Trustworthy AI for Business and Society

Ilker Birbil

Introduction

Logistics

Instructors

Ilker Birbil
(Lectures)

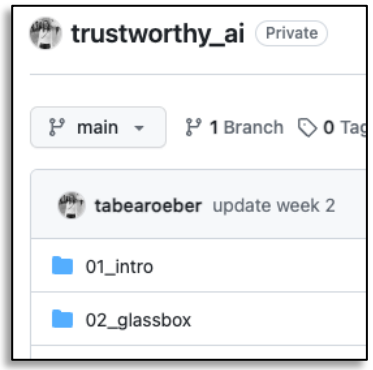
Tabea Röber
(Tutorials)

Prerequisites algorithmic machine learning; basics of optimization and probability; programming

Literature Lecture notes, presentation material, and research papers

Online Canvas and GitHub pages

(link)



(link)

Logistics

Software

Tutorial: 7 February 2024 (W)

Setup for Tutorials and Assignments

Assessment

- Group assignments (40%) - each group with two members
- Final exam (60%)
- Minimum of 5.0 to attend the final exam
- Assignments cannot be retaken - results remain valid for 2023-2024

Tutorials & Assignments

Week 02 – Glassbox Models

In this notebook we'll be applying different glassbox models to the same dataset and compare their performance in terms of accuracy and f1-score and their interpretability.

If you haven't done so yet, please run the following in your command line so you're able to access the relevant packages:

```
# navigate to parent directory using cd and activate environment
cd trustworthy_ai
source trustworthy_ai_venv/bin/activate

# for Rule Generation (RUG)
git clone https://github.com/sibirbil/RuleDiscovery.git

# for General Additive Models (GAMs), Explainable Boosting Machine (EBM), and Decision Lists
pip install interpret statsmodels skope-rules

# for Sparse Decision Trees
pip install cython
git clone https://github.com/ubc-systopia/pydl8.5-lbguess.git
cd pydl8.5-lbguess
python3 setup.py install

# close environment
deactivate
```

import packages

In [1]:
```
import warnings
warnings.filterwarnings("ignore")
import numpy as np
import pandas as pd
from sklearn.metrics import accuracy_score, f1_score, confusion_matrix, roc_auc_score
import os
import sys
import matplotlib.pyplot as plt
```

Load the dataset

We are using the Titanic dataset, which holds data about passangers of the Titanic and whether they survived or not. Passengers are described by 7 features. The response variable is binary (0 – died; 1 - survived).

Make sure to save the dataset in the parent directory or adjust the file path below.

Assignment 01

Deadline: February 21st 2024, 10:00 AM

Deliverables:

Please submit, via Canvas, a **zip file** including the following:

- a .ipynb file (you can use this one) with all of your code included -- `01_assignment_surname1_surname2.ipynb`
- a compiled .html file of your .ipynb which includes all of the output -- `0x_assignment_surname1_surname2.html`
- a pdf file with your written answers to the questions -- `01_assignment_surname1_surname2.pdf`

Make sure to follow the naming convention indicated above. The zip name can be named `01_assignment_surname1_surname2`.

Make sure to annotate your code. We may substract up to ... points if code is not annotated and unclear.

Data

We are using the Statlog (German Credit Data) dataset. The German Credit dataset classifies people described by a set of 20 features as good or bad credit risk.

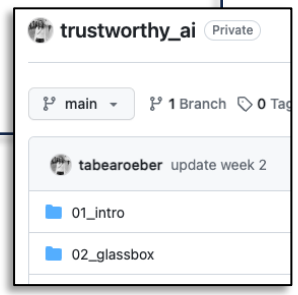
In [1]:
```
import warnings
warnings.filterwarnings("ignore")
import pandas as pd
```

In [ ]:
```
# complete dataset
loan_data = pd.read_csv('../datasets/credit/credit-g_csv.csv')

# train_data
X_train = pd.read_csv('../datasets/credit/credit_g_X_train.csv')
y_train = pd.read_csv('../datasets/credit/credit_g_y_train.csv')

# test data
X_test = pd.read_csv('../datasets/credit/credit_g_X_test.csv')
y_test = pd.read_csv('../datasets/credit/credit_g_y_test.csv')
```

---

1. Preparation

(5 points)



Sample exam will be constructed gradually with each assignment.

Artificial Intelligence

MIT Sloan Management Review

ME, MYSELF, AND AI • EPISODE 209

AI and the COVID-19 Vaccine: Moderna's Dave Johnson

July 13, 2021 / The pharma company's chief data and artificial intelligence officer discusses the digital biotech's platform approach to data science.

WORLD ECONOMIC FORUM

Why artificial intelligence is vital in the race to meet the SDGs

Alice Gast
President, Imperial College London

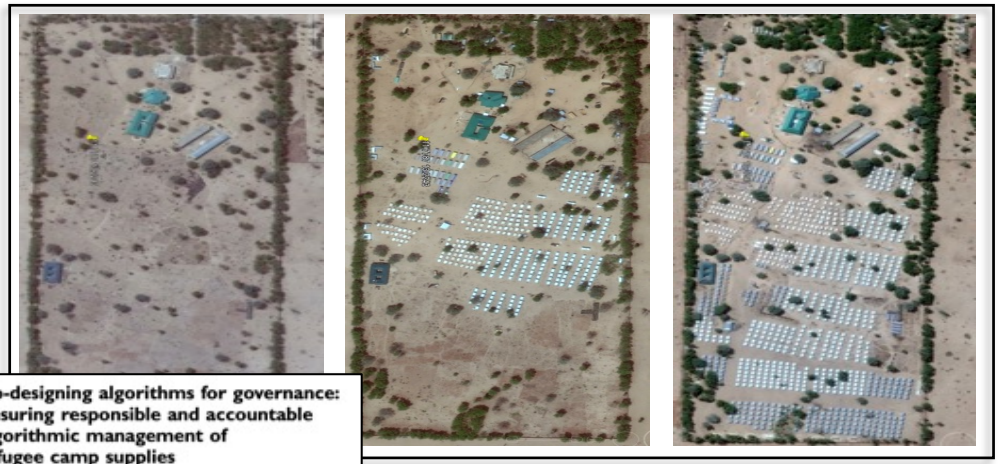
nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄ Subscribe

nature › news › article

…erything': …nakes gigantic l… in structures

…termining the 3D shapes of proteins st…



Co-designing algorithms for governance: Ensuring responsible and accountable algorithmic management of refugee camp supplies

Rianne Dekker[1], Paul Koot[2], S. Ilker Birbil[3] and Mark van Embden Andres[4]

Forbes

The Amazing Opportunities Of AI In The Future Of The Educational Metaverse

Rem Darbinyan Forbes Councils Member
Forbes Technology Council
COUNCIL POST | Membership (Fee-Based)

Apr 27, 2022, 08:00am EDT

However, …

CBC

Hong Kong protesters use laser pointers to deter police, scramble facial recognition

Police label the hand-held lasers 'offensive weapons'

Adam Jacobson · CBC News · Posted: Aug 11, 2019 4:00 AM ET | Last Updated: August 11, 2019

HARVARD LAW REVIEW

PRIVACY

The Dangers of Surveillance

Symposium Article by Neil M. Richards

MAY 20, 2013

126 Harv. L. Rev. 1934

The New York Times

PRIVACY

With a Few Bits of Data, Researchers Identify 'Anonymous' People

BY NATASHA SINGER JANUARY 29, 2015 2:01 PM 12

CGAP

BLOG 05 SEPTEMBER 2019

Algorithm Bias in Credit Scoring: What's Inside the Black Box?

By Maria Fernandez Vidal, Jacobo Menajovsky

> The responsible use of algorithms requires providers to know which variables are being considered in their credit scoring models and how they are affecting people's scores.

de Rechtspraak

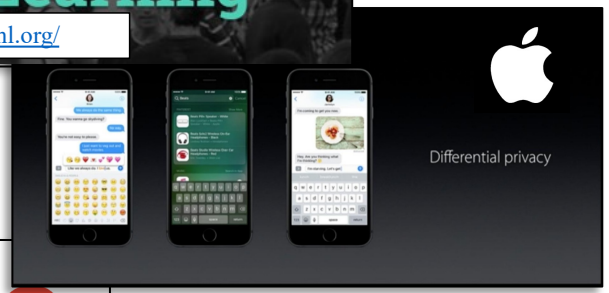
SyRI legislation in breach of European Convention on Human Rights

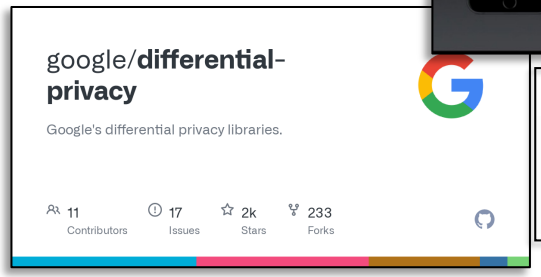
NOS

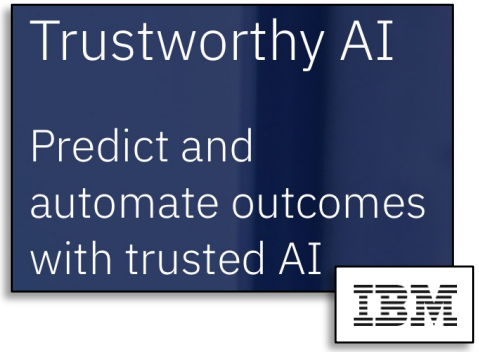
NOS Nieuws • Woensdag 5 februari 2020, 10:56 •
Aangepast woensdag 5 februari 2020, 13:54

Anti-fraudesysteem SyRI moet van tafel, overheid maakt inbreuk op privéleven

Trustworthy Algorithms







THE AI ACT

The Artificial Intelligence Act

European Commission

POLICY AND LEGISLATION | Publication 08 April 2019

Communication: Building Trust in Human Centric Artificial Intelligence

Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Building Trust in Human Centric Artificial Intelligence (COM(2019)168)

2.1. Guidelines for trustworthy AI drafted by the AI high-level expert group

The seven key requirements are:

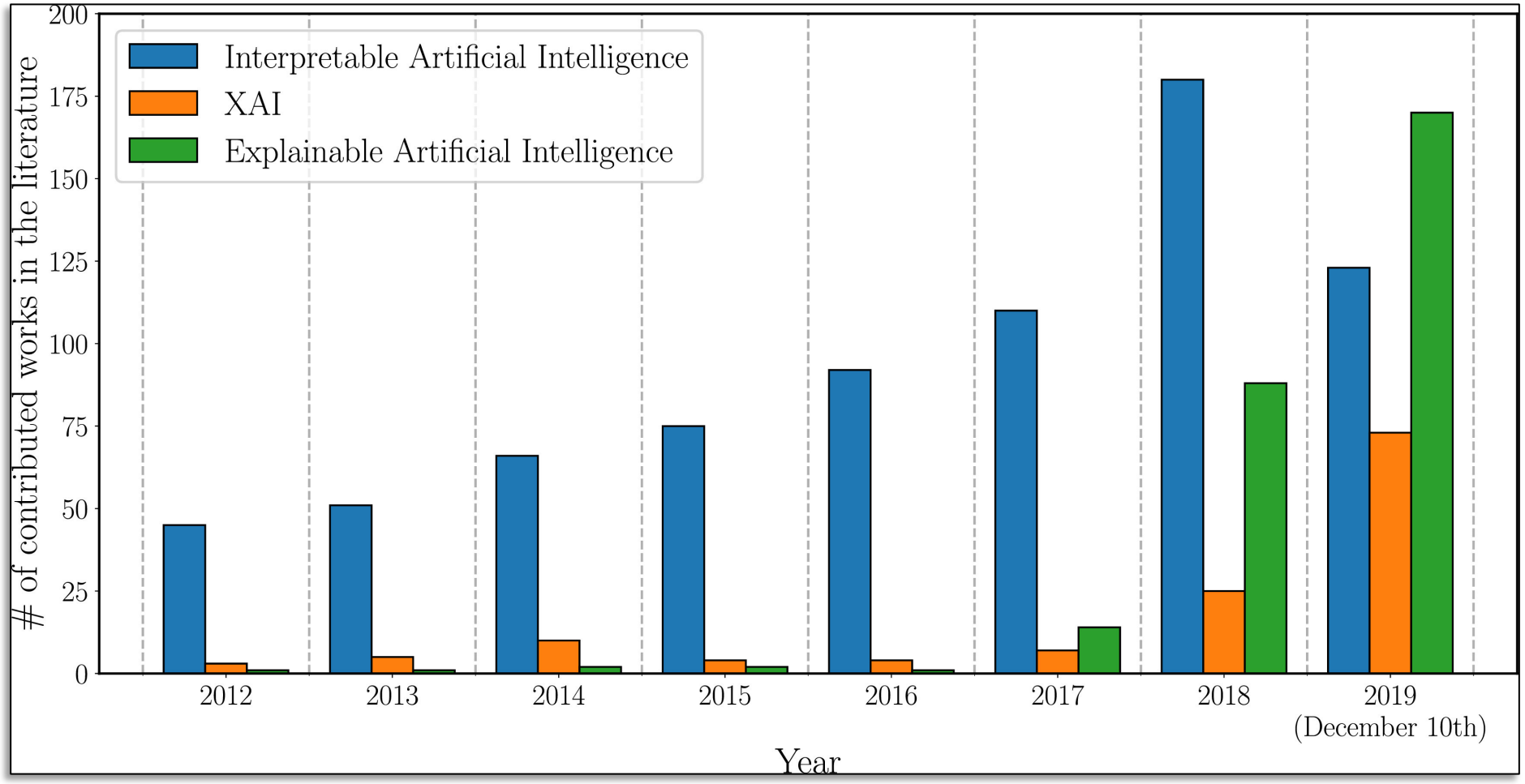
- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental well-being
- Accountability

Right to Explanation

The European Union enacted the **right to explanation** which was incorporated in the EU General Data Protection Regulation (GDPR) in 2018:

[...] In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to **obtain an explanation of the decision reached after such assessment** and to challenge the decision. [...]

Interpretable vs. Explainable

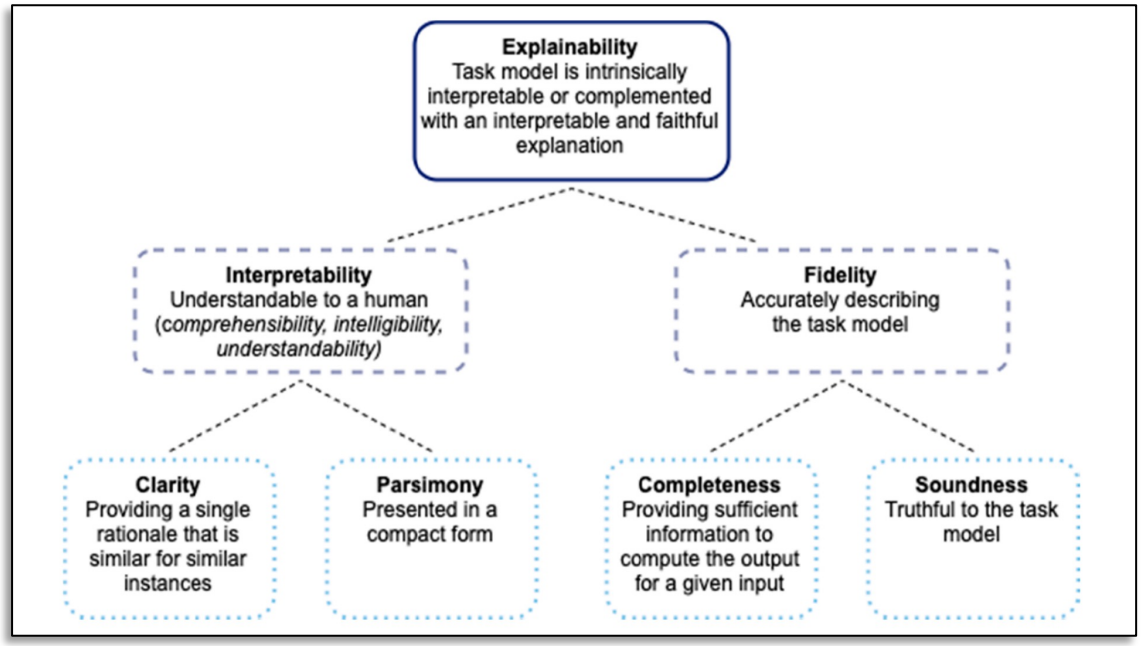


Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, 82-115, 2020.

Interpretable vs. Explainable

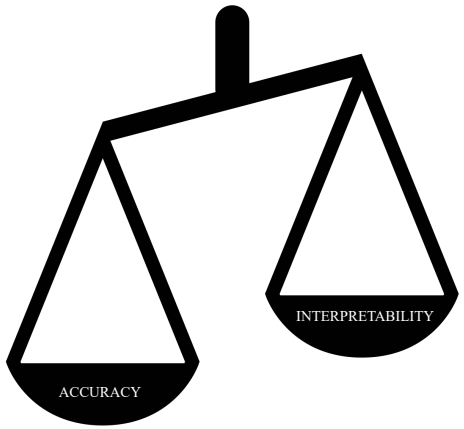
"An AI system is **explainable** if the task model is intrinsically interpretable (here the AI system is the task model) or if the non-interpretable task model is complemented with an interpretable and faithful explanation (here the AI system also contains a post-hoc explanation)."

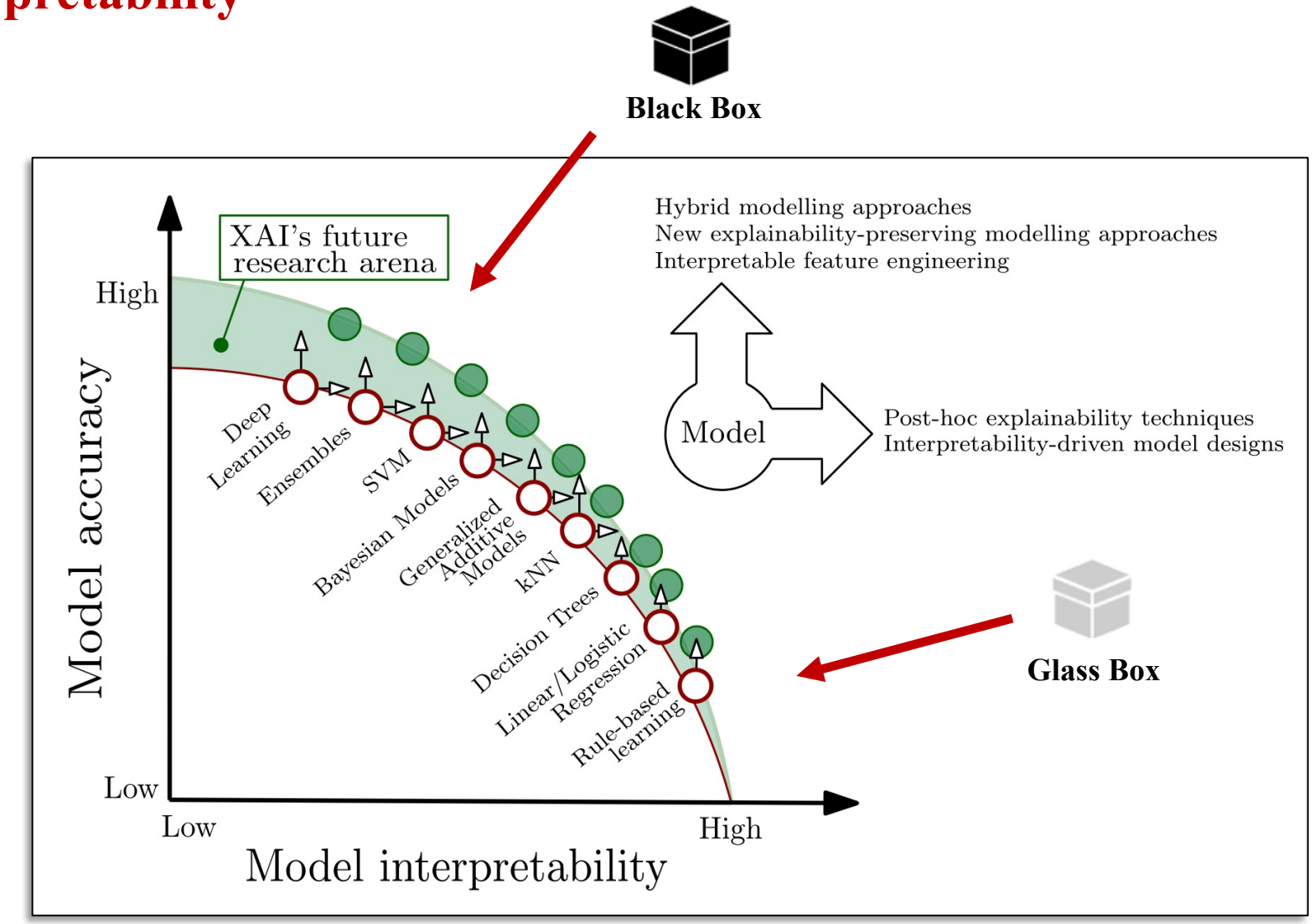
by Markus, Kors and Rijnbeek, 2021



Markus, F. A., Kors J. A. & Rijnbeek, P. R. "The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies," Journal of Biomedical Informatics, (113) 103655, 2021.

Accuracy vs. Interpretability

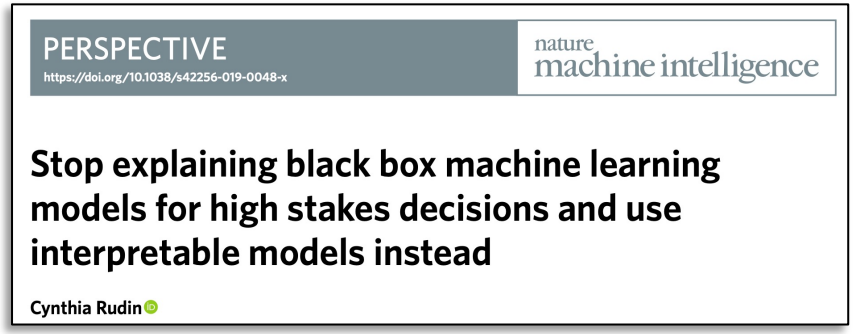




Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, 82-115, 2020.

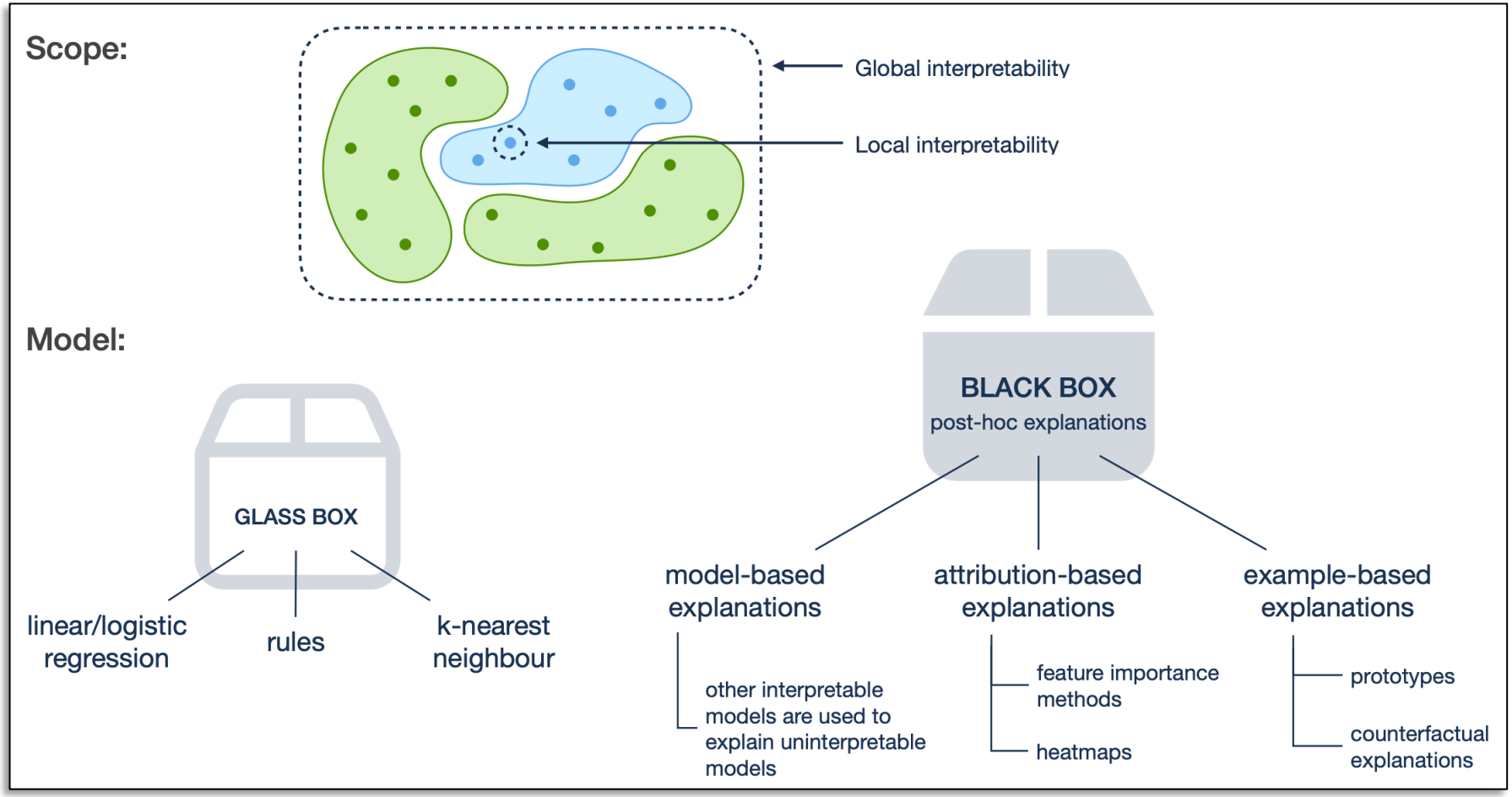
Inherently Interpretable Models

"It is a myth that there is necessarily a trade-off between accuracy and interpretability. There is a widespread belief that more complex models are more accurate, meaning that a complicated black box is necessary for top predictive performance. However, this is often not true, particularly when the data are structured, with a good representation in terms of naturally meaningful features. When considering problems that have structured data with meaningful features, there is often no significant difference in performance between more complex classifiers (deep neural networks, boosted decision trees, random forests) and much simpler classifiers (logistic regression, decision lists) after preprocessing. "



(link)

A Taxonomy of XAI Methods



by T. Röber, 2023

Privacy: Federated Learning

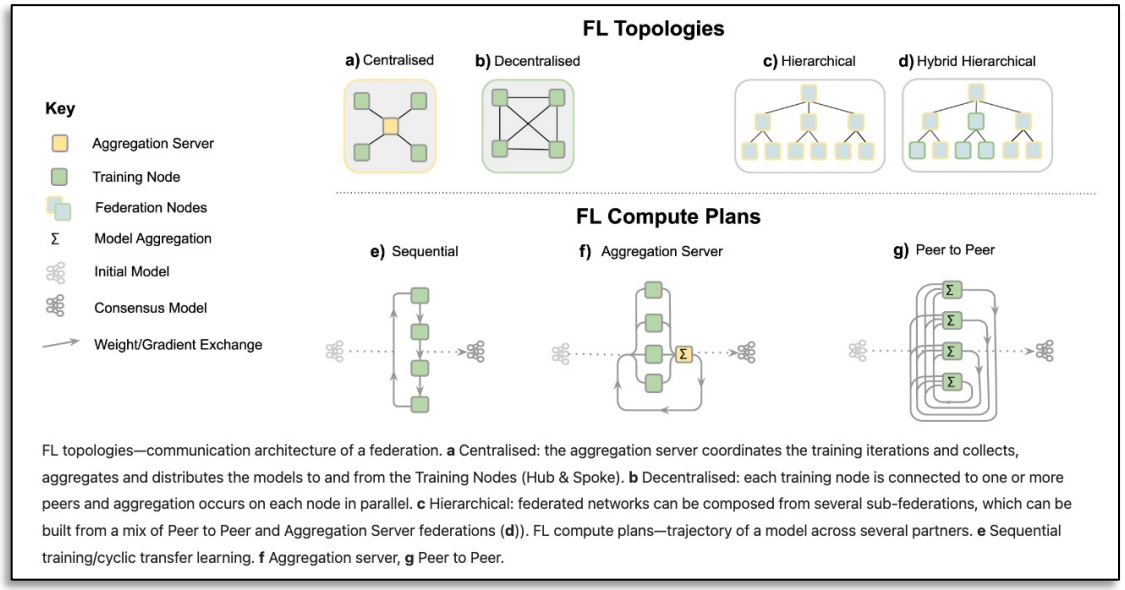
Federated Learning and Privacy: Building privacy-preserving systems for machine learning and data science on decentralized data

Authors: Kallista Bonawitz, Peter Kairouz, Brendan McMahan, Daniel Ramage Authors Info & Claims

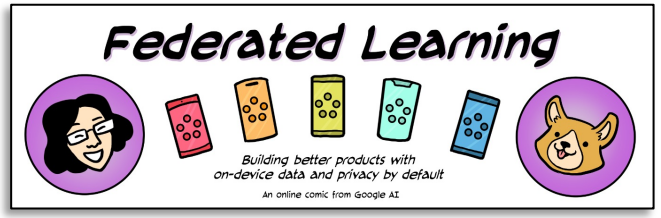
Queue, Volume 19, Issue 5 • Pages: 40, pp 87–114 • https://doi.org/10.1145/3494834.3500240

(link)

"Federated Learning is a machine learning setting where multiple entities (clients) collaborate in solving a machine learning problem, under the coordination of a central server or service provider. Each client's raw data is stored locally and not exchanged or transferred; instead, focused updates intended for immediate aggregation are used to achieve the learning objective."



FL topologies—communication architecture of a federation. **a** Centralised: the aggregation server coordinates the training iterations and collects, aggregates and distributes the models to and from the Training Nodes (Hub & Spoke). **b** Decentralised: each training node is connected to one or more peers and aggregation occurs on each node in parallel. **c** Hierarchical: federated networks can be composed from several sub-federations, which can be built from a mix of Peer to Peer and Aggregation Server federations (**d**)). FL compute plans—trajectory of a model across several partners. **e** Sequential training/cyclic transfer learning. **f** Aggregation server, **g** Peer to Peer.



(link)

Rieke et al. "The future of digital health with federated learning,"npj Digital Medicine, (3)119, 2020.

Privacy: Differential Privacy

"Differential privacy addresses the paradox of learning nothing about an individual while learning useful information about a population."

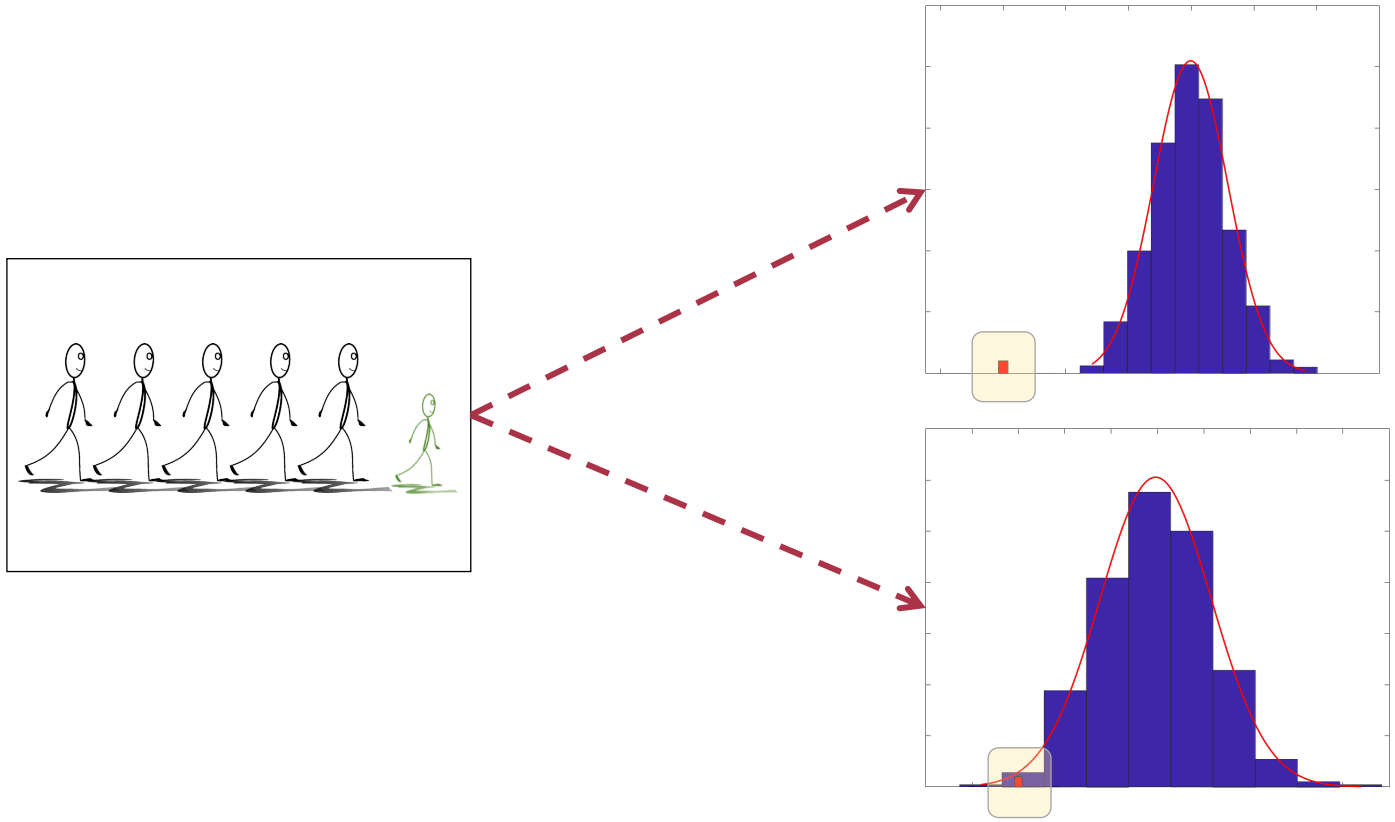
Mathematically provable privacy guarantee!

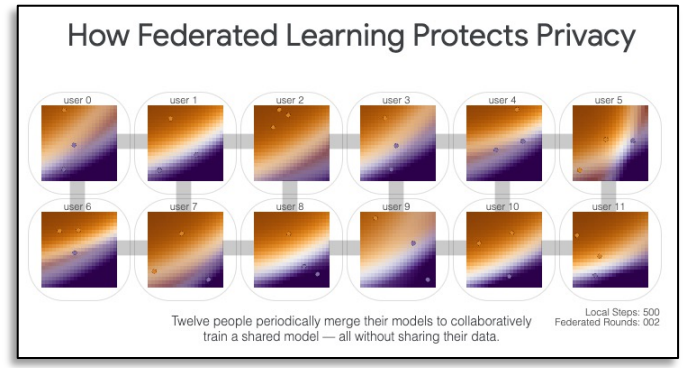
The Algorithmic Foundations of Differential Privacy

Cynthia Dwork
Microsoft Research, USA
dwork@microsoft.com

Aaron Roth
University of Pennsylvania, USA
aaroth@cis.upenn.edu

(link)





(link)

Privacy vs. Explainability

Model Reconstruction from Model Explanations

Smitha Milli
smilli@berkeley.edu
University of California, Berkeley

Ludwig Schmidt
ludwig@berkeley.edu
University of California, Berkeley

Anca D. Dragan
anca@berkeley.edu
University of California, Berkeley

Moritz Hardt
hardt@berkeley.edu
University of California, Berkeley

(link)

"Our work demonstrates that establishing usable explanation methods for machine learning models faces another hurdle in commercial applications. Whatever criteria of explanation quality we choose must be weighed against the risk of model leakage resulting from the method at hand."

"In this work, we demonstrate significant vulnerabilities in post hoc explanation techniques that can be exploited by an adversary to generate classifiers whose post hoc explanations can be arbitrarily controlled."

Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods

Dylan Slack*
University of California, Irvine
dslack@uci.edu

Sophie Hilgard*
Harvard University
ash798@g.harvard.edu

Emily Jia
Harvard University
ejia@college.harvard.edu

Sameer Singh
University of California, Irvine
sameer@uci.edu

Himabindu Lakkaraju
Harvard University
hlakkaraju@seas.harvard.edu

(link)

On the Privacy Risks of Model Explanations

Reza Shokri
reza@comp.nus.edu.sg
National University of Singapore

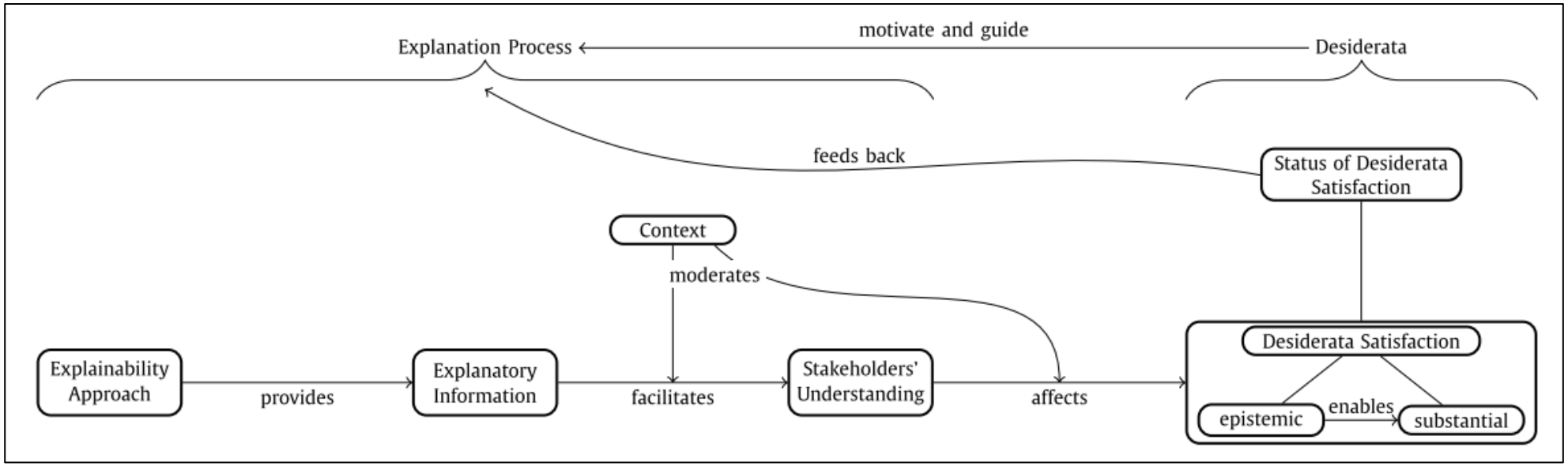
Martin Strobel
mstrobel@comp.nus.edu.sg
National University of Singapore

Yair Zick*
yzick@umass.edu
University of Massachusetts, Amherst

(link)

"Our work is the first to extensively analyze the data privacy risks that arise from releasing model explanations, which can result in a trade-off between transparency and privacy."

User's Perspective



Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research ☆,☆☆

Markus Langer [a,*,1], Daniel Oster [b,1], Timo Speith [b,c,*,1], Holger Hermanns [c,d], Lena Kästner [b], Eva Schmidt [e], Andreas Sesing [f], Kevin Baum [b,c]

(link)

What is this course about?

- Introduction
- XAI Methods
  - Glass Box Methods
  - Unboxing Methods
  - Other Methods
- Privacy
  - Federated Learning
  - Differential Privacy

Anything missing?

Other glass box and unboxing approaches
Other privacy approaches
Discussion on fairness and accountability
…

Trustworthy AI: From Principles to Practices

BO LI, JD Technology, China and Tsinghua University, China
PENG QI, Amazon AWS AI Labs, USA
BO LIU, Walmart Inc., USA
SHUAI DI, JD Technology, China
JINGEN LIU, JD Technology, USA
JIQUAN PEI, JD Technology, China
JINFENG YI, Frontis.AI, China
BOWEN ZHOU, Tsinghua University, China and Frontis.AI, China

(link)

Reading Material