On the challenges of bringing explainable AI to practice

A practitioner's perspective

Hinda Haned, Ph.D.
Owls & Arrows | University of Amsterdam
Transparent ML Symposium, Utrecht October 2022

The Civic AI Lab

UvA

VU

ICAI

Gemeente Amsterdam

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties

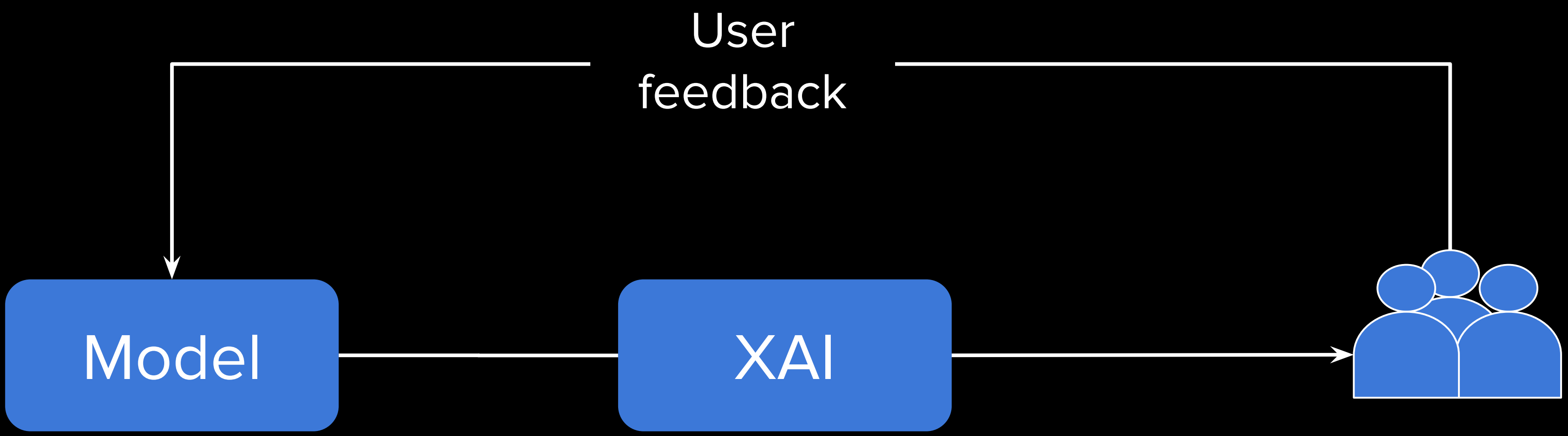
The Civic AI Lab

- Civic-centered and community-minded design, development and deployment of AI technology
- We co-create with Academia, Government, Industry and Civil society
- We also serve as an information point for residents and businesses who have questions about new AI technologies and the ethical and inclusive use of them

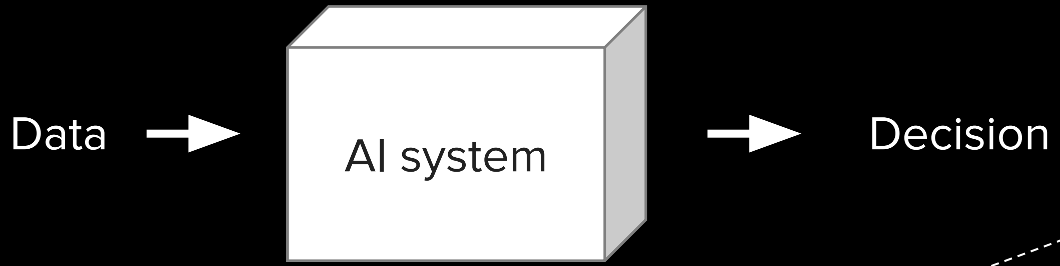
XAI: eXplainable AI

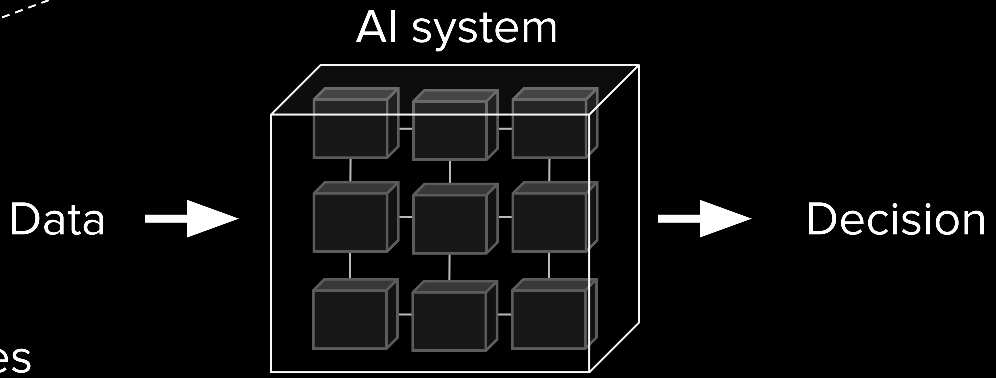
XAI: eXplainable AI

Model outputs must be understandable and transparent to the decision makers and the subjects impacted by them



‘Black-box’ metaphor





The black box nature of AI systems comes from the interaction of many simple components

Complex systems raise concerns

FROM POLITICOPRO

BY MELISSA HEIKKILÄ

MARCH 29, 2022 | 6:14 PM

Dutch scandal serves as a warning for Europe over risks of using algorithms

The Dutch tax authority ruined thousands of lives after using an algorithm to spot suspected benefits fraud – and critics say there is little stopping it from happening again.

Complex systems raise concerns

- Why this ad?
- Why this discount?
- Why this recommendation?
- Why was I rejected?
- Can I change the outcome?
- When will the system fail?

XAI motivators

Model verification

Compliance

User trust

Accountability

Responsible AI

EUROPEAN COMMISSION

Brussels, 21.4.2021

COM(2021) 206 final

2021/0106(COD)

Proposal for a

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS

{SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}

Ethics Guidelines

The EU AI ACT

Compliance - GDPR

ETid	Country	Date of Decision	Fine [€]	Controller/Processor	Quoted Art.	Type	Source
Filter Column	Filter Column		Filter Column	Filter Column		Filter Column	
ETid-1421	SPAIN	2022-10-04	6,000	Club Náutico el Estacio	Art. 5 (1) f) GDPR, Art. 32 GDPR	Insufficient technical and organisational measures to ensure information security	link
ETid-1420	DENMARK	2021-08-17	20,100	Danish Immigration Agency	Art. 5 (1) f) GDPR, Art. 32 GDPR	Insufficient technical and organisational measures to ensure information security	link
ETid-1419	AUSTRIA	2021	600	Private individual	Art. 5 (1) a) GDPR, Art. 9 (1), (2) GDPR	Non-compliance with general data processing principles	link
ETid-1418	AUSTRIA	2021	Unknown	Private individual	Art. 5 (1) a), c) GDPR	Non-compliance with general data processing principles	link
ETid-1417	SPAIN	2022-09-28	31,200	BAYARD REVISTAS, S.A.	Art. 5 (1) f) GDPR, Art. 32 GDPR, Art. 33 GDPR	Insufficient technical and organisational measures to ensure information security	link
ETid-1416	ITALY	2022-07-21	3,000	Azienda Socio Sanitaria Territoriale Rhodense	Art. 5 (1) f) GDPR, Art. 32 GDPR	Insufficient technical and organisational measures to ensure information security	link

https://www.enforcementtracker.com/; 04/10/22

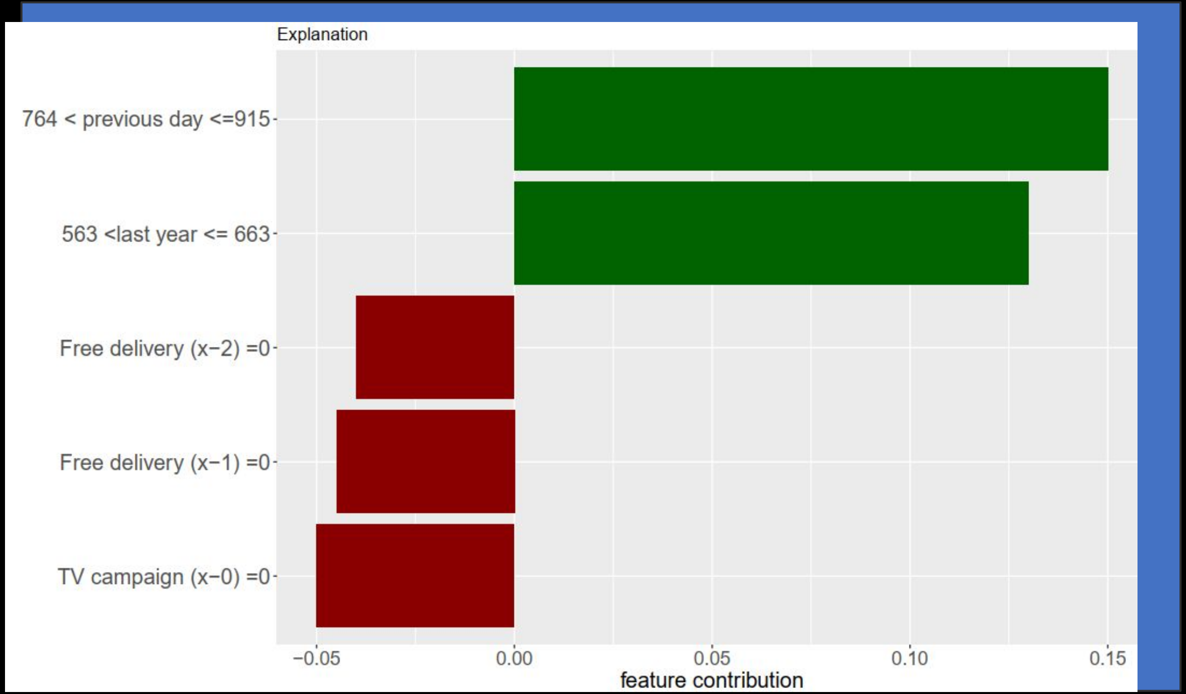
XAI approaches

Explanations types

Global explanations
Explain a model's decision-making process in general. Typically: feature importance.
Treeinterpreter, PDP, feature importance

Local explanations
Explain a single prediction. Since it remains challenging to establish fidelity to black box models in globally interpretable approximations, much attention is put on local explanations.
LIME, SHAP, Skater

Feature attribution



Counterfactual explanations

A counterfactual describes the smallest required change to a feature value that changes the prediction to a predefined desired output

- **Model**: forecast for next week is 5,000 orders
- **Question**:Which feature values must be changed to decrease the forecast to 4,000?
- **Counterfactual**: If your delivery on the weekend is no longer free, you will decrease the forecast to below 4,000 transactions

XAI in practice: challenges

Algorithmic aversion

"We show that people are especially averse to algorithmic forecasters after seeing them perform, even when they see them outperform a human forecaster. This is because people more quickly lose confidence in algorithmic than human forecasters after seeing them make the same mistake"

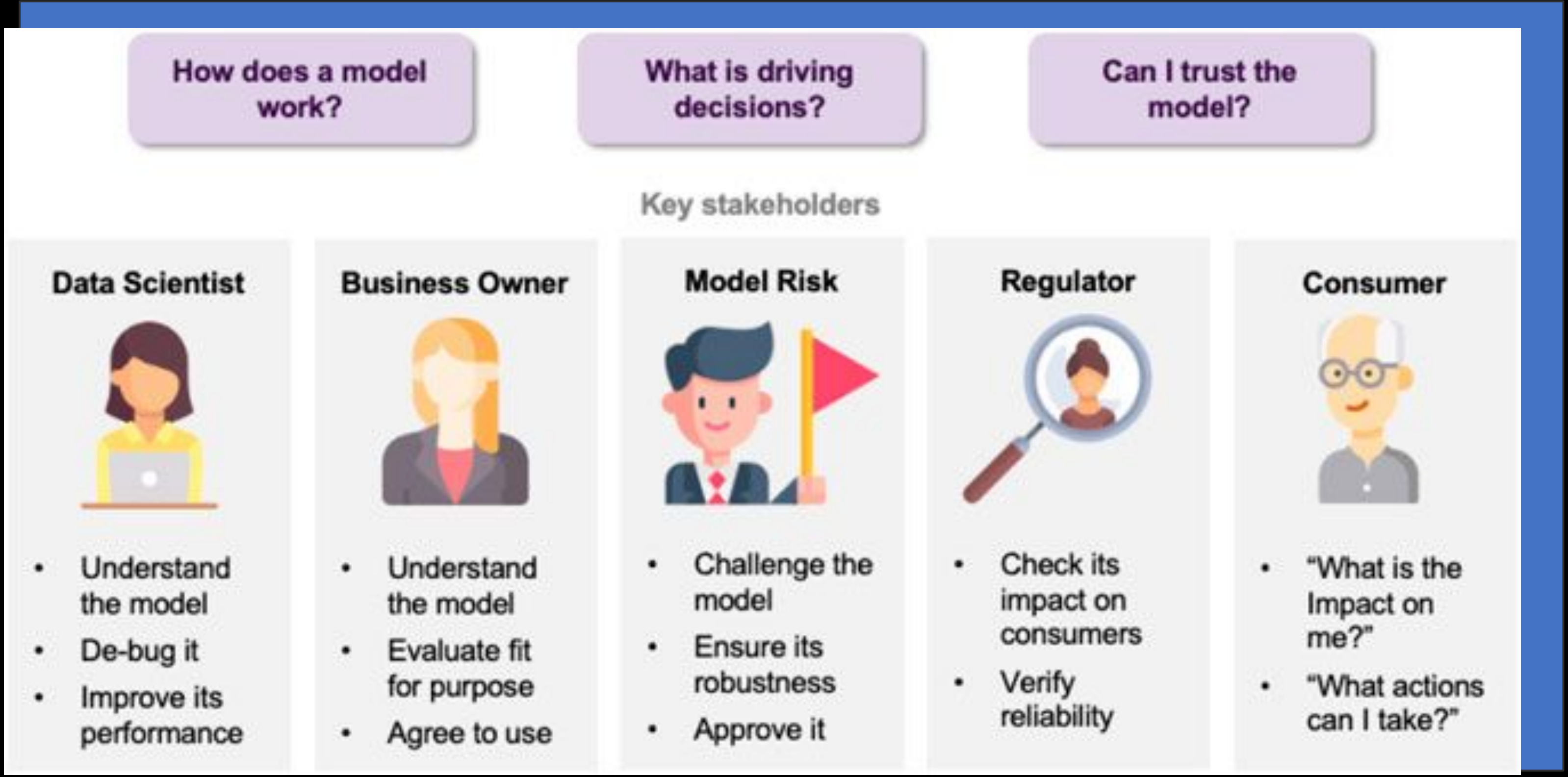
Dietvorst et al. Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology, 2015.

XAI tools: deployed on already-built models

Library Name	Type of Explanation	Regression	Text	Images	Distributed	Licence
AI Explainability 360 (AIX360)	Local and Global	No	No	Yes	No	Apache 2.0
Alibi	Global explanation	Yes	No	No	No	Apache 2.0
Captum	Local and Global	Yes	Yes	Yes	Yes	BSD 3-Clause
Dalex	Local and Global	Yes	No	No	No	GPL v3.0
Eli5	Local and Global	Yes	Yes	Yes	No	MIT License
explainX	Local and Global	Yes	No	No	No	MIT License
LIME	Local and Global	No	Yes	Yes	-	BSD 2-Clause "Simplified" License
InterpretML	Local and Global	Yes	No	No	-	MIT License
SHAP	Local and Global	Yes	Yes	Yes	-	MIT License
TensorWatch	Local explanation	Yes	Yes	Yes	-	MIT License
tf-explain	Local explanation	Yes	Yes	Yes	-	MIT License

Gashi, M.; et al. .State-of-the-Art Explainability Methods with Focus on VisualAnalytics Showcased by Glioma Classification. Biomedinformatics 2022,2, 139–158.

Different stakeholders require different explanations



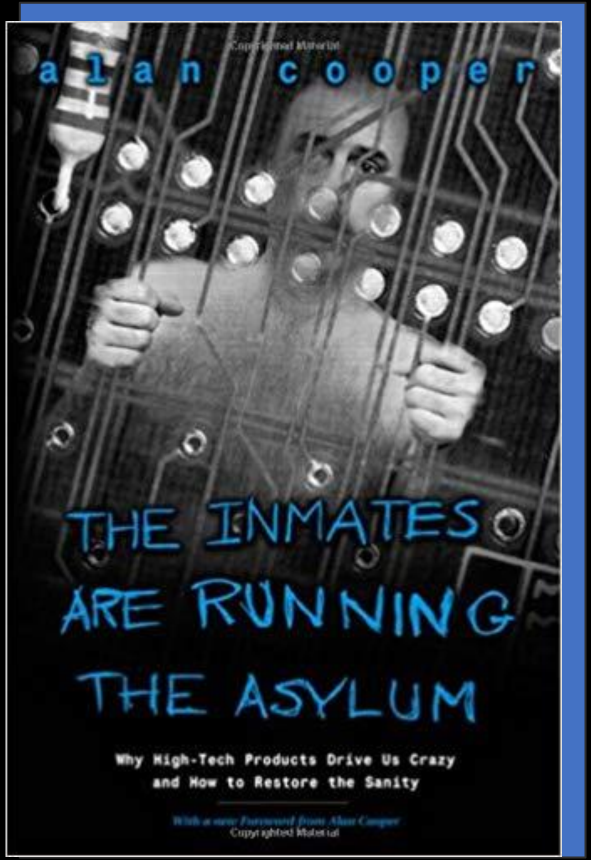
From: Belle V, Papantonis I. Principles and Practice of Explainable Machine Learning. Front Big Data. 2021

Disconnect between algorithmic research and in-deployment contexts

- Counter-intuitive or difficult to understand explanations
- Deployment/usability constraints (computing-time, UI)

Disconnect between algorithmic research and in-deployment contexts

"Most of us as AI researchers are building explanatory agents for ourselves, rather than for the intended users"

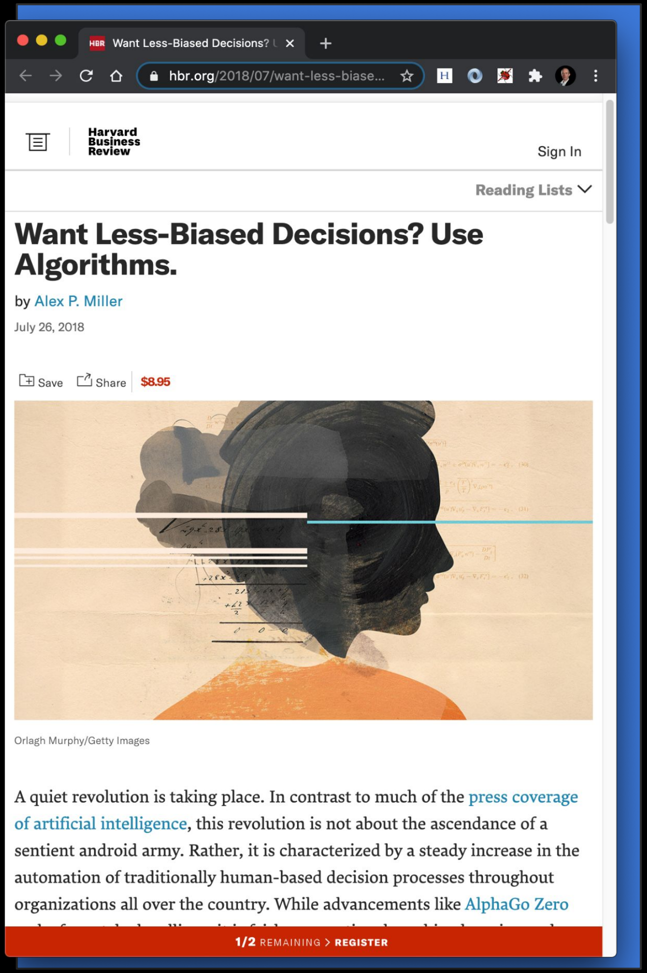


T. Miller et al. Beware of inmates running the Asylum, IJCAI Workshop on explainable AI, 2017.

Common setting

- Automate tedious or repetitive task
- System acquired or co-designed AI system
- Challenged by end-user adoption and acceptance

How can we make this system explainable in deployment?



Want Less-Biased Decisions? U ×

hbr.org/2018/07/want-less-biase...

Harvard Business Review

Sign In

Reading Lists

Want Less-Biased Decisions? Use Algorithms.

by Alex P. Miller

July 26, 2018

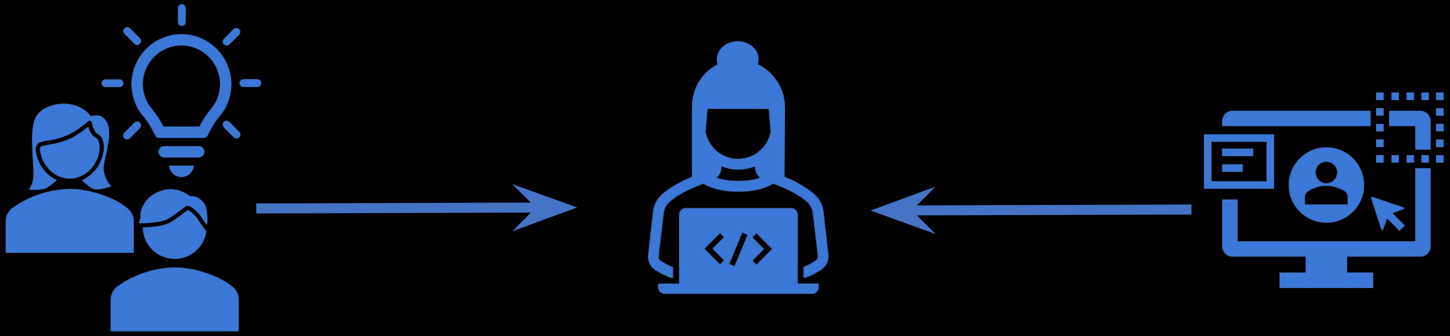
Save Share $8.95

Orlagh Murphy/Getty Images

A quiet revolution is taking place. In contrast to much of the press coverage of artificial intelligence, this revolution is not about the ascendance of a sentient android army. Rather, it is characterized by a steady increase in the automation of traditionally human-based decision processes throughout organizations all over the country. While advancements like AlphaGo Zero

1/2 REMAINING › REGISTER

Limited agency to enable explainability



Business Practitioner Vendor

- Regulatory constraints
- Deployment/maintenance costs

- Limited agency
- Mitigation after system is built

XAI as an afterthought

- Deployment constraints make ad hoc explanations tedious and costly in practice
- Generating explanations is no guarantee for transparent outcomes
- Cost-benefit balance to consider when XAI is a desirable property for a product

End users are not helped by XAI

The way forward

Need for culture shift

How can we make this system explainable in deployment?

How do we establish SOPs to ensure transparency & accountability throughout the development process?

Ask fundamental questions

- Why do you need AI for this task?
- Is the system transparent?
- When and how does the system fail?
- What are the potential harms that could occur?
- What types of explanations are needed? for whom?
- Can we ensure explainable outcomes?
- Who is responsible for ensuring transparency/XAI?

XAI as a process rather than a product

- Mobilise the AI community to develop useful XAI tools that help solve realistic and relevant problems while embracing the challenges of real world datasets and collaboration with domain experts
- Encourage and create meaningful incentives for a stakeholder-centric approach to create useful applications and systems
- Embrace and normalise direct communication between stakeholders and AI developers/researchers

Thank you!

h.haned@uva.nl
https://hindantation.github.io/