanation

AI METHOD

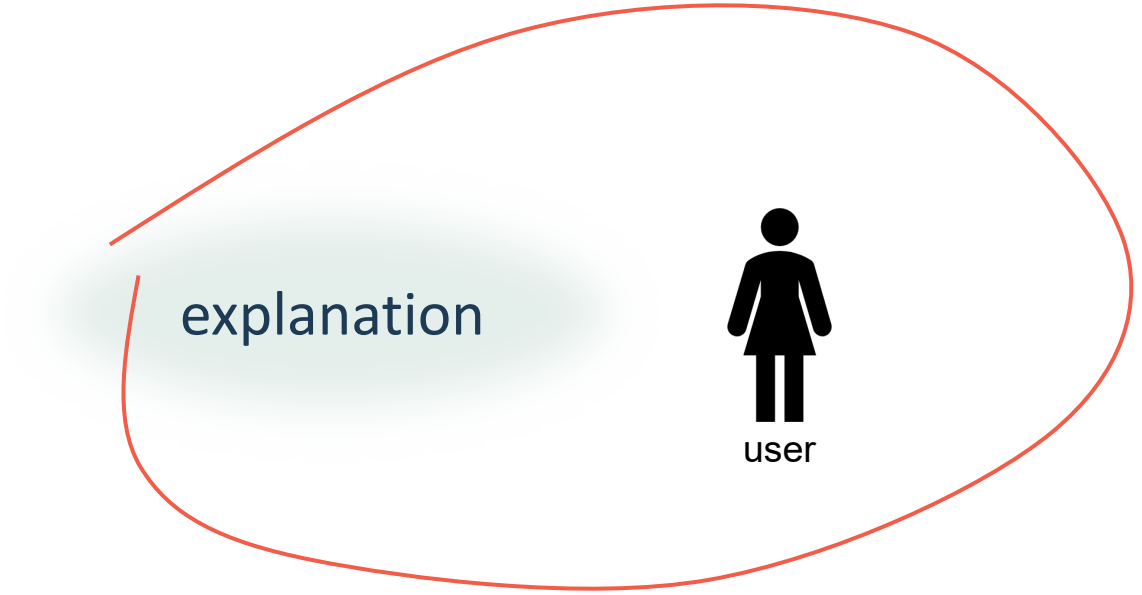


The explanation context of an AI method can be seen as the set of information contents offered as output in addition to its precisely defined output.

The set of information contents to populate the explanation context is limited to the outputs produced by the AI method (e.g., global importance factors—make a model interpretable) and the information contents produced by all the compatible/applicable XAI methods.

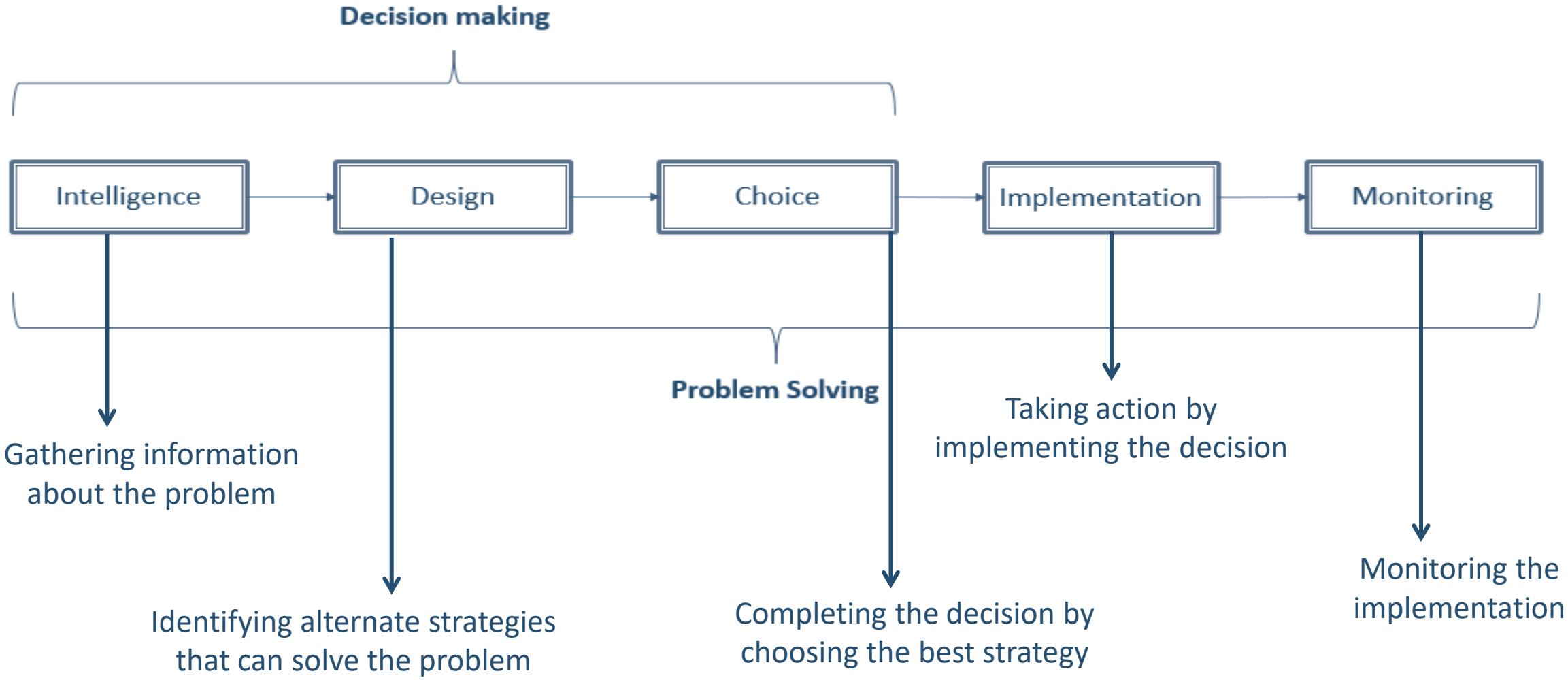
DIRECTION IV: Investigate how to precisely define the explanation context from the perspective of the AI method.

USER IN A DECISION-MAKING CONTEXT



DECISION-MAKING MODEL

Simon (1957) Huber (1980)



HUMAN USERS MAKE THE DECISIONS OR OBTAIN DECISIONS FROM HUMANS

human

explanation

decision maker is human

HUMAN USERS OBTAIN DECISIONS FROM HUMANS

Human decision makers communicate multiple information types and not always include an explicit decision

human

Humans describe problems

Human decision maker communicates decision

decision maker is human

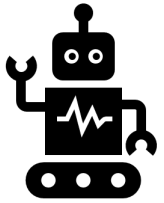
HUMAN USERS OBTAIN DECISIONS FROM OTHERS

AI methods communicate the output of an AI task, e.g., a class, prediction, or plan such as "application rejected".

human

Humans describe problems

AI methods return decisions via solutions to AI tasks, e.g., class is mammal, plus aspects that make the model interpretable



decision maker is AI method

Human decision makers communicate multiple information types and not always include an explicit decision

human

Humans describe problems

Human decision maker communicates decision

decision maker is human

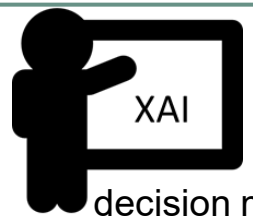
HUMAN USERS OBTAIN DECISIONS FROM OTHERS

There is a limited number of information types that can be produced as outputs of XAI methods, they are feature attributions (from which visualizations like salience maps can be bult), instance attributions, examples, rules, and counterfactuals.

human

Humans describe problems

AI methods return decisions as in the box below, e.g., class is mammal, number of legs has highest contribution followed by lacto-beverage production



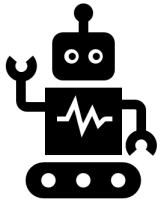
decision maker is XAI-equipped AI method

AI methods communicate the output of an AI task, e.g., a class, prediction, or plan such as "application rejected". This is the only part that if formalized in the design of the AI method.

human

Humans describe problems

AI methods return decisions via solutions to AI tasks, e.g., class is mammal, plus aspects that make the model interpretable



decision maker is AI method

Human decision makers communicate multiple information types and not always include an explicit decision

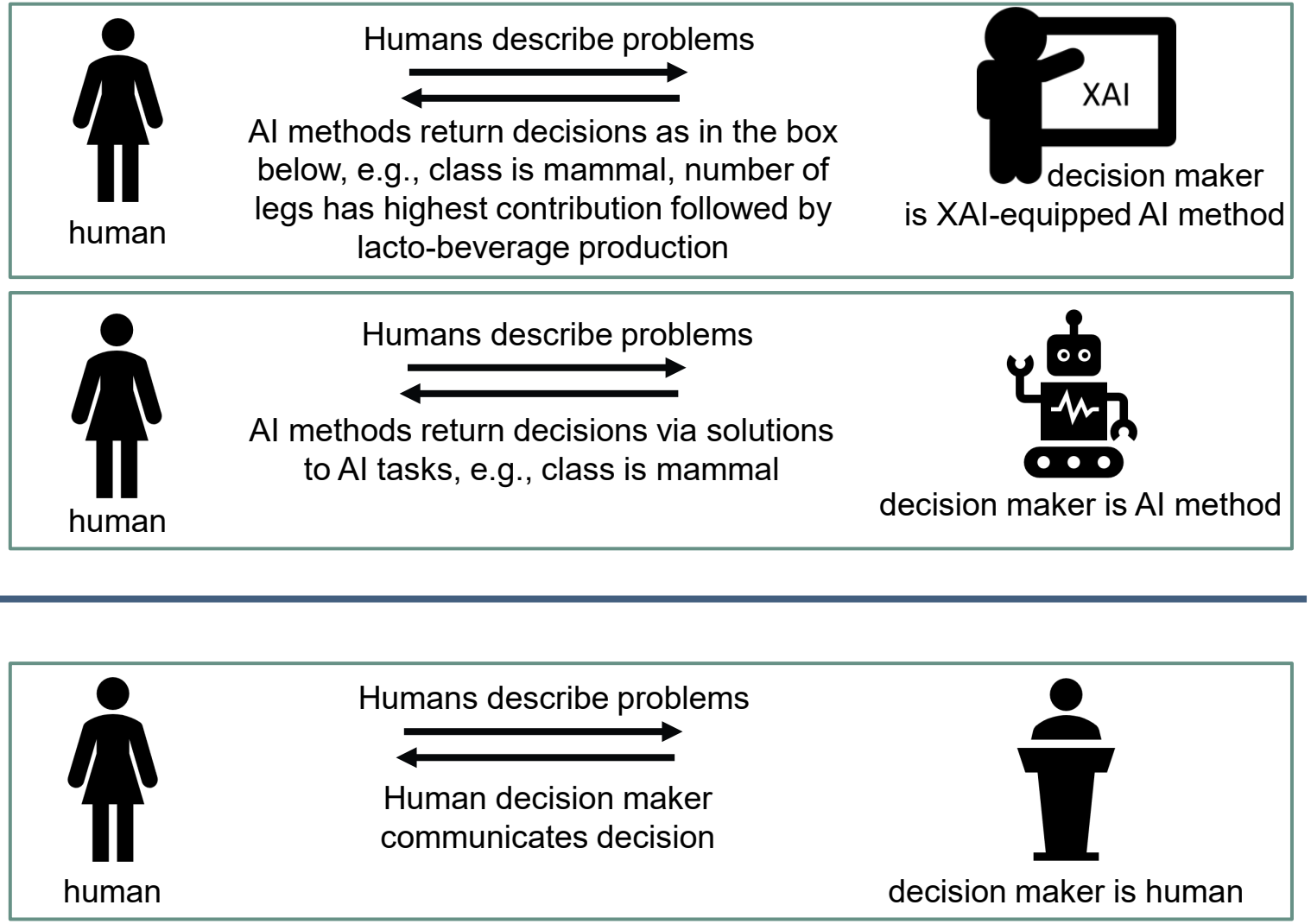
human

Humans describe problems

Human decision maker communicates decision

decision maker is human

EXPLANATION CONTENTS PRODUCED BY AI METHODS DO NOT MATCH EXPLANATION CONTENTS USERS EXPECT



LACK OF CONSENSUS ON FOUNDATIONAL TERM DEFINITIONS: EXPLANATION

Lim 2012, Nunes and Jannach 2017, Chari et al. 2020 provide multiple information contents users expect as explanations that AI methods can use to explain themselves. Other contributions come from Gunning (2017), Gilpin et al. (2018), and Gallant (1988).

How certain are you?

What data did you use?

What would it take for me to get another decision?

Are you sure it is not something else ?

How did you make that decision?

Why did you make that decision?

What background and complementary information do you use?

When do you succeed?

When do you fail?

When can I trust you?

How do I correct an error?

What happens before you make a decision?

Is there scientific evidence for this result?

What do you know?

LACK OF CONSENSUS ON FOUNDATIONAL TERM DEFINITIONS: EXPLANATION

Considering the information produced as output, XAI methods can be grouped by:

Feature attribution:
(additive) SHAP Lundberg & lee (2017), LIME Ribeiro, Singh, and

Guestrin (2016), DeepLift Shrikumar, Greenside, & Kundaje (2017), LRP Bach et al. (2015);
(non-additive) GradCAM Selvaraju et al. (2017), Integrated Gradients Sundararajan, Taly, and Yan (2017), SmoothGrad Smilkov et al. (2017)

Instance attribution:
Influence functions Koh & Liang (2017), representer points Yeh et al. (2018), HYDRA Chen et al. (2021)

Example-based, Prototype-based:
CBR, CBR Twins, Prototypes, Bayesian-based, etc.

Counterfactuals:
DICE, MACE, VLK, case-based (Smyth): See Keane et al. 2021 for a review

Rules, paths, etc.

Rule extractors, decision-tree paths Izza, Ignatiev, & Marques-Silva (2020)

Feature attribution:

(the role played by different features in classifying an instance.)

Instance attribution:

(the role played by different training instances in classifying an instance.)

Example-based:

(instances that are similar to the instance being explained.)

Counterfactuals:

(neighbor instances that are produce different outcome class)

Why did you make that decision?

How did you make that decision?

What data did you use?

How certain are you?

Are you sure it is not something else ?

What would it take for me to get another decision?

LITERATURE DESCRIBES QUESTIONS USERS WOULD LIKE ANSWERED THAT ARE NOT PROVIDED BY XAI METHODS

When do you succeed?

When do you fail?

When can I trust you?

How do I correct an error?

What happens before you make a decision?

Is there scientific evidence for this result?

What do you know?

What background and complementary information do you use?

DIRECTION V:

Investigate methods to produce **the information contents** users want that are not yet available.

MISLEADING MOTIVATIONS

Accuracy interpretability tradeoff

Users do not trust AI agents because they are black-boxes

Image by Pablo Jimeno from Pixabay

ACCURACY INTERPRETABILITY TRADEOFF

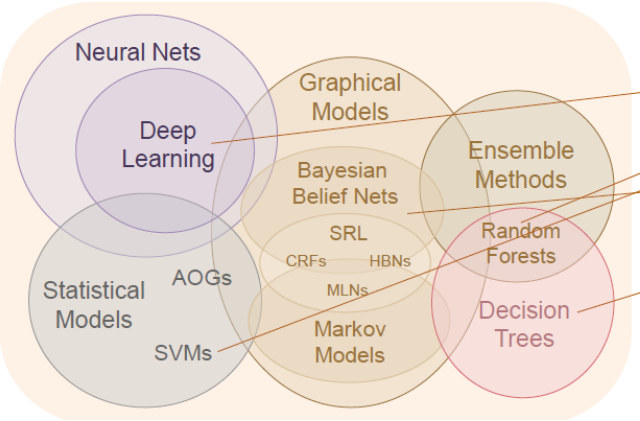
ACCURACY INTERPRETABILITY TRADEOFF

DARPA Explainable AI – Performance vs. Explainability

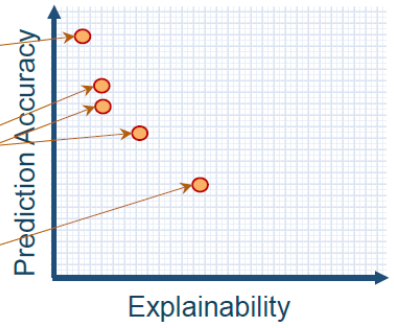
New Approach

Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



Explainability (notional)



Gunning et al XAI—Explainable artificial intelligence.pdf - Adobe Acrobat Pro DC (32-bit)

File Edit View E-Sign Plug-Ins Window Help

Home Tools Gunning et al XAI... ×

1 / 3

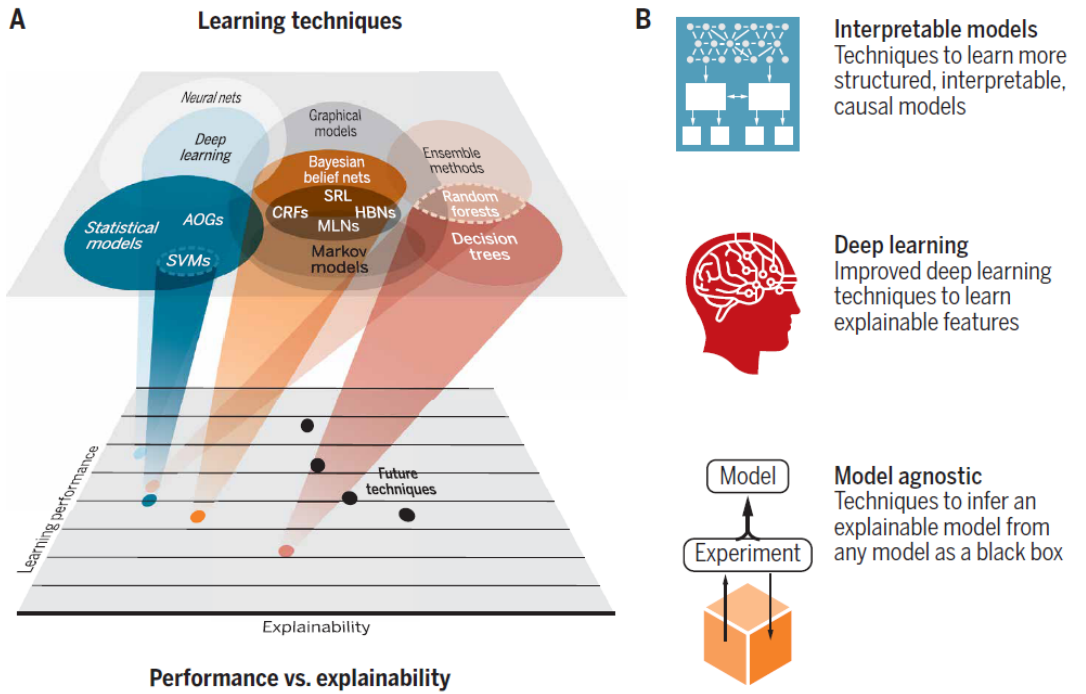


Fig. 1. Performance versus explainability tradeoff for ML techniques. (**A**) Learning techniques and explainability. Concept adapted from (*9*). (**B**) Interpretable models: ML techniques that learn more structured, interpretable, or causal models. Early examples included Bayesian rule lists, Bayesian program learning, learning models of causal relationships, and using stochastic grammars to learn more interpretable structure. Deep learning: Several design choices might produce more explainable representations (e.g., training data selection, architectural layers, loss functions, regularization, optimization techniques, and training sequences). Model agnostic: Techniques that experiment with any given ML model, as a black box, to infer an approximate explainable model.

AUTHORS WHO DEMONSTRATE THE TRADE-OFF DOES NOT HOLD

"These two data extremes show that in machine learning, the dichotomy between the accurate black box and the less-accurate interpretable model is false" Rudin et al. 2022.

Murdoch et al. 2019;

Dziugaite, Ben-David and Roy 2020;

Rudin et al. 2022;

Bell et al. 2022;

Ahmed et al. 2022

USERS DO NOT TRUST AI AGENTS BECAUSE THEY ARE BLACK-BOXES

Photo by Max Fischer:

Feature attribution:

(the role played by different features in classifying an instance.)

Instance attribution:

(the role played by different training instances in classifying an instance.)

Example-based:

(instances that are similar to the instance being explained.)

Counterfactuals:

(neighbor instances that are produce different outcome class)

Why did you make that decision?

How did you make that decision?

What data did you use?

How certain are you?

Are you sure it is not something else ?

What would it take for me to get another decision?

LITERATURE DESCRIBES QUESTIONS USERS WOULD LIKE ANSWERED THAT ARE NOT PROVIDED BY XAI METHODS

When do you succeed?

When do you fail?

When can I trust you?

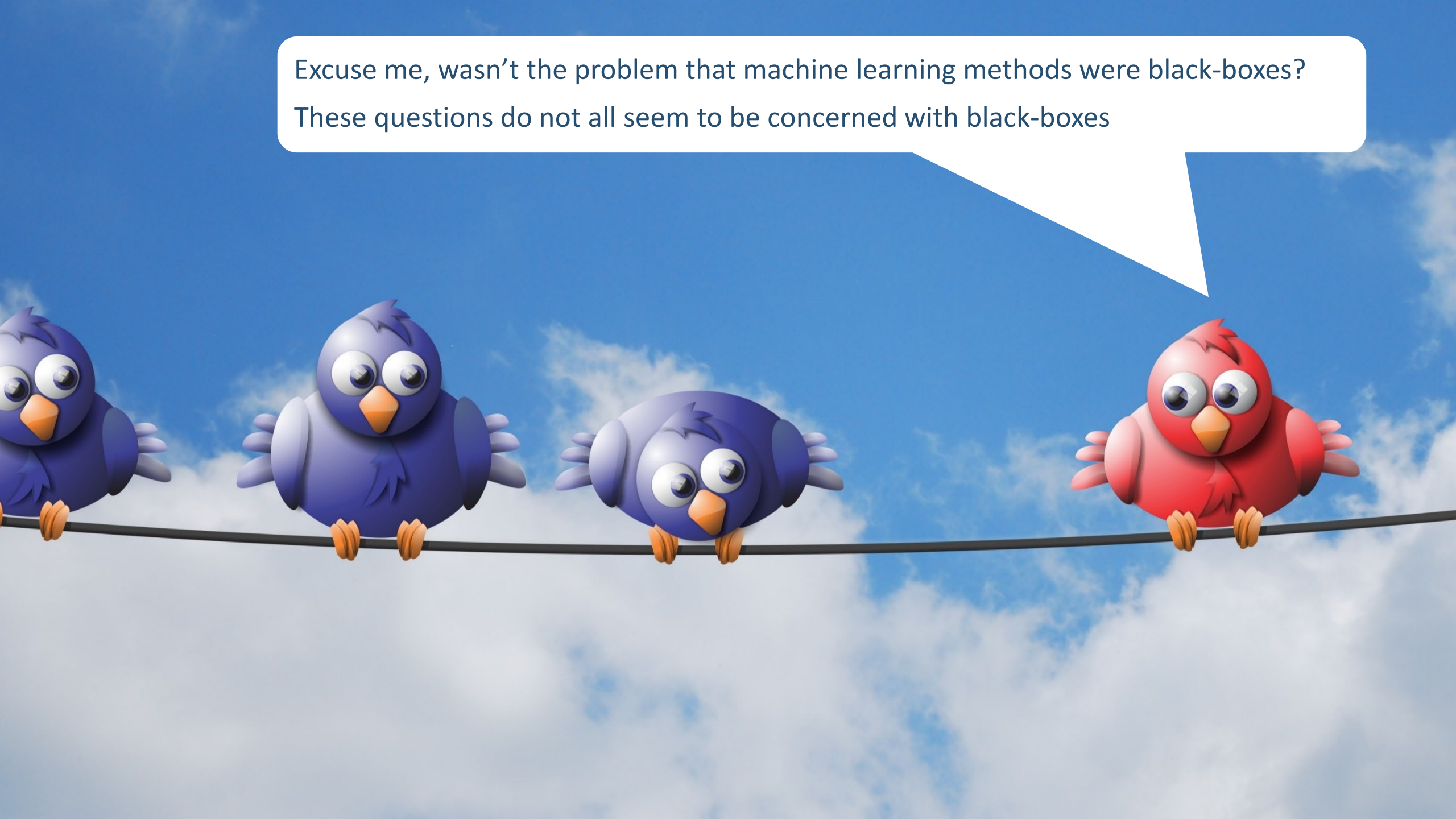
How do I correct an error?

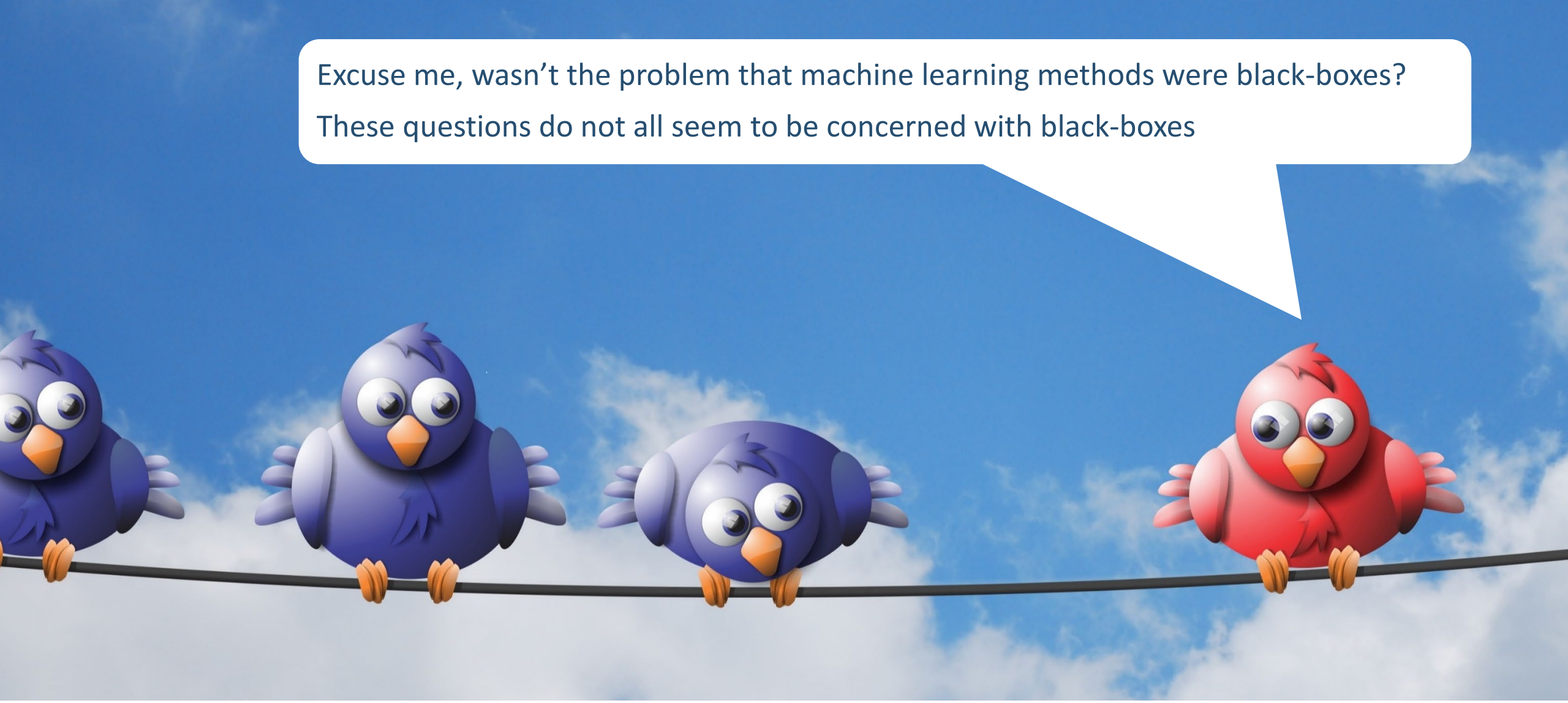
What happens before you make a decision?

Is there scientific evidence for this result?

What do you know?

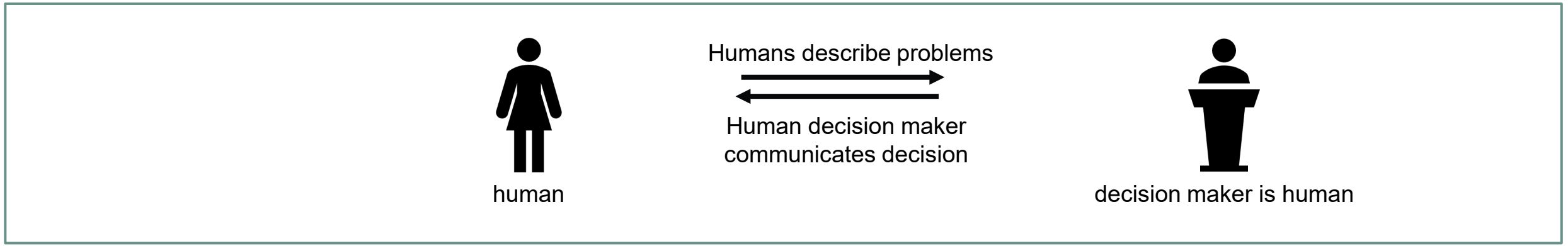
What background and complementary information do you use?


Excuse me, wasn’t the problem that machine learning methods were black-boxes?
These questions do not all seem to be concerned with black-boxes

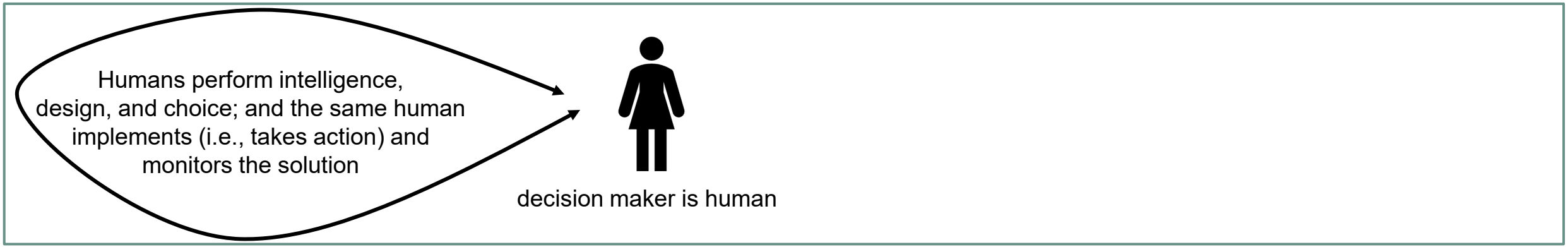


Because users require multiple information contents, then simply using interpretable methods will not suffice to provide users with the information contents they want!

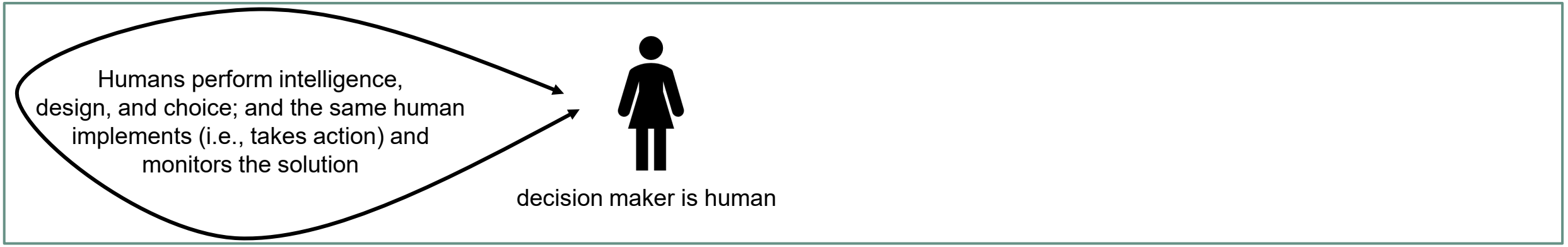
USERS DO NOT TRUST AI AGENTS BECAUSE THEY ARE BLACK-BOXES

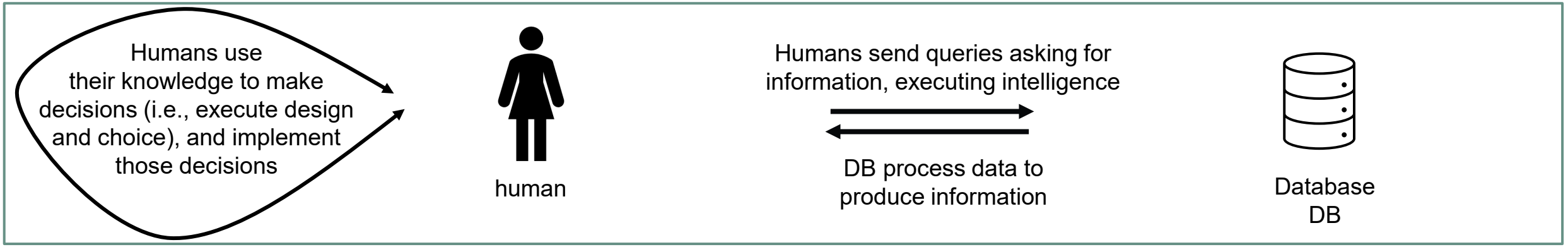


USERS DO NOT TRUST AI AGENTS BECAUSE THEY ARE BLACK-BOXES



USERS DO NOT TRUST AI AGENTS BECAUSE THEY ARE BLACK-BOXES





USERS DO NOT TRUST AI AGENTS BECAUSE THEY ARE BLACK-BOXES

Humans perform intelligence, design, and choice; and the same human implements (i.e., takes action) and monitors the solution

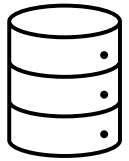
decision maker is human

Humans use their knowledge to make decisions (i.e., execute design and choice), and implement those decisions

human

Humans send queries asking for information, executing intelligence

DB process data to produce information



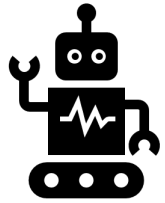
Database DB

Humans implement decisions (i.e., take action)

human

Humans describe problems

AI methods return decisions via solutions to AI tasks



AI method

For non-ai experts, could this change in paradigm be the cause of resistance?

If XAI is a sub-field dedicated to open black-boxes because:
1. humans do not trust AI methods because they are black-boxes, and
2. there is a tradeoff between accuracy and interpretability

then I agree this field should not exist!

We need a sub-field to study how AI methods explain themselves.

DIRECTION VI:
What are the motivations for the field of XAI?

DEPENDENCIES ON MULTIPLE DISCIPLINES

- Multi and interdisciplinarity
- Barriers to multi-disciplinarity
- What to avoid
- What to do

MULTI AND INTERDISCIPLINARITY

Multidisciplinarity

Juxtaposition of disciplines in both education and research without integration and with limited interaction characterizes *multidisciplinarity* (Lattuca 2001, Klein 2010).

Juxtaposition of disciplines does not mean that researchers can execute the research methods outside their own expertise.

MULTI AND INTERDISCIPLINARITY

Multidisciplinarity

Juxtaposition of disciplines in both education and research without integration and with limited interaction characterizes *multidisciplinarity* (Lattuca 2001, Klein 2010).

Interdisciplinarity

Interdisciplinarity is characterized by juxtapositions that entail integration, interaction, linking, focusing and blending (Klein 2010). Choi and Pak (2006) further describes linking as supporting a coherent whole where disciplinary boundaries eroded.

HAS XAI SUCCEEDED IN ERODING THE BOUNDARIES AND CREATING A COHERENT WHOLE THROUGH THE JUXTAPOSITION OF COMPUTER SCIENCE AND SOCIAL SCIENCES AROUND EXPLAINABILITY TO END USERS?

ticularly those using Machine Learning (ML), should be able to "explain" their behavior. Unfortunately, there is little agreement as to what constitutes an "explanation." This has caused a disconnect between the explanations that systems produce in service of explainable Artificial

Gilpin et al. 2022

about what this means and how to achieve it. Authors disagree on *what* should be explained (topic), *to whom* something should be explained (stakeholder), *how* something should be explained (instrument), and *why* something should be explained (goal). In this paper, I em-

Buchholz 2022

There is now a vast and confusing literature on some combination of interpretability and explainability. Much literature on explainability confounds it with interpretability/comprehensibility, thus obscuring the arguments (and thus detracting from their precision), and failing to convey the relative importance and use-cases of the two topics in practice. Some of the literature discusses topics in such generality that its lessons have little bearing on any specific problem. Some of it aims to design taxonomies that miss vast topics within interpretable ML. Some of it provides definitions that we disagree with. Some of it even provides guidance that could perpetuate bad practice. Importantly, most of it

Rudin et al. 2022

in interpretable machine learning. The literature currently being generated on interpretable and explainable AI can be downright confusing. The sheer diversity of individuals weighing in on this field includes not just statisticians and computer scientists but legal experts, philosophers, and graduate students, many of whom have not either built or deployed a machine learning model ever. It is

DEPENDENCIES ON MULTIPLE DISCIPLINES

When computer scientists/mathematicians/statisticians/engineers have their submissions to AAAI/IJCAI rejected on the basis that they do not include validation via user studies,

they are conducting these studies without the help of qualified social scientists.

This unsuccessful lack of boundaries is also causing papers written by social scientists being reviewed by non-social scientists

THE RESULT IS LACK OF RIGOR

Johs et al. 2022 surveyed papers and observed large part lacked details required to assess qualitative research rigor.

Non-experts in qualitative research should not be encumbered with the additional burden of designing, conducting, and analyzing the results of qualitative investigations in XAI.

We underscore the standpoints of Miller, Payrovnaziri et al., Bhatt et al., and Xu and call for the XAI community to collaborate with experts from social disciplines toward bolstering rigor and effectiveness in user studies.

BARRIERS TO INTERDISCIPLINARITY

Lélé and Norgaard 2005 Haythornthwaite et al. 2006 Wagner et al. 2011

Researchers in one discipline do not even know about the research interests, research questions, and theories the researchers in other disciplines rely on.

Researchers in each discipline have their own culture and values that impact decisions at every step.

Researchers in each discipline have their own value judgements that can manifest in different interpretations of reality.

The steps pursued by a given culture and value judgement are interdependent and such interdependencies are not obvious or apparent.

Organizational barriers such as difficulties stemming from disciplines not being organized based on societal problems and overhead imposed by infrastructure and logistics of collaboration.

Perceptions that interdisciplinary work is of lesser value, and the fact that it is harder to reproduce.

HOW TO COUNTERACT INTERDISCIPLINARY BARRIERS

Make every step (research goals, research questions, theories), concept, interpretation, and their interdependency explicit (Bauer 1990) .

Make your discipline, research goals, research questions, and theories explicit and keep it multidisciplinary.

Avoid the risks of interdisciplinarity.

DIRECTION VII:

Make explicit what your discipline is and indicate the AI method, the AI task, the XAI aspect you are investigating. Delimit the scope of each expertise!

TRENDS

Social scientists continue to advance their studies, e.g., personalized XAI (Conati et al. 2021b; Vasileiou and Yeoh 2022)

Authors continue to identify new criteria for evaluation, but no benchmarks (Weber, Amir, and Miller 2022)

There is a new trend to use XAI methods to improve model performance

No papers addressing any of the roadblocks except for one exception for evaluating counterfactuals (Keane et al. 2021)

Image by StockSnap from Pixabay

Photo by Charles Parker

DIRECTION I: Engage the XAI community to describe and make explicit their broad view of the sub-field of XAI.

DIRECTION II: Investigate a precise means to describe and recognize interpretability aspects of a model both at the global and local levels so it can be determined when explanation methods for the model are needed.

DIRECTION III: Investigate approaches to evaluate the competence of XAI methods to produce each type of information content that can have explanatory value including benchmark datasets.

DIRECTION IV: Investigate how to precisely define the explanation context from the perspective of the AI method.

DIRECTION V: Investigate methods to produce **the information contents** users want that are not yet available.

DIRECTION VI: What are the motivations for the field of XAI?

DIRECTION VII: Make explicit what your discipline is and indicate the AI method, the AI task, the XAI aspect you are investigating. Delimit the scope of each expertise!

REFERENCES CITED

Cited on interpretability and interpretable model:

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistics Surveys, 16, 1-85.

Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients. Methods in Ecology and Evolution, 1(2), 103-113.

Lou, Y., Caruana, R. and Gehrke, J., 2012, August. Intelligible models for classification and regression. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 150-158).

Doshi-Velez, Finale, and Been Kim. "Towards A Rigorous Science of Interpretable Machine Learning." stat 1050 (2017): 2.

Drumond, T. F., Viéville, T., & Alexandre, F. (2017, December). Using prototypes to improve convolutional networks interpretability. In NIPS 2017-31st Annual Conference on Neural Information Processing Systems: Transparent and interpretable machine learning in safety critical environments Workshop.

Zhang, Q. S., & Zhu, S. C. (2018). Visual interpretability for deep learning: a survey. Frontiers of Information Technology & Electronic Engineering, 19(1), 27-39.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018, October). Explaining explanations: An overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA) (pp. 80-89). IEEE.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3), 31-57.

Chen C, Li O, Tao D, Barnett A, Rudin C, Su JK. This looks like that: deep learning for interpretable image recognition. Advances in neural information processing systems. 2019;32.

Lalor, J. P., & Guo, H. (2022). Measuring algorithmic interpretability: A human-learning-based framework and the corresponding cognitive complexity score. arXiv preprint arXiv:2205.10207.

REFERENCES CITED

Articles on explaining decision trees:

Izza, Y., Ignatiev, A., & Marques-Silva, J. (2020). On explaining decision trees. arXiv preprint arXiv:2010.11034.

Izza, Y., Ignatiev, A., & Marques-Silva, J. (2022). On Tackling Explanation Redundancy in Decision Trees. arXiv preprint arXiv:2205.09971.

Ambiguity and definition of explanation:

Buchholz O. A Means-End Account of Explainable Artificial Intelligence. arXiv preprint arXiv:2208.04638. 2022 Aug 9.

Gilpin, L. H., et al. " Explanation" is Not a Technical Term: The Problem of Ambiguity in XAI. arXiv preprint arXiv:2207.00007 (2022).

Mueller et al., 2019, p. 86; Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI

Definition of AI:

Russell and Norvig, 2009, p 1. Artificial intelligence: a modern approach. Prentice-Hall.

Decision-making and problem-solving model:

Simon, H. A. 1957. Models of man; social and rational.

Huber, G.P. (1980) Managerial decision making. Scott, Foresman and Company.

REFERENCES CITED

Articles on explaining Evaluations

Yang, F.; Du, M.; and Hu, X. 2019. Evaluating explanation without ground truth in interpretable machine learning. arXiv preprint arXiv:1907.06831.

Tomsett, R.; Harborne, D.; Chakraborty, S.; Gurram, P.; and Preece, A. 2019. Sanity Checks for Saliency Metrics. arXiv:1912.01451.

Hooker, S.; Erhan, D.; Kindermans, P.-J.; and Kim, B. 2019. A benchmark for interpretability methods in deep neural networks. In Advances in Neural Information Processing Systems, 9737–9748.

Montavon, G. 2019. Gradient-Based Vs. Propagation-Based Explanations: An Axiomatic Comparison. In Samek, W.; Montavon, G.; Vedaldi, A.; Hansen, L. K.; and M¨uller, K.- R., eds., Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, 253–265. Cham: Springer International Publishing. ISBN 978-3-030-28954-6.

Hooker, S.; Erhan, D.; Kindermans, P.-J.; and Kim, B. 2019. A benchmark for interpretability methods in deep neural networks. In Advances in Neural Information Processing Systems, 9737–9748.

Barr, B.; Xu, K.; Silva, C.; Bertini, E.; Reilly, R.; Bruss, C. B.; and Wittenbach, J. D. 2020. Towards Ground Truth Explainability on Tabular Data.

Yang, M.; and Kim, B. 2019. Benchmarking Attribution Methods with Relative Feature Importance. arXiv preprint arXiv:1907.09701

Amiri, S. S.; Weber, R. O.; Goel, P.; Brooks, O.; Gandley, A.; Kitchell, B.; and Zehm, A. 2020. Data representing ground-truth explanations to evaluate xai methods. arXiv preprint arXiv:2011.09892.

Zhou, Y.; Booth, S.; Ribeiro, M. T.; and Shah, J. 2022. Do feature attribution methods correctly attribute features? In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 9623–9633.

REFERENCES CITED

Articles on what users expect as explanations:

Brian Y Lim. Improving understanding and trust with intelligibility in context-aware applications. PhD thesis, Carnegie Mellon University, 2012.

I Nunes and D Jannach. A systematic review and taxonomy of explanations in decision support and recommender systems. User Modeling and User-Adapted Interaction, 27(3):393–444, 2017.

S Chari, O Seneviratne, D M Gruen, M A Foreman, A K Das, and DL McGuinness. Explanation ontology: A model of explanations for user-centered ai. In International Semantic Web Conference, pages 228–243. Springer, 2020.

D. Gunning, Explainable artificial intelligence (XAI), DARPA/I2O; www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf.

Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning. (2018). https://doi.org/arXiv:1806.00069v2 arXiv:1806.00069

Gallant SI. Connectionist expert systems. Communications of the ACM. 1988 Feb 1;31(2):152-69.

REFERENCES CITED

Articles on feature attribution methods

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, 1135–1144. New York, NY, USA: Association for Computing Machinery. ISBN 9781450342322.

Shrikumar, A., Greenside, P. and Kundaje, A., 2017, July. Learning important features through propagating activation differences. In International conference on machine learning (pp. 3145-3153). PMLR.

S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," PLoS One, vol. 10, no. 7, p. e0130140, Jul. 2015, doi: 10.1371/journal.pone.0130140.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In Precup, D.; and Teh, Y. W., eds., Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, 3319–3328. PMLR.

Smilkov, D.; Thorat, N.; Kim, B.; Vi´egas, F.; and Wattenberg, M. 2017. SmoothGrad: removing noise by adding noise. arXiv:1706.03825.

REFERENCES CITED

Articles on instance attribution methods

Koh, P. W.; and Liang, P. 2017. Understanding Black-box Predictions via Influence Functions. In Precup, D.; and Teh, Y. W., eds., Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, 1885–1894. PMLR.

Yeh, C.-K.; Kim, J.; Yen, I. E.-H.; and Ravikumar, P. K. 2018. Representer point selection for explaining deep neural networks. Advances in neural information processing systems, 31.

Chen, Y.; Li, B.; Yu, H.; Wu, P.; and Miao, C. 2021. HyDRA: Hypergradient Data Relevance Analysis for Interpreting Deep Neural Networks. Proceedings of the AAAI Conference on Artificial Intelligence, 35(8): 7081–7089.

Articles on example- and prototype-based

Kenny, E. M.; and Keane, M. T. 2021. Explaining Deep Learning using examples: Optimal feature weighting methods for twin systems using post-hoc, explanation-by example in XAI. Knowledge-Based Systems, 233: 107530.

Chen, C.; Li, O.; Tao, D.; Barnett, A.; Rudin, C.; and Su, J. K. 2019. This looks like that: deep learning for interpretable image recognition. Advances in neural information processing systems, 32.

Articles on counterfactuals methods

Keane, M. T., Kenny, E. M., Delaney, E., & Smyth, B. (2021). If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques. arXiv preprint arXiv:2103.01035.

REFERENCES CITED

Trade-off exists:

Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; and Yang, G.-Z. 2019. XAI—Explainable artificial intelligence. Science robotics, 4(37): eaay7120.

Articles stating/showing accuracy/interpretability tradeoff does not hold:

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistics Surveys, 16, 1-85.

Bell, A., Solano-Kamaiko, I., Nov, O., & Stoyanovich, J. (2022). It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy. In 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 248-266).

Dziugaite, G. K., Ben-David, S., & Roy, D. M. (2020). Enforcing Interpretability and its Statistical Impacts: Trade-offs between Accuracy and Interpretability. arXiv preprint arXiv:2010.13764.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, 116(44), 22071-22080.

Mobyen Uddin Ahmed, Shaibal Barua, Shahina Begum, Mir Riyanul Islam and Rosina O Weber. When a CBR in Hand is Better than Twins in the Bush. ICCBR XCBR'22: 4th Workshop on XCBR: Case-based Reasoning for the Explanation of Intelligent Systems at ICCBR-2022, CEUR-WS.org

REFERENCES CITED

Articles on multi and interdisciplinarity:

Lattuca, L. (2001). Creating interdisciplinarity: interdisciplinary research and teaching among college and university faculty. Nashville, TN: Vanderbilt University Press.

Klein, J. T. (2010). A taxonomy of interdisciplinarity. In R Frodeman, J T Klein, and C Mitcham, eds. The Oxford Handbook of Interdisciplinarity. Chapter 2. Oxford University Press, Oxford.

Choi, B. C., & Pak, A. W. (2006). Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness. Clinical and Investigative medicine. Medecine Clinique et Experimentale, 29(6), 351-364.

Lélé, S. and Norgaard, R.B. (2005). Practicing interdisciplinarity. BioScience, 55, 11. 967-975

Haythornthwaite, C., Lunsford, K. J., Bowker, G. C., and Bruce, B. C. (2006). Challenges for research and practice in distributed, interdisciplinary collaboration. New infrastructures for science knowledge production, 143-166.

Bauer, H. H. (1990). Barriers against Interdisciplinarity: Implications for Studies of Science, Technology, and Society (STS). Science, Technology & Human Values, 15(1), 105-119.

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., ... & Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. Journal of Informetrics, 5(1), 14-26.

Bauer, H. H. 1990. Barriers against interdisciplinarity: Implications for studies of science, technology, and society (STS. Science, Technology, & Human Values, 15(1): 105–119.

REFERENCES CITED

Study showing computer scientists conduct qualitative user studies without rigor:

Johs, A. J., Agosto, D., & Weber, R. O. (2022) Explainable artificial intelligence and social science: Further insights for qualitative investigation. Applied AI Letters. 2022; e64. https://onlinelibrary.wiley.com/doi/10.1002/ail2.64

Articles that call for the XAI community to collaborate with experts from social disciplines toward bolstering rigor and effectiveness in user studies:

Miller T. Explanation in artificial intelligence: insights from the social sciences. Artif Intell. 2019; 267: 1-38.doi:10.1016/j.artint.2018.07.007

Payrovnaziri SN, Chen Z, Rengifo-Moreno P, et al. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. J Am Med Inform Assoc. 2020; 27: 1173- 1185. doi:10.1093/jamia/ocaa053

Bhatt U, Andrus M, Weller A, Xiang A. Machine learning explainability for external stakeholders. Proceedings of the IJCAI-PRICAI 2020 Workshop on eXplainable Artificial Intelligence, 2020.

Xu W. Toward human-centered AI: a perspective from human-computer interaction. Interactions. 2019; 26(4): 42-46. doi:10.1145/3328485

REFERENCES CITED

Trends:

Personalized XAI:

Conati, C.; Barral, O.; Putnam, V.; and Rieger, L. 2021. Toward personalized XAI: A case study in intelligent tutoring systems. Artificial Intelligence, 298: 103503.

Vasileiou, S. L.; and Yeoh, W. 2022. On Generating Abstract Explanations via Knowledge Forgetting. In ICAPS 2022 Workshop on Explainable AI Planning.

XAI for Performance:

Weber L, Lapuschkin S, Binder A, Samek W. Beyond Explaining: Opportunities and Challenges of XAI-Based Model Improvement. arXiv preprint arXiv:2203.08008. 2022 Mar 15.

Evaluation:

See Rizzo et al. and Li et al. In:

Weber, R. O.; Amir, O.; and Miller, T. 2022. IJCAI Workshop on Explainable Artificial Intelligence (XAI). https: //sites.google.com/view/xai2022. Accessed: 2022-09-01.

Exception:

Keane, M. T., Kenny, E. M., Delaney, E., & Smyth, B. (2021). If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques. arXiv preprint arXiv:2103.01035.

Adam J Johs and Prateek Goel, Information Science, College of Computing & Informatics,

Sponsors: NIH/NCATS, DARPA, VINNOVA

VINNOVA and Mälardalens University, Sweden

Colleagues, RAs, Students, and Coauthors

Now I want to hear what you think!

GUIDANCE FROM SOCIAL SCIENTISTS

"Material that is offered as an explanation, no matter its medium, format, or reference, is only an explanation if it results in good effect, that is, it has explanatory value for particular individuals.

Technically, the property of "being an explanation" is not a property of text, statements, narratives, diagrams, or other forms of material.

It is an interaction of:

(1) the offered explanation,

(2) the learner's knowledge and beliefs,

(3) the context or situation and its immediate demands, and

(4) the learner's goals or purposes in that context. This explains why it is possible that purely descriptive statements, not primarily intended to serve as explanations, can nevertheless have explanatory value" Mueller et al., 2019, p. 86.