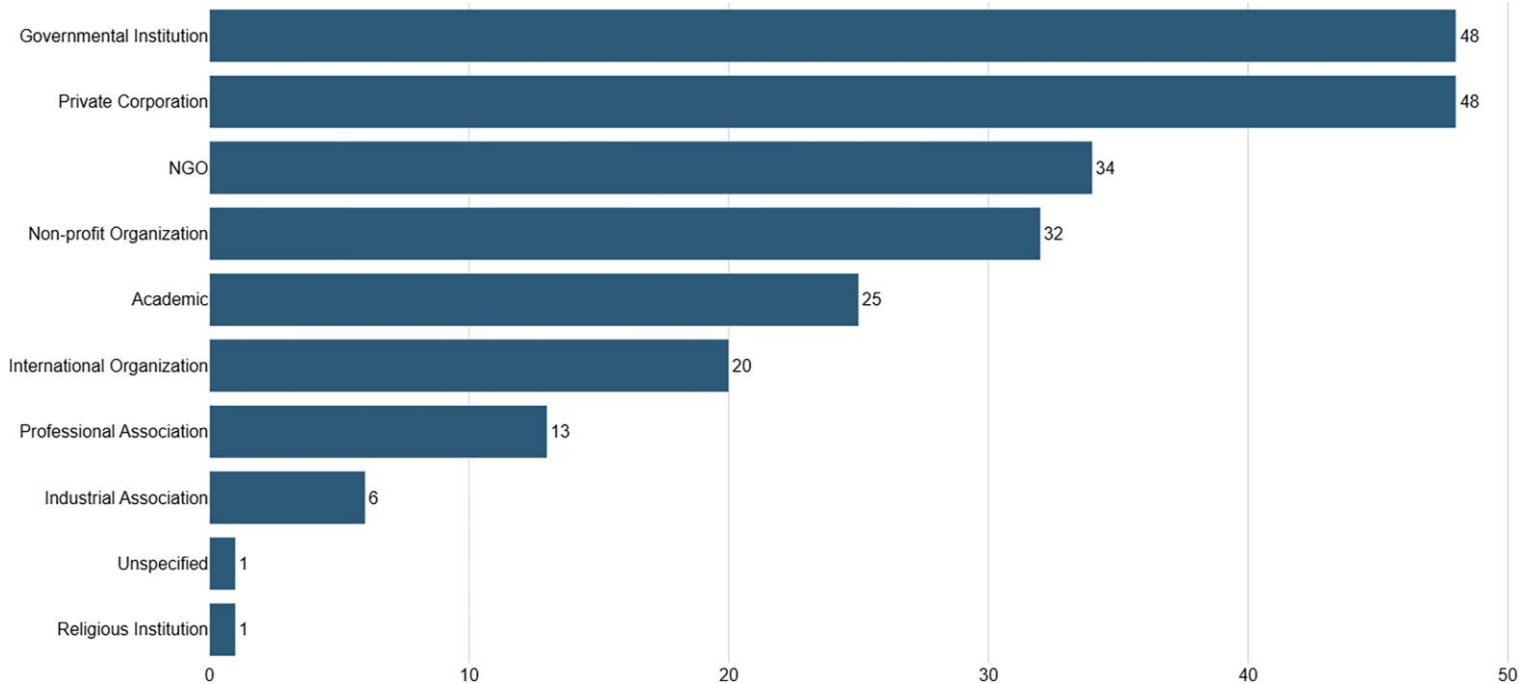




A Good Explanation?

Philosophical Foundations for
Explainable Artificial Intelligence

Rune Nyrup
Centre for Science Studies
Aarhus University



Article

Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance

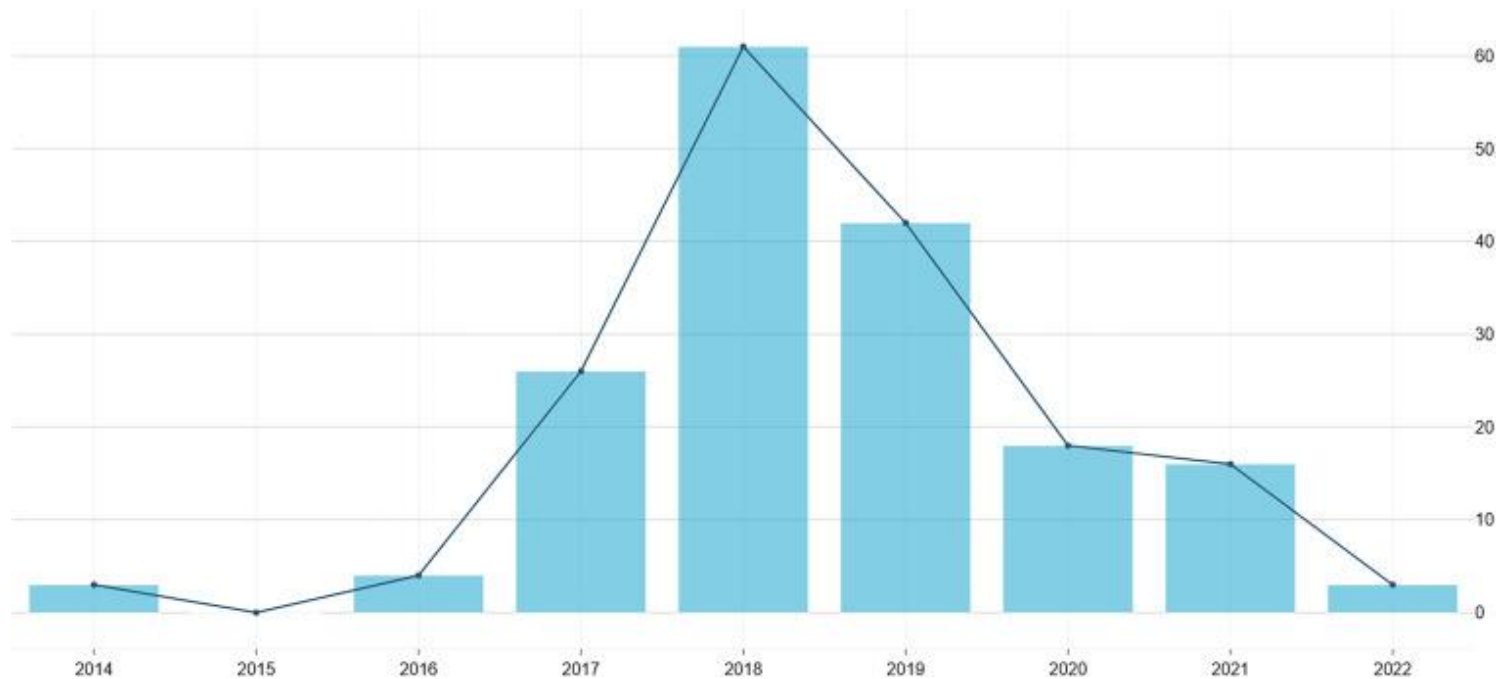
Nicholas Kluge Corrêa^{1,2,7}, Camila Galvão³, James William Santos¹  , Carolina Del Pino⁴, Edson Pontes Pinto³, Camila Barbosa¹, Diogo Massmann¹, Rodrigo Mambrini³, Luiza Galvão⁵, Edmund Terem⁶, Nythamar de Oliveira¹

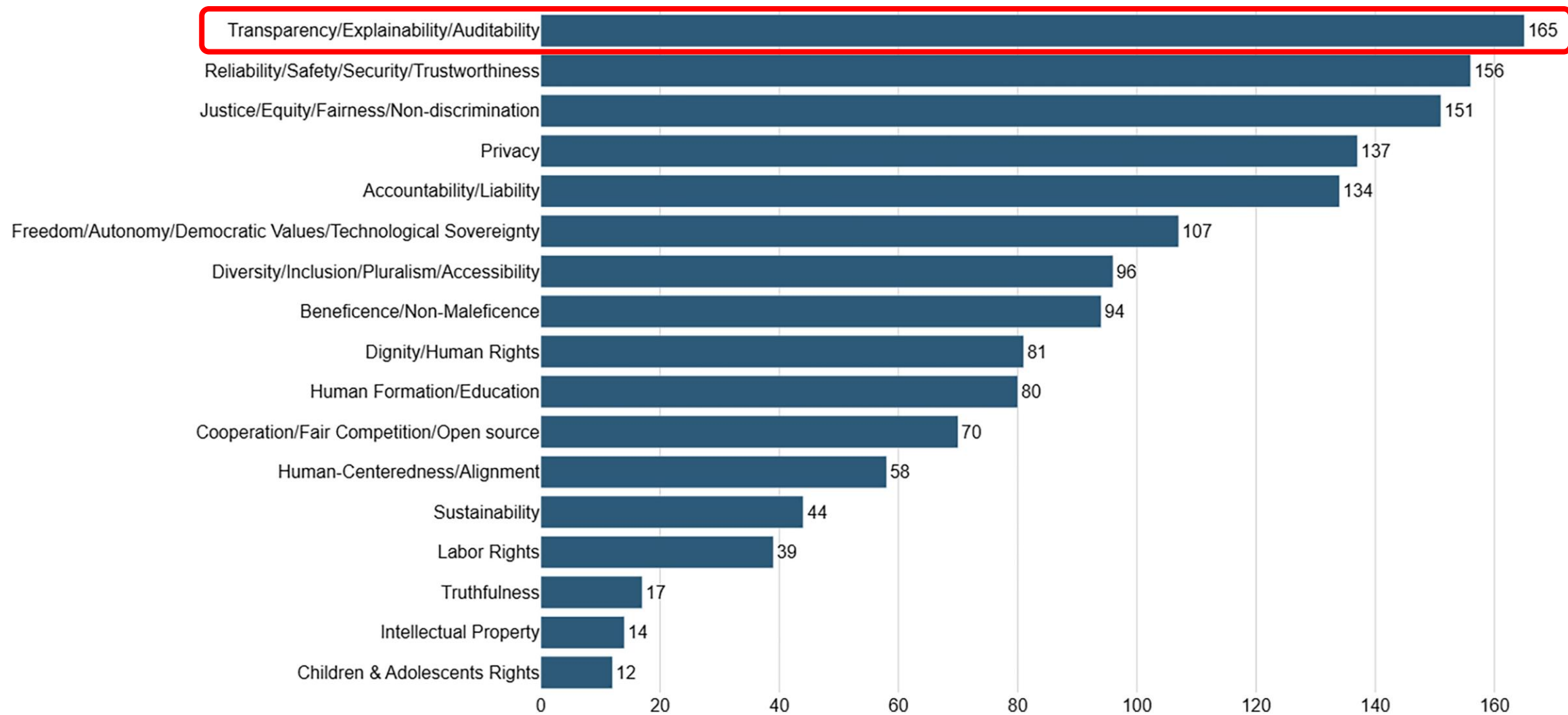
Show more 

 Add to Mendeley  Share  Cite

<https://doi.org/10.1016/j.patter.2023.100857>

[Get rights and content](#)





ETHICS AND GOVERNANCE
OF ARTIFICIAL INTELLIGENCE
FOR HEALTH

WHO GUIDANCE



5. Key ethical principles for use of artificial intelligence for health.....

- 5.1 Protect autonomy.....
- 5.2 Promote human well-being, human safety and the public interest.....
- 5.3 Ensure transparency, explainability and intelligibility.....
- 5.4 Foster responsibility and accountability.....
- 5.5 Ensure inclusiveness and equity.....
- 5.6 Promote artificial intelligence that is responsive and sustainable.....

“AI technologies should be explainable to the extent possible and according to the capacity of those to whom the explanation is directed.”

Proceedings > KDD '16 > "Why Should I Trust You?": Explaining the Predictions of Any Classifier

RESEARCH-ARTICLE



"Why Should I Trust You?": Explaining the Predictions of Any Classifier

Authors: Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin [Authors Info & Claims](#)

KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining • August 2016 • Pages 1135–1144 • <https://doi.org/10.1145/2939672.2939778>

Harvard Journal of Law & Technology

Volume 31, Number 2 Spring 2018

COUNTERFACTUAL EXPLANATIONS WITHOUT OPENING THE BLACK BOX: AUTOMATED DECISIONS AND THE GDPR

Sandra Wachter, Brent Mittelstadt,** & Chris Russell****

NeurIPS Proceedings

Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles

Part of [Advances in Neural Information Processing Systems 30 \(NIPS 2017\)](#)

Bibtex

Metadata

Paper

Reviews

Supplemental

Authors

Balaji Lakshminarayanan, Alexander Pritzel, Charles Blundell

NeurIPS Proceedings

Neural Additive Models: Interpretable Machine Learning with Neural Nets

Part of [Advances in Neural Information Processing Systems 34 \(NeurIPS 2021\)](#)

Bibtex

Paper

Reviews And Public Comment »

Supplemental

Authors

Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, Geoffrey E. Hinton

Local Interpretable Model-agnostic Explanations (LIME)

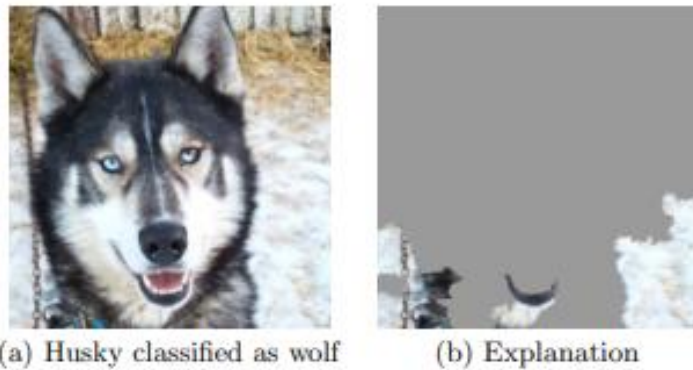
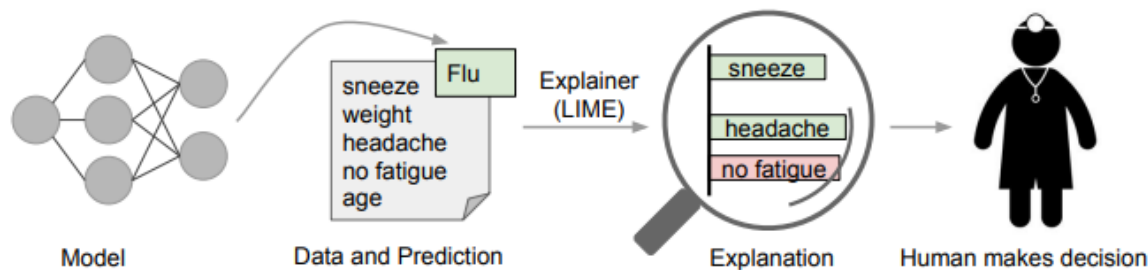
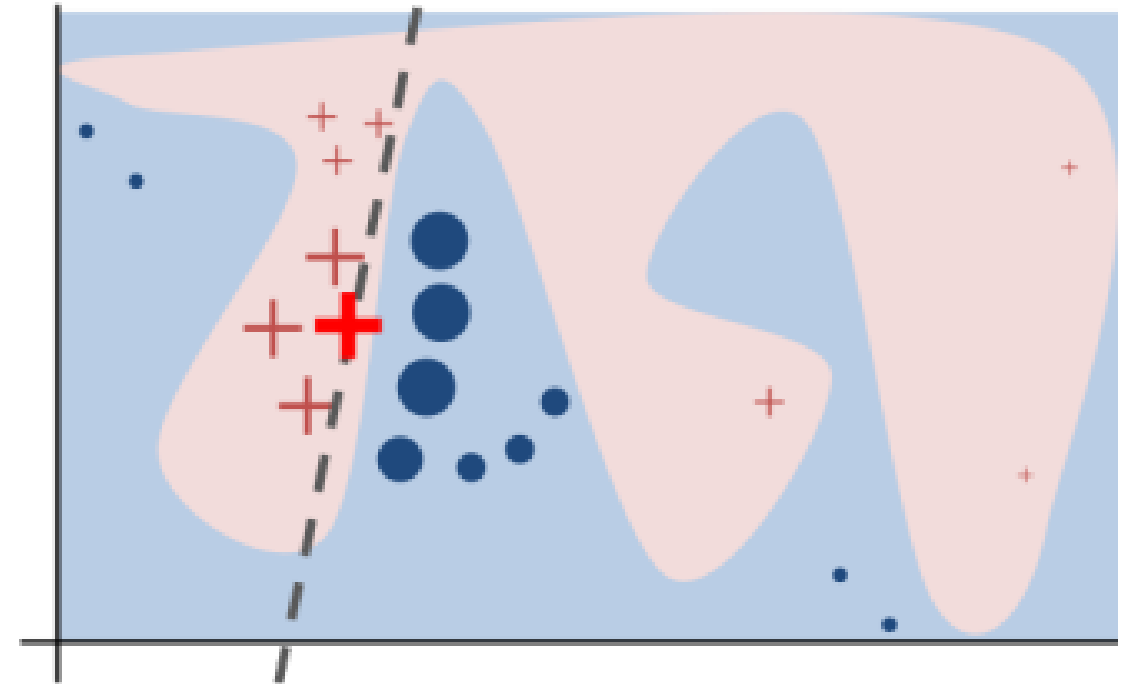
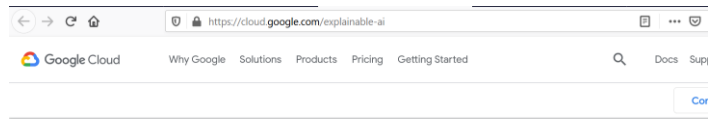


Figure 11: Raw data and explanation of a bad model's prediction in the “Husky vs Wolf” task.





EXPLAINABLE AI

Overview

Benefits

Features

Customers

Resources

Pricing

Explainable AI^{BETA}

Tools and frameworks to understand and interpret your machine learning models.

[Try it free](#)

[View documentation](#)

Understand AI output
and build trust



Model Explainability

Shapley Values

SHAP Baselines for
Explainability

Partial dependence
plots

Computer Vision

Create Feature Attribute
Baselines and
Explainability Reports

Explainability Reports
for NLP Models

Amazon SageMaker Clarify Model Explainability

[PDF](#) | [RSS](#)

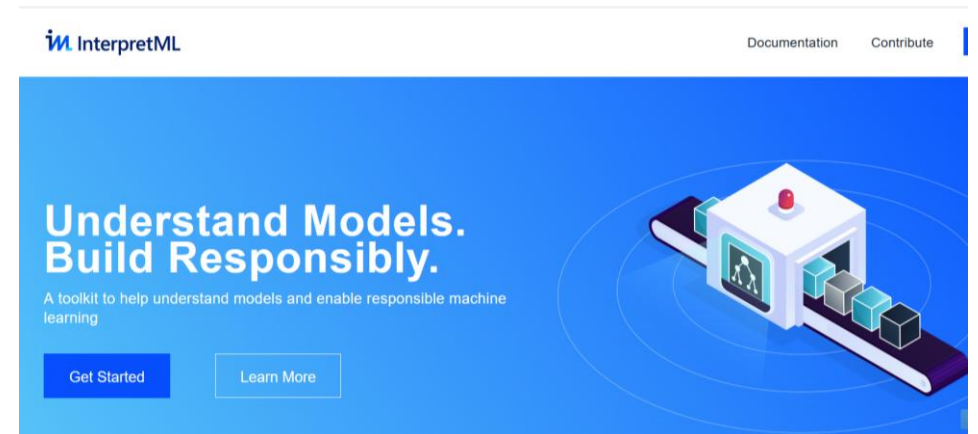
Amazon SageMaker Clarify provides tools to help explain how machine learning (ML) models make predictions. These tools can help ML modelers and developers and other internal stakeholders understand model characteristics as a whole prior to deployment and to debug predictions provided by the model after it's deployed. Transparency



[Facebook app](#)

Why Am I Seeing This? We Have an Answer for You

March 31, 2019



PERSPECTIVE

<https://doi.org/10.1038/s42256-019-0048-x>

nature
machine intelligence

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin 

Explainable AI is Dead, Long Live Explainable AI!: Hypothesis-driven Decision Support using Evaluative AI

Author:  [Tim Miller](#) | [Authors Info & Claims](#)

FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency • Pages 333 - 342
<https://doi.org/10.1145/3593013.3594001>

The Mythos of Model Interpretability

**IN MACHINE LEARNING, THE
CONCEPT OF INTERPRETABILITY IS
BOTH IMPORTANT AND SLIPPERY.**

ZACHARY C. LIPTON

Communications of the ACM, Volume 61, Issue 10

<https://doi.org/10.1145/3233231>

Foundations

~~Frameworks~~ for Explainable AI

- What does it mean for AI to be ‘explainable’?
- Why is explainability ethically important?
- Is explainability always an ethical good?

Three ideas from philosophy of science

- Explanatory pluralism
- Explanations are idealized
- Communicative view of explanation
- Inferentialist view of understanding



Physiological explanation

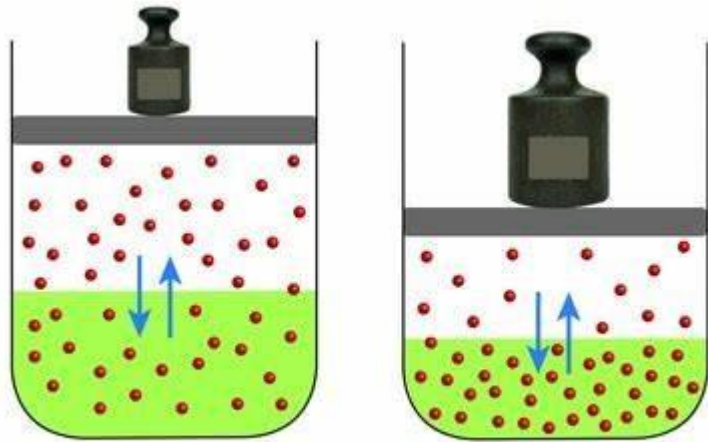
Ecological explanation

Phylogenetic evolutionary
explanation

Molecular-genetic
explanation

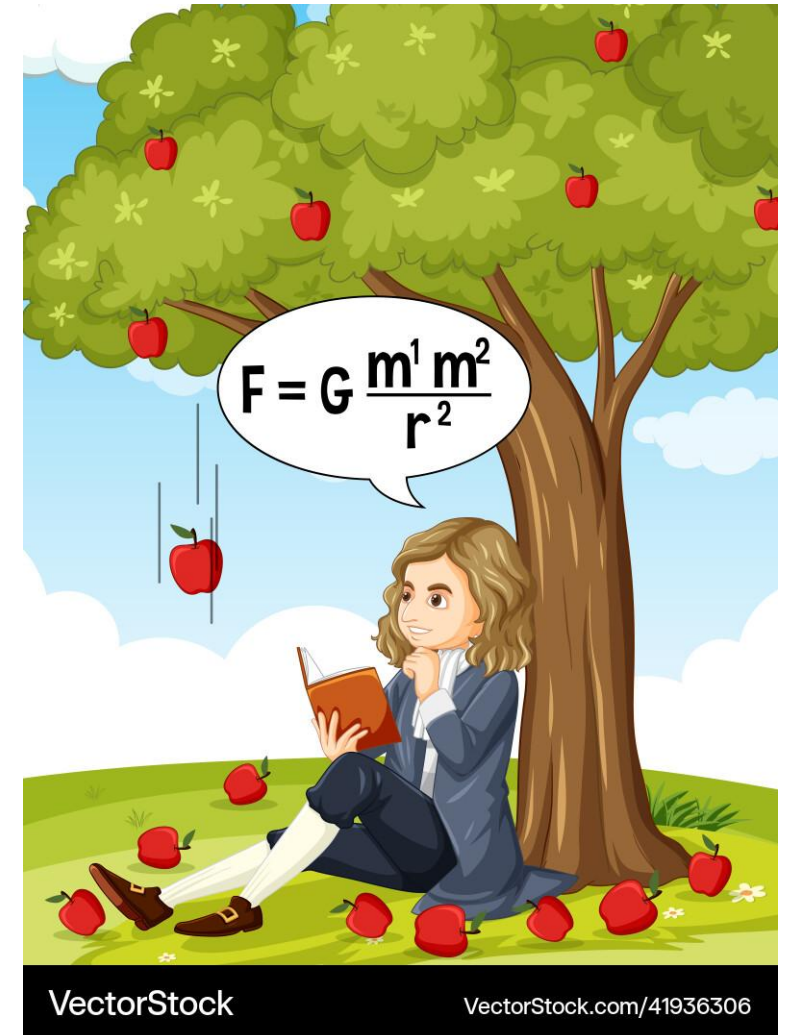
Explanations are idealized

= selective and simplified representations



shutterstock.com · 350946824

$$PV = nRT$$

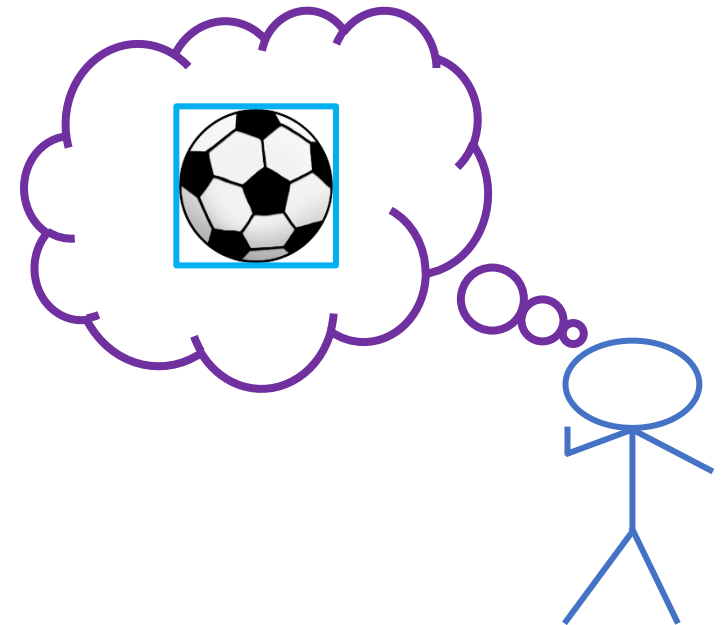
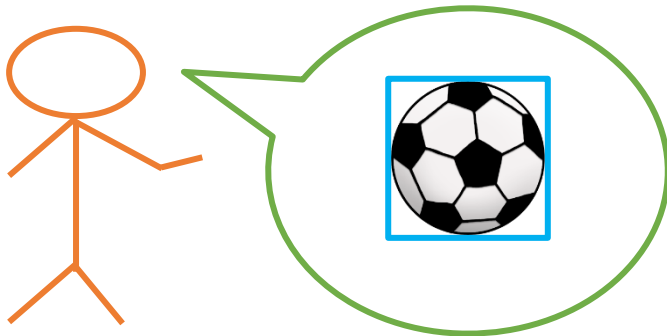


VectorStock

VectorStock.com/41936306

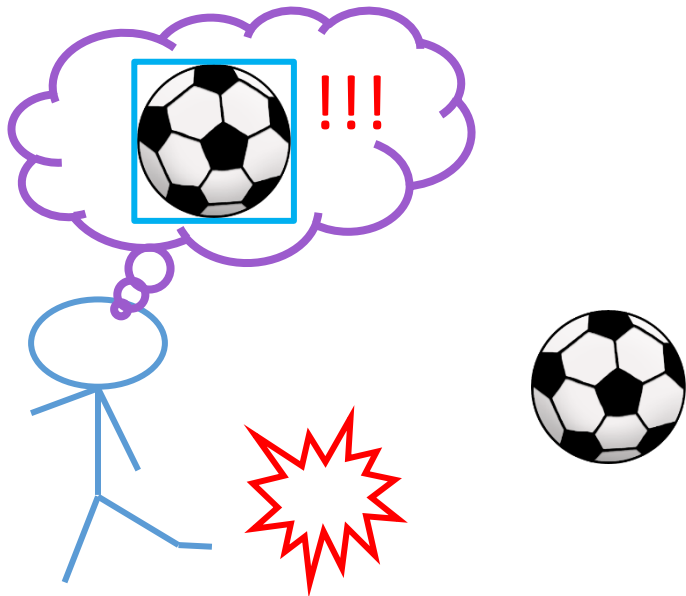
Communicative View of Explanation

S conveys some **information** about an **explanatory target x**
to an **audience A** , in order for A to obtain
a better **understanding** of x



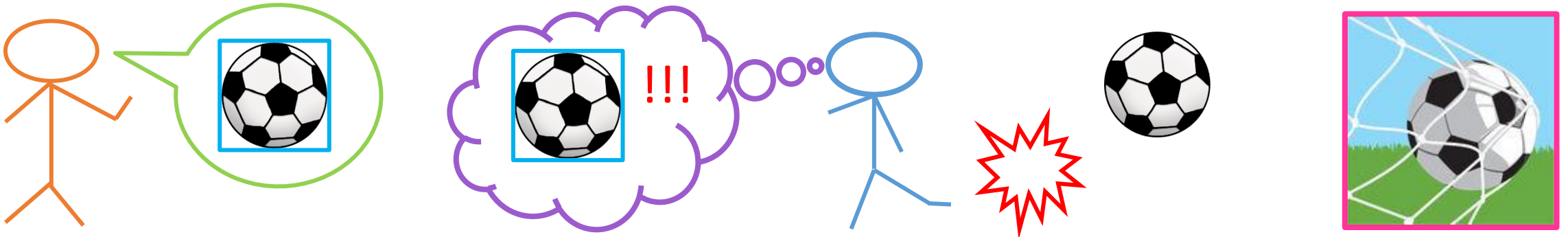
Inferentialist View of Understanding

A understands x in a given context to the extent that
 A is able to competently draw inferences about x that
are needed to achieve contextually salient purposes



Explanatory Pragmatism

x is explainable by S to A in a **given context** to the extent that
 S is able to convey **information** about x to A which
enables A to competently **draw inferences** about x that
are needed to achieve **contextually salient purposes**



Implications

What does it mean for AI to be 'explainable'?

- Always relative to an explainer, audience and contextual purpose

Why is explainability ethically important?

- No general reason
- Relies on assumptions about contextually relevant purposes

Practically: Heuristics for XAI evaluation

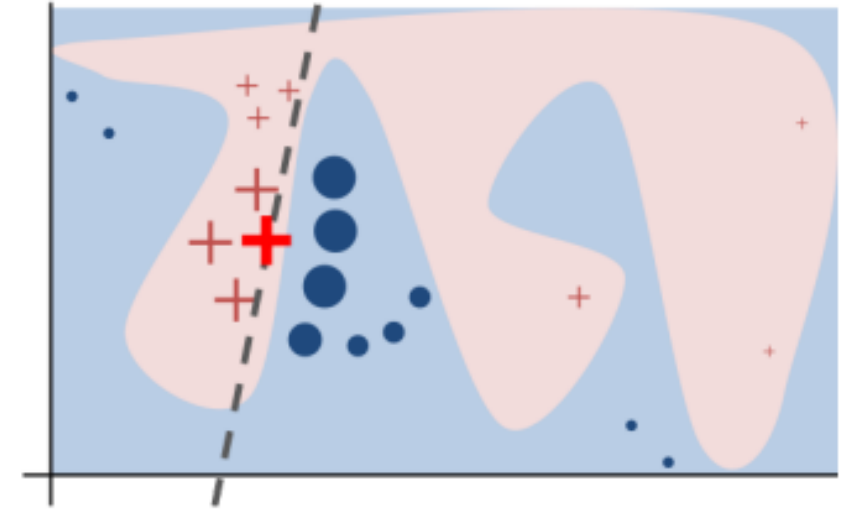
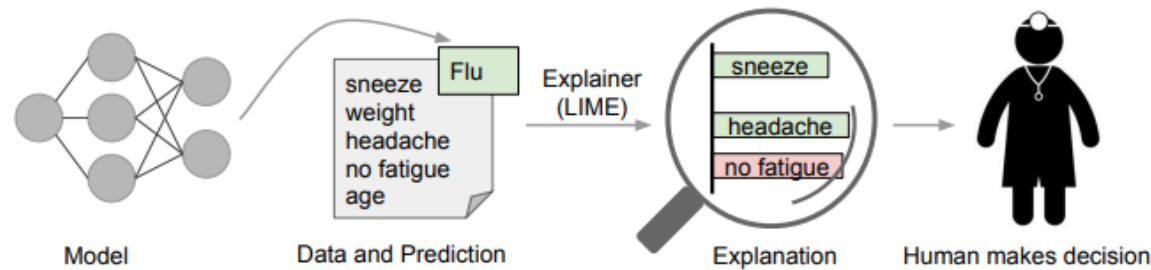
Heuristic 1: Purpose-to-explanation

x is explainable by S to A in a **given context** to the extent that
 S is able to convey **information** about x to A which
enables A to competently **draw inferences** about x that
are needed to achieve **contextually salient purposes**

For a given AI system

- Who needs to be able to do what and for what purpose?
- What information do they need to do this competently?
- How is this information best conveyed to this audience?

LIME: A good explanation?

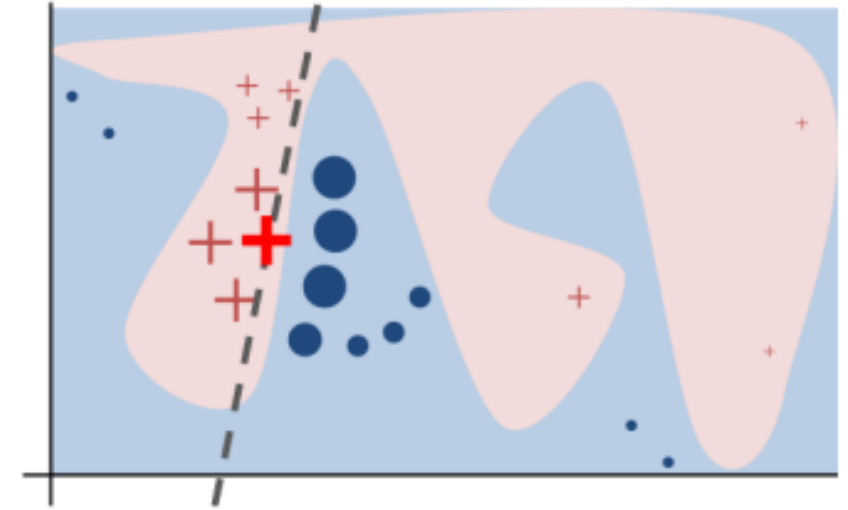
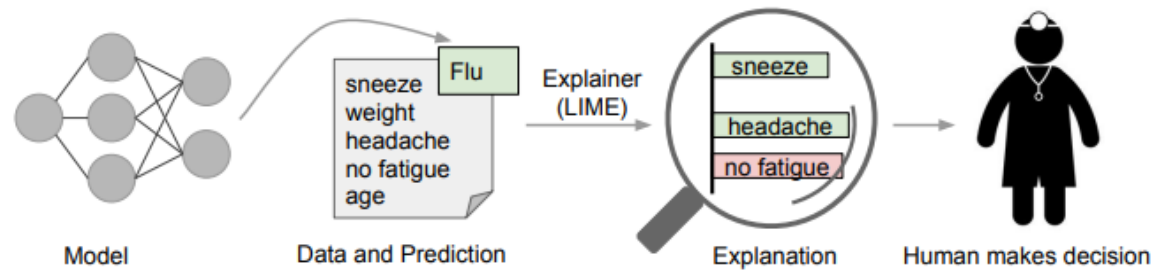


Audience 1: Human clinicians

Purpose 1: Detect faulty model-decisions

Information needed: Which factors made the greatest difference to this decision?

LIME: A good explanation?

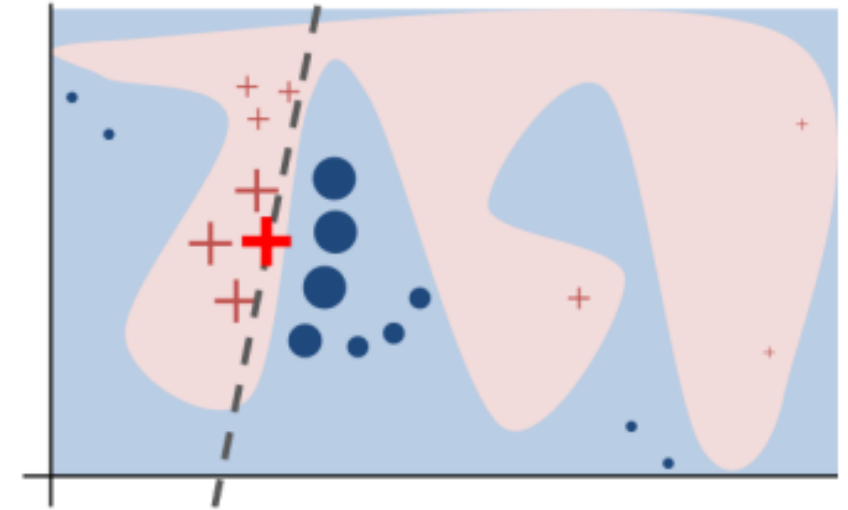
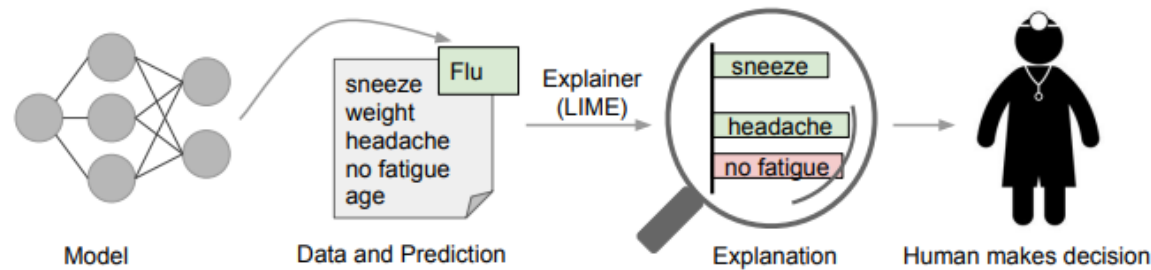


Audience 2: Hospital managers

Purpose 2: Deployment of model at a different hospital

Information needed: What are correlations and causal structures underwrite the reliability of the model?

LIME: A good explanation?



Audience 3: Patients

Purpose 3: Informed consent to proposed medical treatment

Information needed: What values are implicitly prioritised by the model?

Heuristic 2: Explanation-to-purpose

x is explainable by S to A in a **given context** to the extent that
 S is able to convey **information** about x to A which
enables A to competently **draw inferences** about x that
are needed to achieve **contextually salient purposes**

For a given XAI technique

- What information does the technique provide?
- What can different audiences competently infer from this information?
- What purposes can those inferences serve?

Is explainability always an
ethical good?



Interested in encouraging compliance and ensuring a fair and efficient use of health resources across all patients

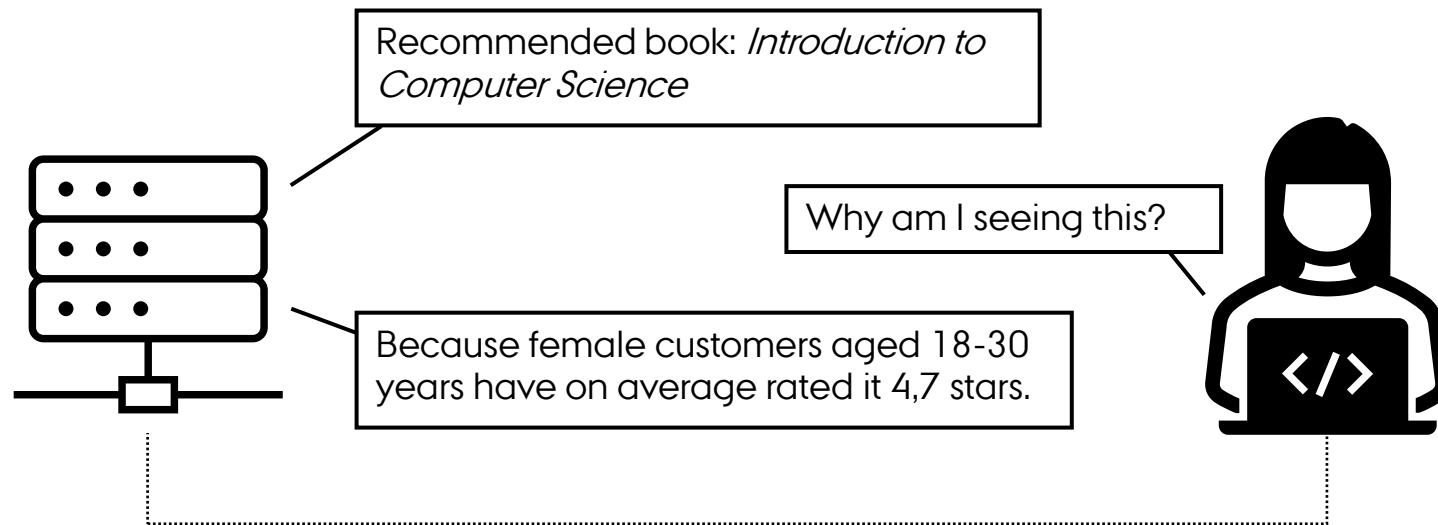
You're being moved from weekly check ups to a yearly review

Why?

Because your situation is sufficiently stable



Interested in receiving the best possible health care



Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations

Athanasios Andreou[§], Giridhari Venkatadri[†], Oana Goga^{*}, Krishna P. Gummadi[‡], Patrick Loiseau^{*‡}, Alan Mislove[†]

“Our results show that ad explanations are often incomplete and sometimes misleading while data explanations are often incomplete and vague”

COGNITIVE SCIENCE

A Multidisciplinary Journal

[Free Access](#)

The misunderstood limits of folk science: an illusion of explanatory depth

Leonid Rozenblit [✉](#) Frank Keil [✉](#)

First published: 11 February 2010 | https://doi.org/10.1207/s15516709cog2605_1 | Citations: 282

'ejournals@cambridge - find full text'

[PDF](#) [TOOLS](#) [SHARE](#)

Abstract

People feel they understand complex phenomena with far greater precision, coherence, and depth than they really do; they are subject to an illusion—an illusion of explanatory depth. The illusion is far stronger for explanatory knowledge than many other kinds of

Journal of Cognitive Neuroscience

Issues [Online Early](#)

Volume 20, Issue 3

March 2008

March 01 2008

The Seductive Allure of Neuroscience Explanations

Deena Skolnick Weisberg, Frank C. Keil, Joshua Goodstein, Elizabeth Rawson, Jeremy R. Gray

[Author and Article Information](#)

Journal of Cognitive Neuroscience (2008) 20 (3): 470–477.

<https://doi.org/10.1162/jocn.2008.20040>



[Cite](#) [PDF](#) [Permissions](#) [Share](#) [v](#)

Abstract

Explanations of psychological phenomena seem to generate more public interest in neuroscientific information. Even irrelevant neuroscience information in an explanation may interfere with people's abilities to critically consider the explanation. We tested this hypothesis by giving naïve adults, students in a

[Previous Article](#) [Next Article](#)

Article Contents

[Abstract](#)

Illusory understanding in XAI

Fig. 2: Saliency does not explain anything except where the network is looking.

From: [Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead](#)



We have no idea why this image is labelled as either a dog or a musical instrument when considering only saliency. The explanations look essentially the same for both classes. Credit: Chaofen Chen, Duke University

Towards Responsible Explainable AI Technologies



In what ways do and don't existing XAI techniques adequately represent their explanatory targets?



Where is the line between helpful simplification, deceit and manipulation?



How do we find ethically legitimate compromises between the interests of explainer(s) and audience(s)?

Towards Responsible Explainable AI Technologies (TREAT)

TREAT

> People



Torben Esbo Agergaard

PhD Student

✉ ta@css.au.dk

🏠 1532, 313



Rune Nyrup

Associate Professor

✉ rune.nyrup@css.au.dk

🏠 1530, 233

☎ +4587150076



Molly Harriet Powell

Postdoc

✉ molly@css.au.dk

🏠 1532, 313



THANKS!