



The Opacity Problem in ML

Slides from Yeji Streppel, with small
adaptations by Nickel.

Today

- **Why is opacity a problem?**
- **What does the opacity problem look like in AI?**
- **How to achieve transparency?**
- **Conclusions / takeaways**



**Why is opacity a
problem?**

Increasing interest in XAI (eXplainable Artificial Intelligence)

- **Omnipresent**
AI is everywhere



Increasing interest in XAI (eXplainable Artificial Intelligence)

- **Omnipresent**
- **High stakes contexts**
Health care, finance,
criminal justice,
employment, housing...

THE LANCET
Digital Health

This journal Journals Publish Clinical Global health Multimedia Events About

ARTICLES · Volume 7, Issue 8, 100882, August 2025 · *Open Access* [Download Full Issue](#)

AI as an independent second reader in detection of clinically relevant breast cancers within a population-based screening programme in the Netherlands: a retrospective cohort study

[Suzanne L van Winkel, MSc](#)^a · [Jim Peters, MSc](#)^b · [Natasja Janssen, PhD](#)^c · [Jaap Kroes, PhD](#)^c · [Elizabeth A Loehrer, PhD](#)^{a,d} · [Jessie Gommers, MSc](#)^a · et al. [Show more](#)

[Affiliations & Notes](#) [Article Info](#)

[Download PDF](#) [Cite](#) [Share](#) [Set Alert](#) [Get Rights](#) [Reprints](#)

» Summary

Show Outline

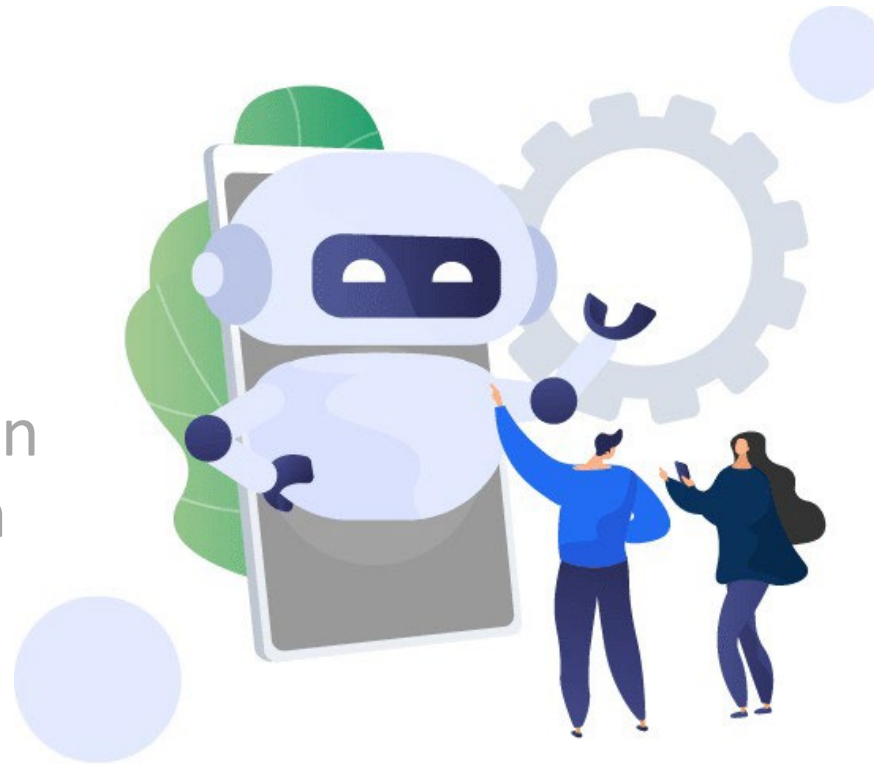
Background

Breast cancer screening programmes have shown to reduce mortality, but current methods face challenges such as limited mammographic sensitivity, limited resources, and variability in radiologist expertise. Artificial intelligence (AI) offers potential to improve screening accuracy and efficiency. This study simulated different screening scenarios, evaluating the performance of

Increasing interest in XAI (eXplainable Artificial Intelligence)

- **Omnipresent**
- **High stakes contexts**
- **Automation bias**

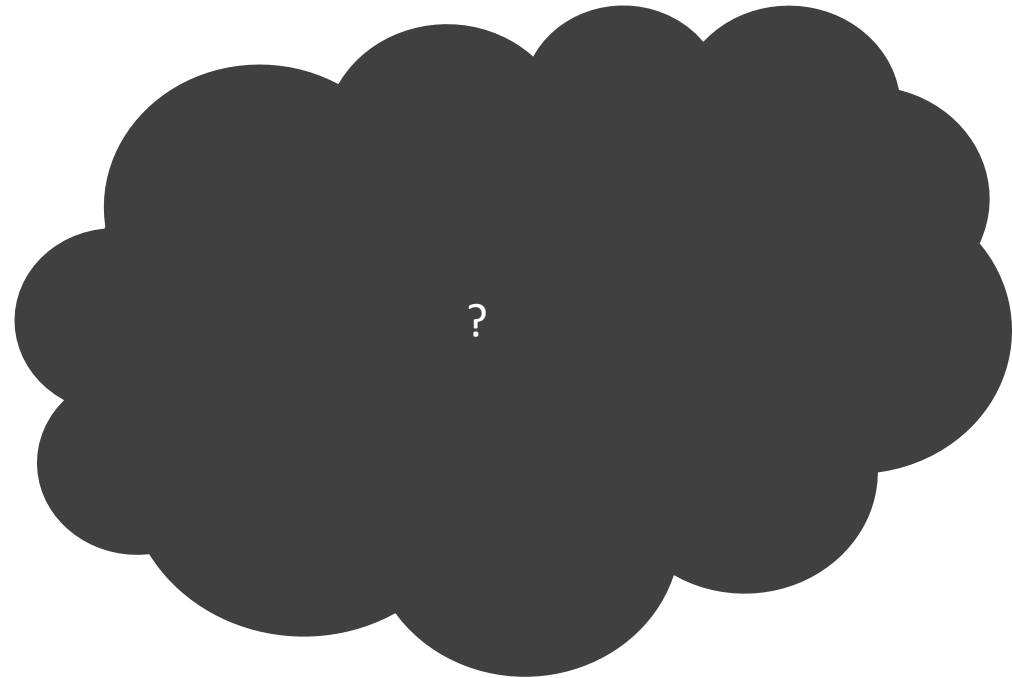
People tend to over-rely on automated systems, even when they know it's faulty



Increasing interest in XAI (eXplainable Artificial Intelligence)

- **Omnipresent**
- **High stakes contexts**
- **Automation bias**
- **Opacity**

Humans often don't understand the model's decision logic



Example

NEWS > TECHNOLOGY

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.

The Dutch tax authorities falsely accused thousands of families of fraud, based on an opaque “risk assessment” tool.

Example

HireVue's AI system, used for job screening, faces criticism for its opaque decision-making process, potential biases, and lack of transparency in how it evaluates applicants.

Rights group files federal complaint against AI-hiring firm HireVue, citing 'unfair and deceptive' practices

The Electronic Privacy Information Center urged the FTC to investigate HireVue's business practices, saying its face-scanning technology threatens job candidates' privacy rights and livelihoods.

Example

Amazon's recruiting tool learned an undesirable relationship between words related to women and the likelihood of a resume to be selected.

In effect, Amazon's system taught itself that male candidates were preferable. It penalized resumes that included the word "women's," as in "women's chess club captain." And it downgraded graduates of two all-women's colleges, according to people familiar with the matter. They did not specify the names of the schools.

Amazon edited the programs to make them neutral to these particular terms. But that was no guarantee that the machines would not devise other ways of sorting candidates that could prove discriminatory, the people said.

Why is opacity a problem?

The need for interpretability comes from **incompleteness** of a machine learning task formulation in relation to other things we want, for example:

Model Validation

Validate and debug the model.

User Acceptance

Convince users that the model makes sense.

Assess Fairness

Determine whether and how the model uses discriminating or arbitrary features.

Knowledge Discovery

Generate hypotheses based on the identified relationships.

Justification

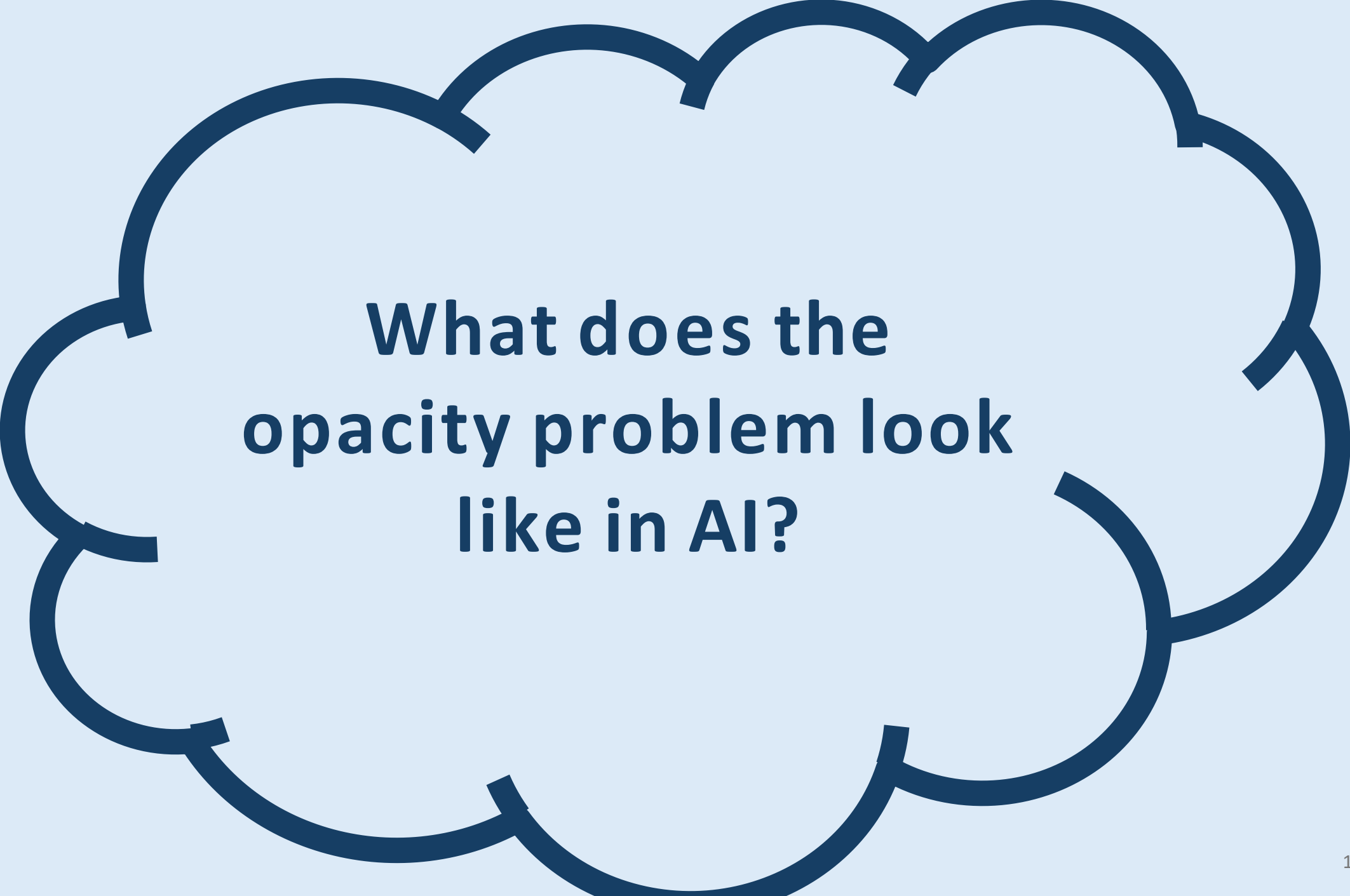
Justify that a prediction is good.

Assess Trustworthiness

Determine whether we can trust a particular prediction.

Why is opacity a problem?

- If we don't know how a system works, we cannot trust, challenge, improve, or intervene on its outputs and internal decision logic. We have no control / oversight.
- **Opacity in automated decision support systems causes harm,** especially in high stakes contexts (finance, employment, health care, education, housing...)
- Harm can look like
 - Injustices
 - Lack of recourse, contestability, accountability
 - Unsafety
 - etc.

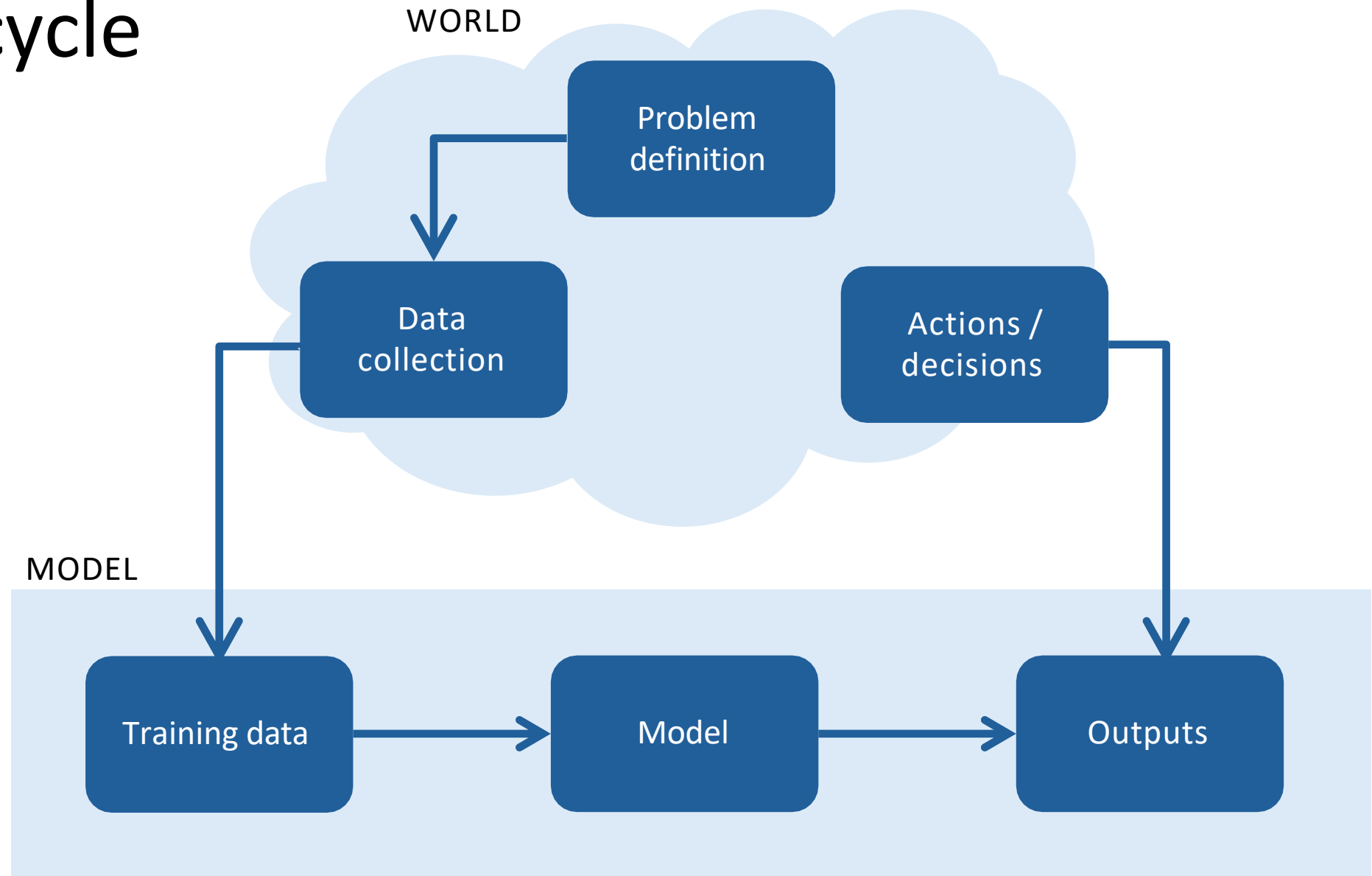


**What does the
opacity problem look
like in AI?**

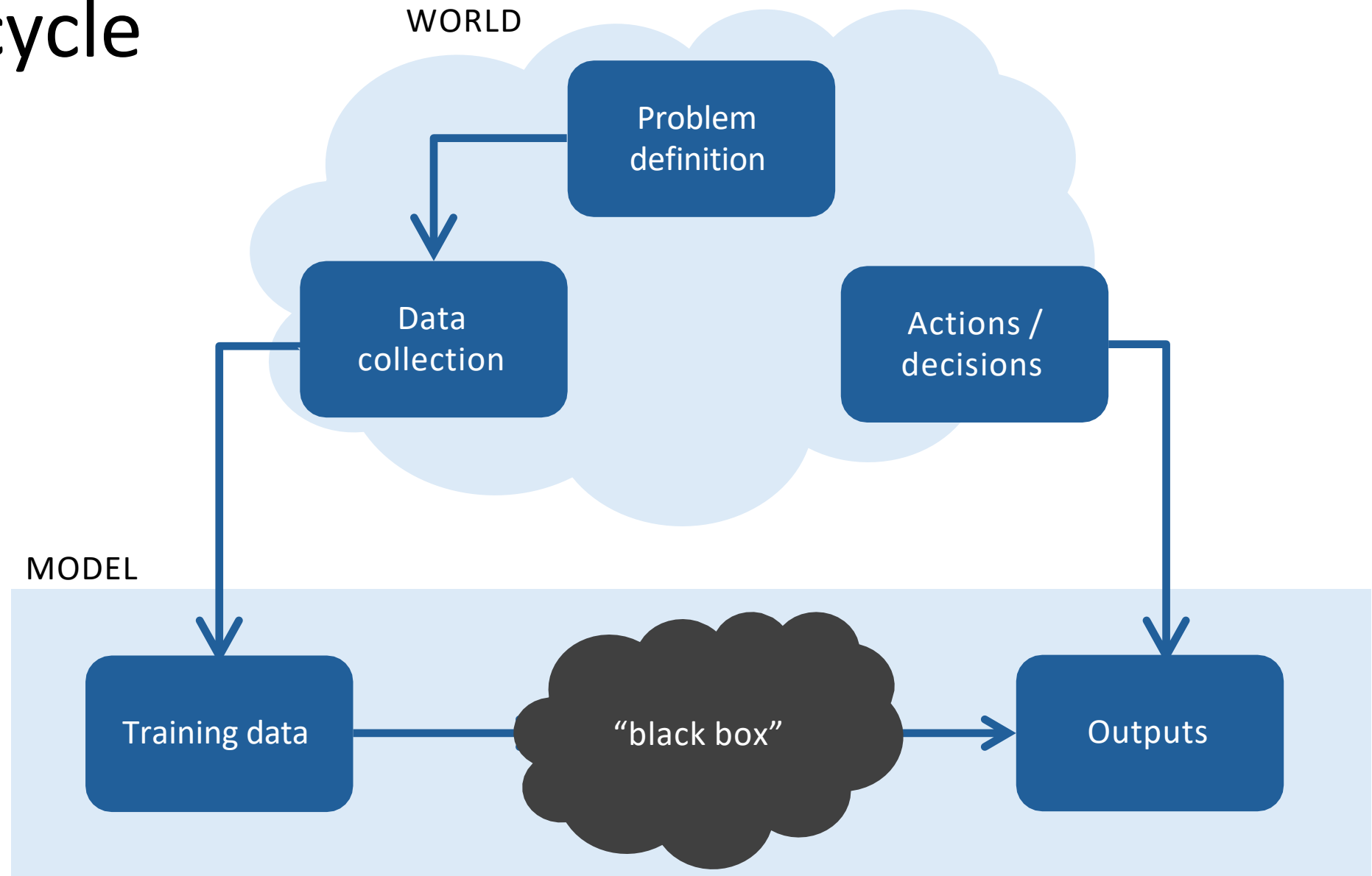
Black boxes?

When we talk about opacity in AI, we often think about “black box” systems and XAI techniques.

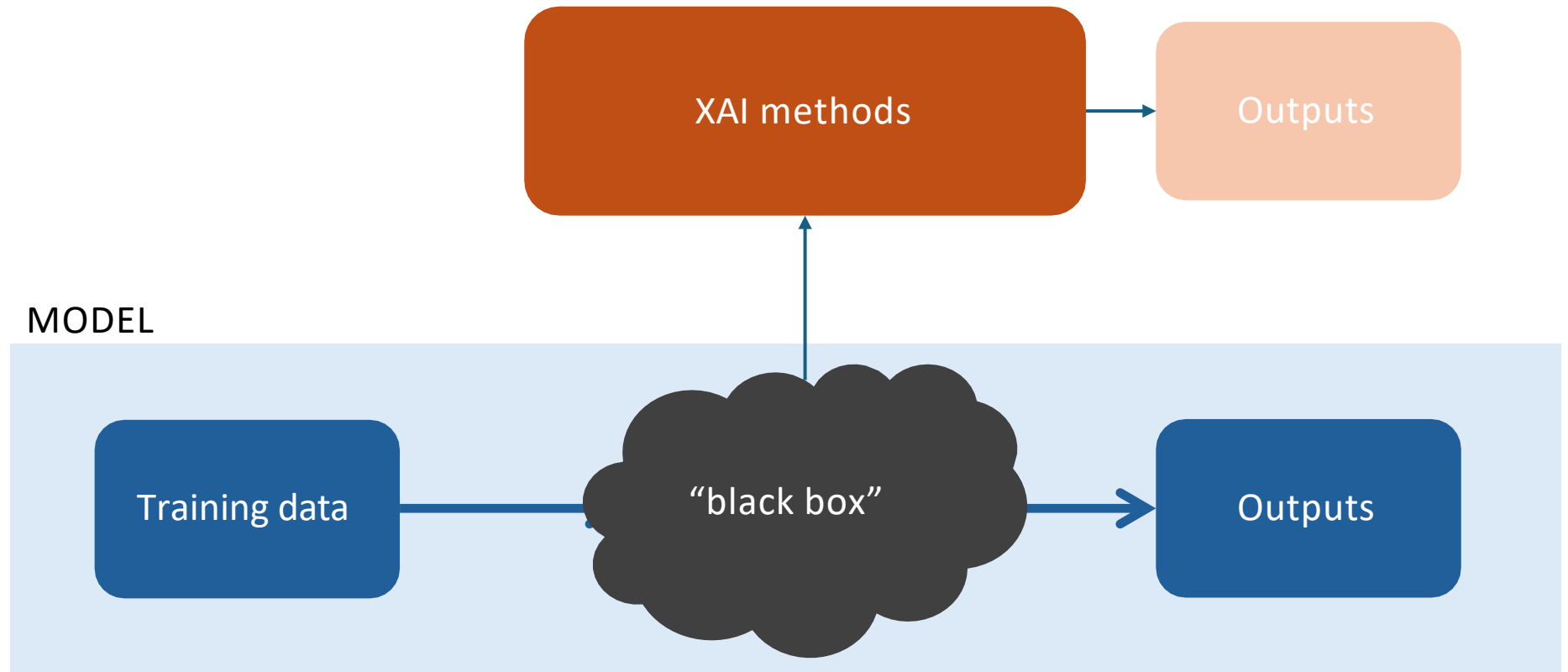
ML lifecycle



ML lifecycle



Technical explainability problem

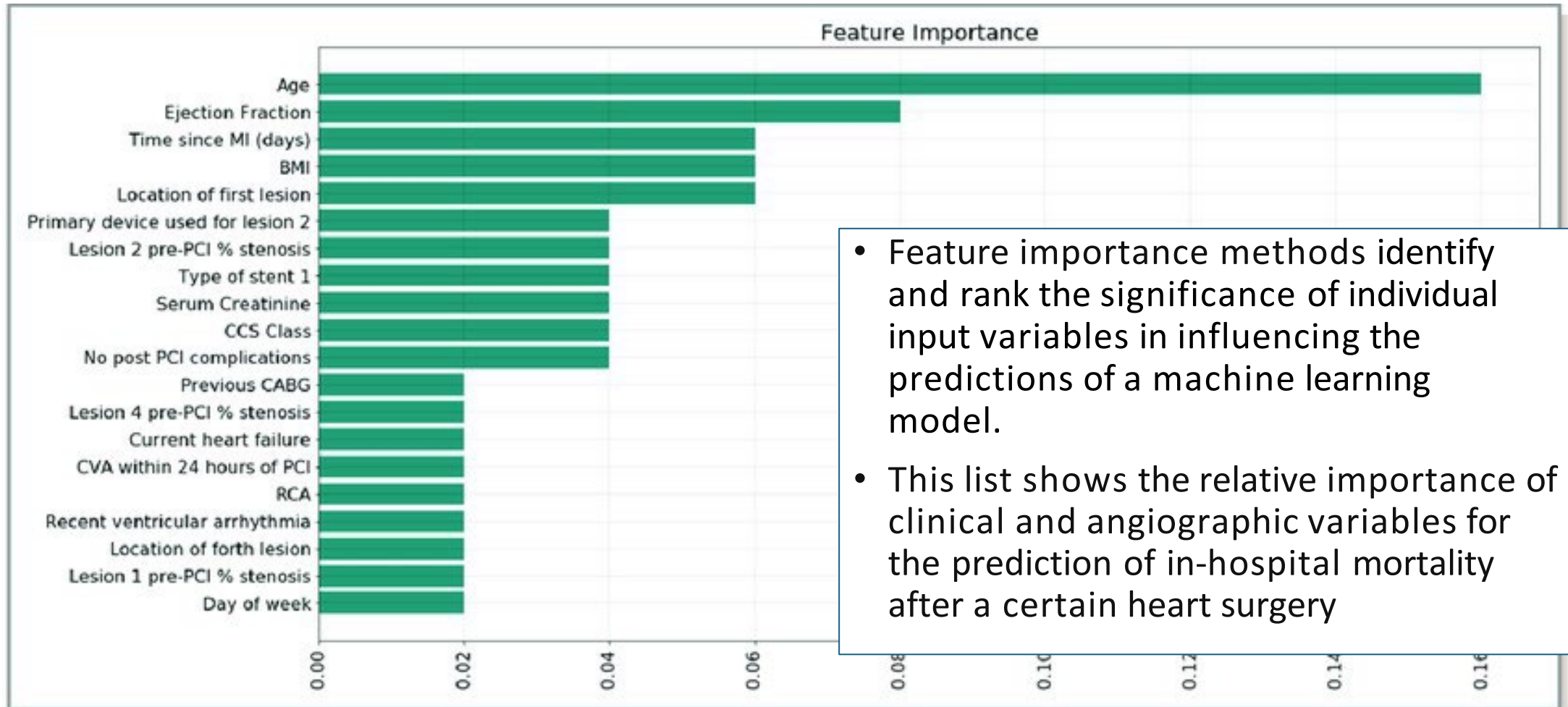


Example: saliency / heatmaps

- Heatmaps show which parts of an input (like an image) are most important for the model's decision-making process.
- E.g. if a system is identifying a dog in a picture, the heatmap might highlight the dog's face/eyes in warm colors. Less important regions, like the background, are shown in cooler colors.



Example: feature importance ranking



- Feature importance methods identify and rank the significance of individual input variables in influencing the predictions of a machine learning model.
- This list shows the relative importance of clinical and angiographic variables for the prediction of in-hospital mortality after a certain heart surgery

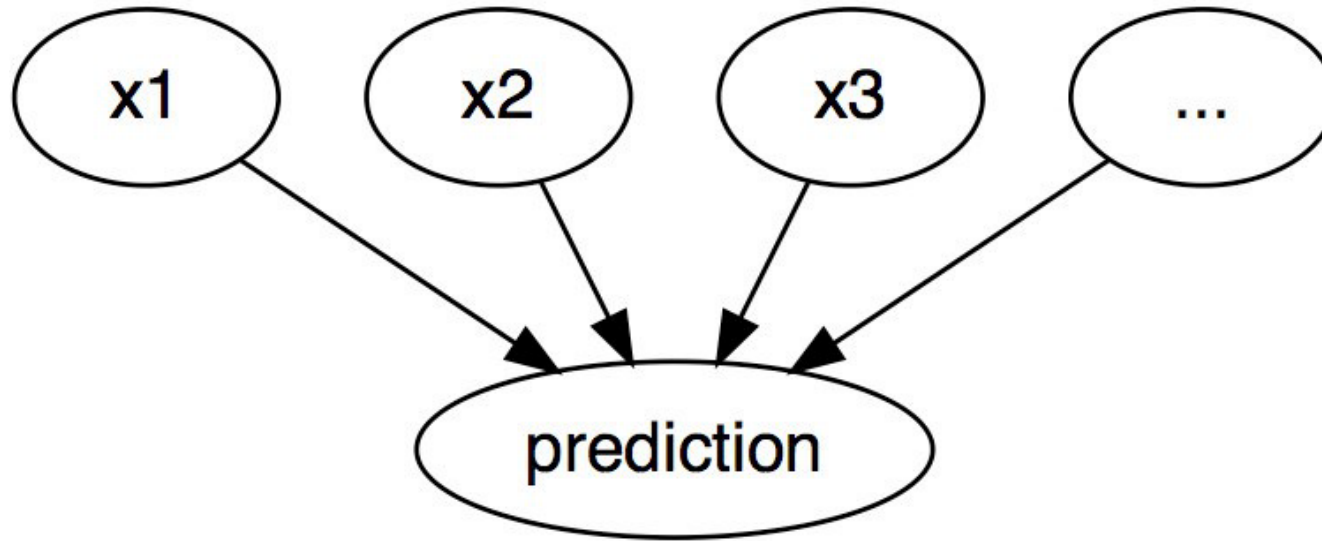
Example: counterfactual explanations

- A counterfactual explanation describes a causal situation in the form: “If **cause X** had not occurred, **event Y** would not have occurred”
- In ML, counterfactual explanations can be used to explain predictions of individual instances.
 - **Event Y** = predicted outcome of an instance
 - **Cause X** = particular feature values of this instance that were input to the model and “caused” a certain prediction
- A counterfactual explanation of a prediction describes the smallest change to the feature values that changes the prediction to a predefined output.

Example: counterfactual explanations

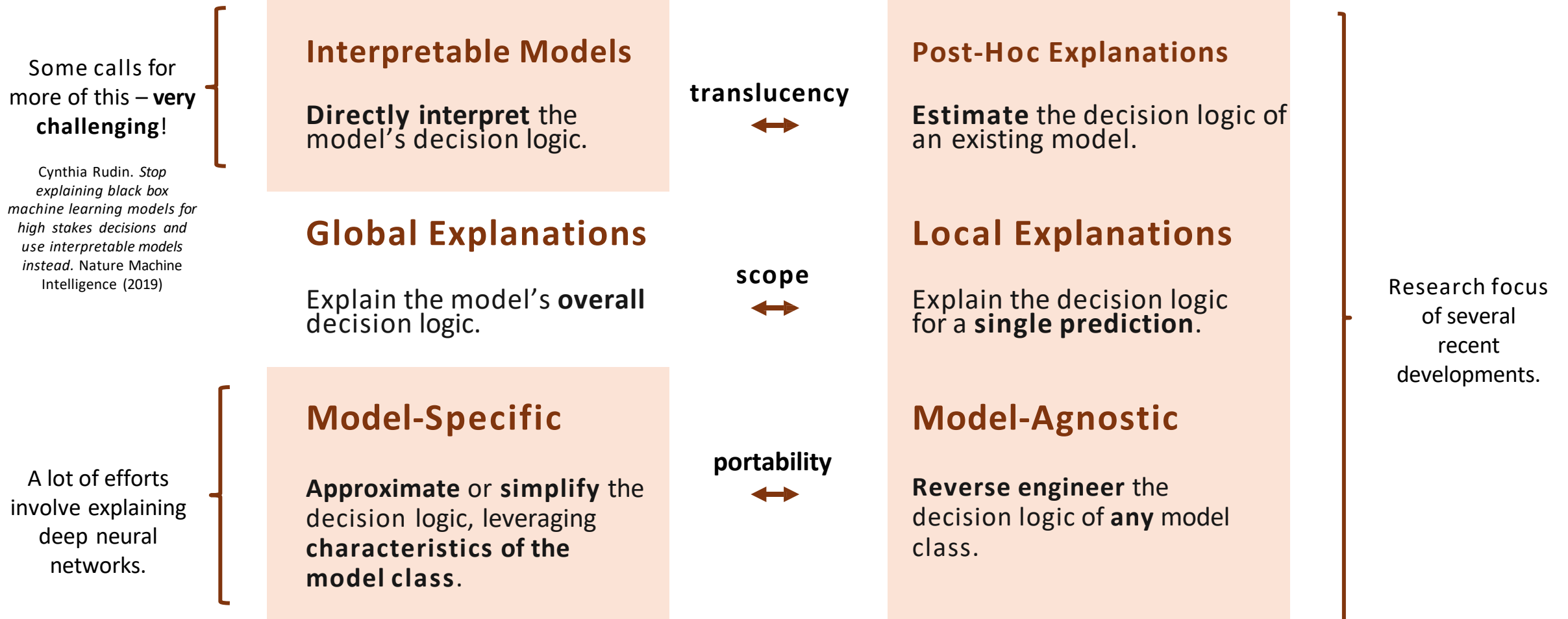
- A customer applies for a loan, but the application is rejected.
- Prediction (event Y): “Loan rejected”
- Cause (feature values):
 - Income: €40.000
 - Age: 52
 - Employment status: full-time
- Counterfactual explanation: “If the applicant’s income was €45.000, the loan would be approved.”
- This is the smallest change to the feature values that would have changed the prediction from “loan rejected” to “loan approved”.

Example: counterfactual explanations



- Popular because they are often seen as **intuitive, actionable**, and they directly target **causal relationships** in the model (as opposed to other methods that are more correlation-based)

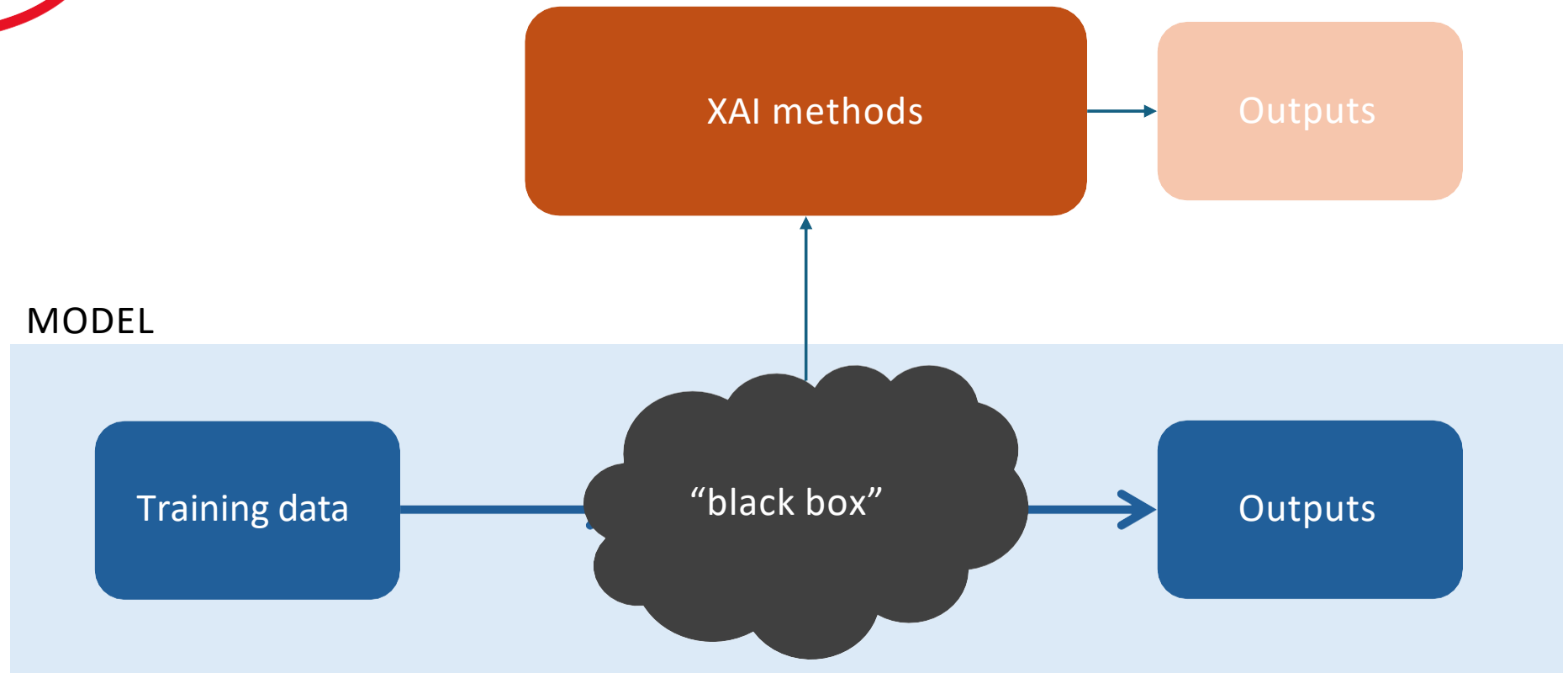
XAI Research



Technical explainability problem

XAI research

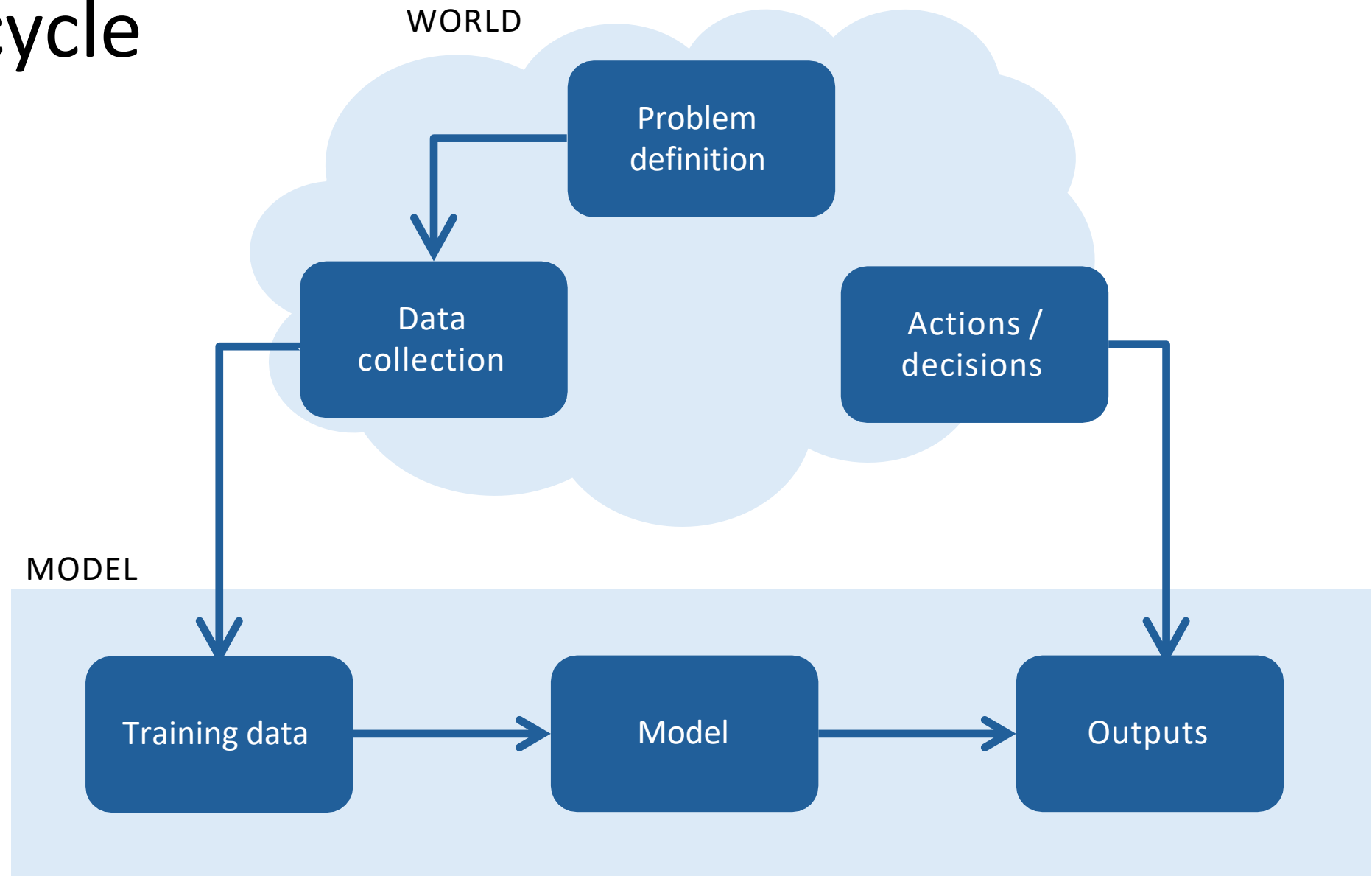
- Interpretable vs post-hoc
- Global vs local
- Model-specific vs model-agnostic



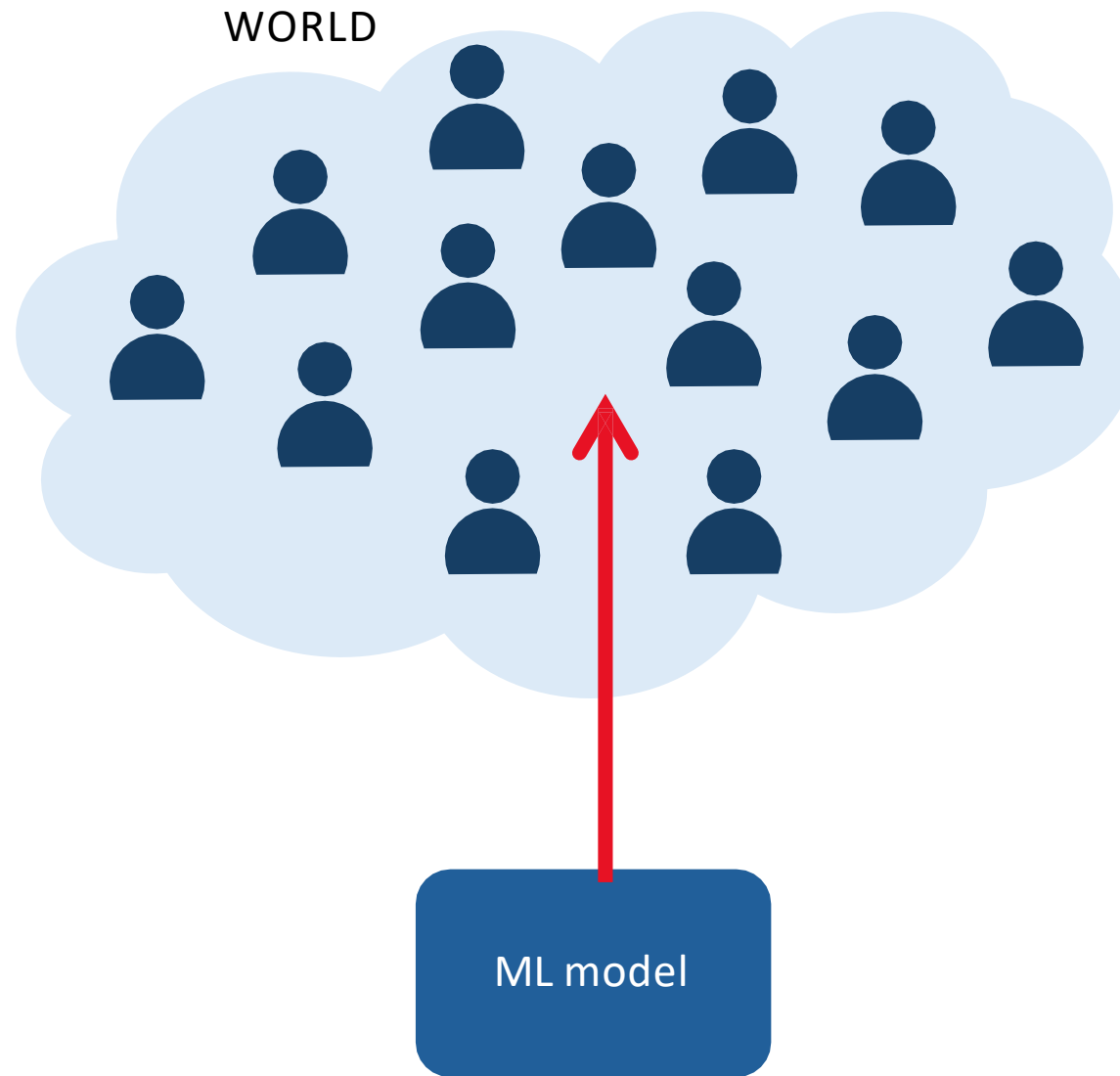
AI systems in context

- The opacity problem is much larger than just opaque algorithms.
- Like all technology, AI & data science work **for people**: they can make people's lives easier and more pleasant, but they can also pose a risk to their well-being.
- Like all technology, AI & data science is produced **by people**: they are not neutral, but reflect the attributes, values, habits and convictions of its creators.

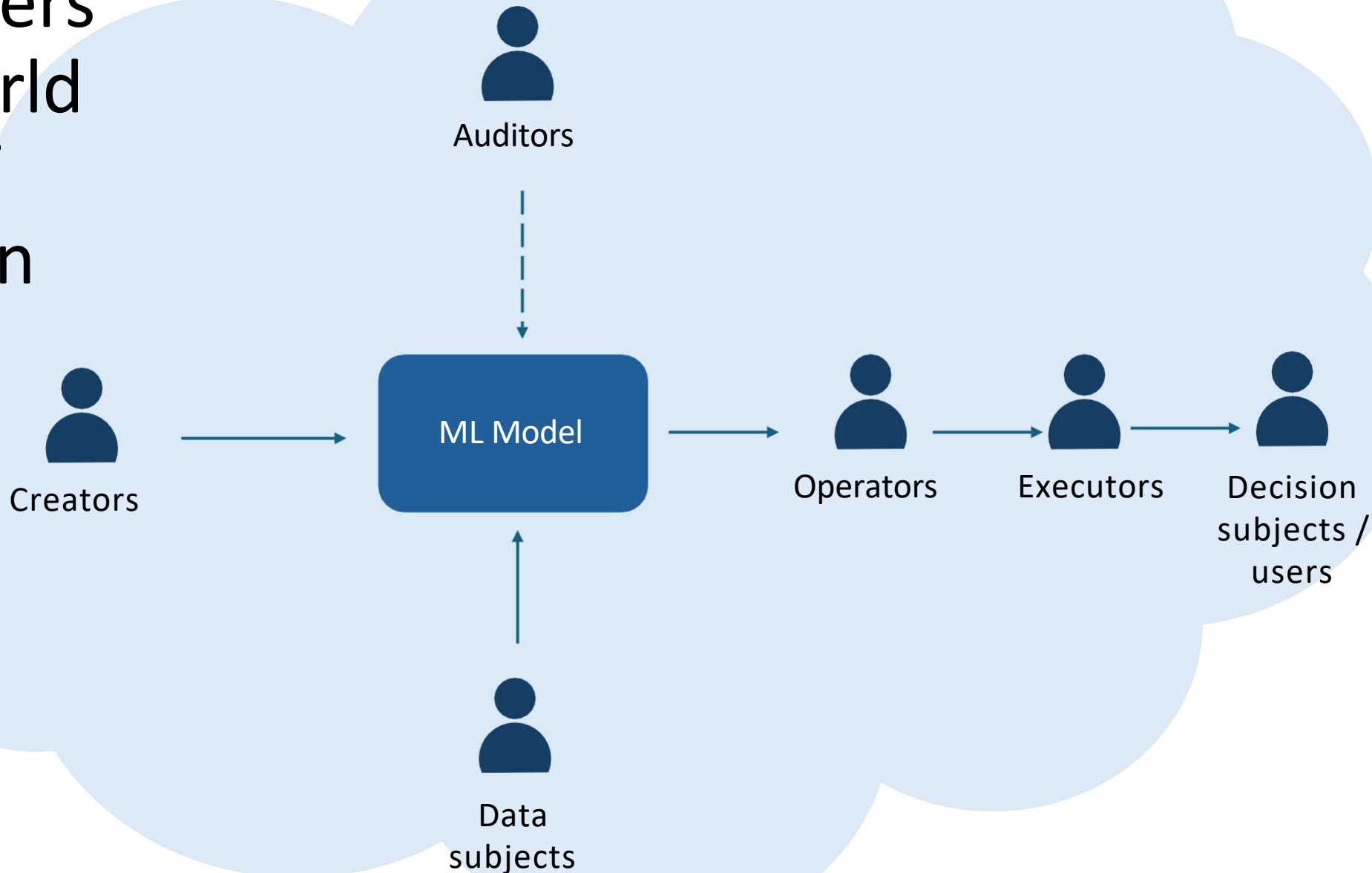
ML lifecycle



Modelling the real world: sociotechnical problem



Stakeholders in real-world context of application

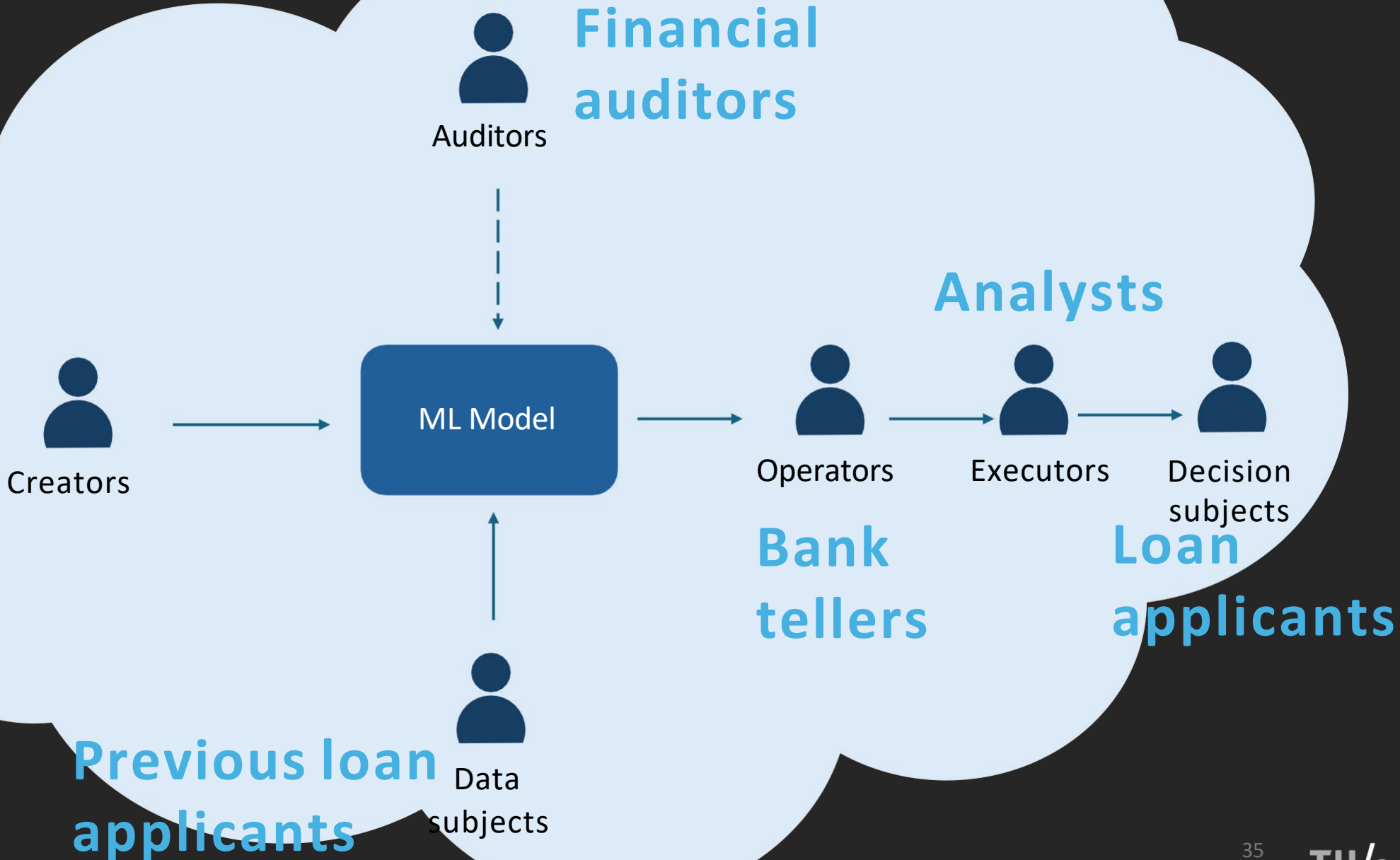


Example: credit scoring

- A **customer** goes to a bank to apply for a loan.
- A front-office **bank teller** takes their data and enters it into a software tool to compute a credit score.
- The tool was created by **software developers**, who trained it using **previous customers' data**, including demographics and repayment behavior.
- A back-office **analyst** uses the credit score to make a data-driven decision about the customer's application.
- At the end of the fiscal year, **financial auditors** evaluate the tool on the basis of functional criteria (e.g. profitability) and ethical norms (e.g. possible discrimination).

Example: credit scoring

Software
developers



Example: genAI for design

- A designer uses GenAI (e.g. DALL·E, MidJourney) for a branding project.
- The model was trained by **software developers** on datasets curated by **data labellers/annotators**.
- The datasets contain thousands of images and designs, coming from **data subjects** (such as artists, photographers, and designers)
- The service is hosted on a platform operated by a company like **OpenAI**.
- **Auditors** evaluate the model's compliance with IP laws, ethical norms etc.
- The designer refines the AI-generated content and presents final concepts to their **client** for feedback.

Opacity is stakeholder-dependent

- Stakeholders can be distinguished by their **prior knowledge** and **abilities** (Zednik 2021).
- Different **sources** of opacity (Burrell 2016):
 - ▶ Technical illiteracy
 - ▶ Intentional secrecy
 - ▶ System complexity

Sources of opacity

Stakeholder knowledge/abilities

- **Technical illiteracy**

- A system is opaque to a certain stakeholder because the stakeholder does not have enough technical skills to understand relevant information about the system.
- Likely affects data subjects, decision subjects; less likely to affect software developers.

- Intentional secrecy

- System complexity

Sources of opacity

Stakeholder knowledge/abilities

- Technical illiteracy
- **Intentional secrecy**
 - A system is opaque to a certain stakeholder because a third party, usually a developer, deployer or executor, deliberately holds back information to various levels of detail.
 - Likely affects data subjects, decision subjects; sometimes operators, executors.
- System complexity

Sources of opacity

Stakeholder knowledge/abilities

- Technical illiteracy
- Intentional secrecy
- **System complexity**
 - A system is opaque to a certain stakeholder because of its characteristics (size, complexity).
 - Affects all stakeholders



Data collection

- **Non-disclosure of data collection methods** – intentional secrecy
- **Dynamic / unstructured data sources** (e.g. social media) – system complexity
- **Complexity in data cleaning and preprocessing** – system complexity / intentional secrecy

- Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in neural information processing systems*, 28.
- Driscoll, K., & Walker, S. (2014). Big data, big questions| working within a black box: Transparency in the collection and production of big twitter data. *International Journal of Communication*, 8, 20.
- Laoutaris, N. (2018). Data transparency: Concerns and prospects [point of view]. *Proceedings of the IEEE*, 106(11), 1867-1871.

In fact...

?

In fact...

I Don't Know, Is AI Also Used in Airbags?

An Empirical Study of Folk Concepts and People's Expectations of Current and Future Artificial Intelligence

Fatemeh Alizadeh , Gunnar Stevens and Margarita Esau

expectations of AI. Our results revealed that for most, AI is a dazzling concept that ranges from a simple automated device up to a full controlling agent and a self-learning superpower. We explain how these folk concepts shape users'

Public understanding

- With the public release of powerful generative tools such as ChatGPT, Midjourney, and the like, **public understanding** is more urgent than ever.
- Public misunderstanding of AI causes significant harm, and with every new technological development, **the gaps in knowledge and understanding** are growing.

Example: lawyer

Here's What Happens When Your Lawyer Uses ChatGPT

A lawyer representing a man who sued an airline relied on artificial intelligence to help prepare a court filing. It did not go well.

ChatGPT had invented everything.

Example: lawyer

the court or the airline. Mr. Schwartz said that he had never used ChatGPT, and “therefore was unaware of the possibility that its content could be false.”

He had, he told Judge Castel, even asked the program to verify that the cases were real.

It had said yes.

Public understanding

- **Knowledge is power**
- Inequalities in AI knowledge and understanding exacerbate existing social inequalities
 - As AI literacy becomes more and more important for high-paying jobs and shaping public policy, people that have access to elite education and resources have more opportunities to develop these skills, while **marginalized communities are left behind**
- **Power concentrates** around technocratic elite
- Gaps between **have and have-nots** grow

What does the opacity problem look like?

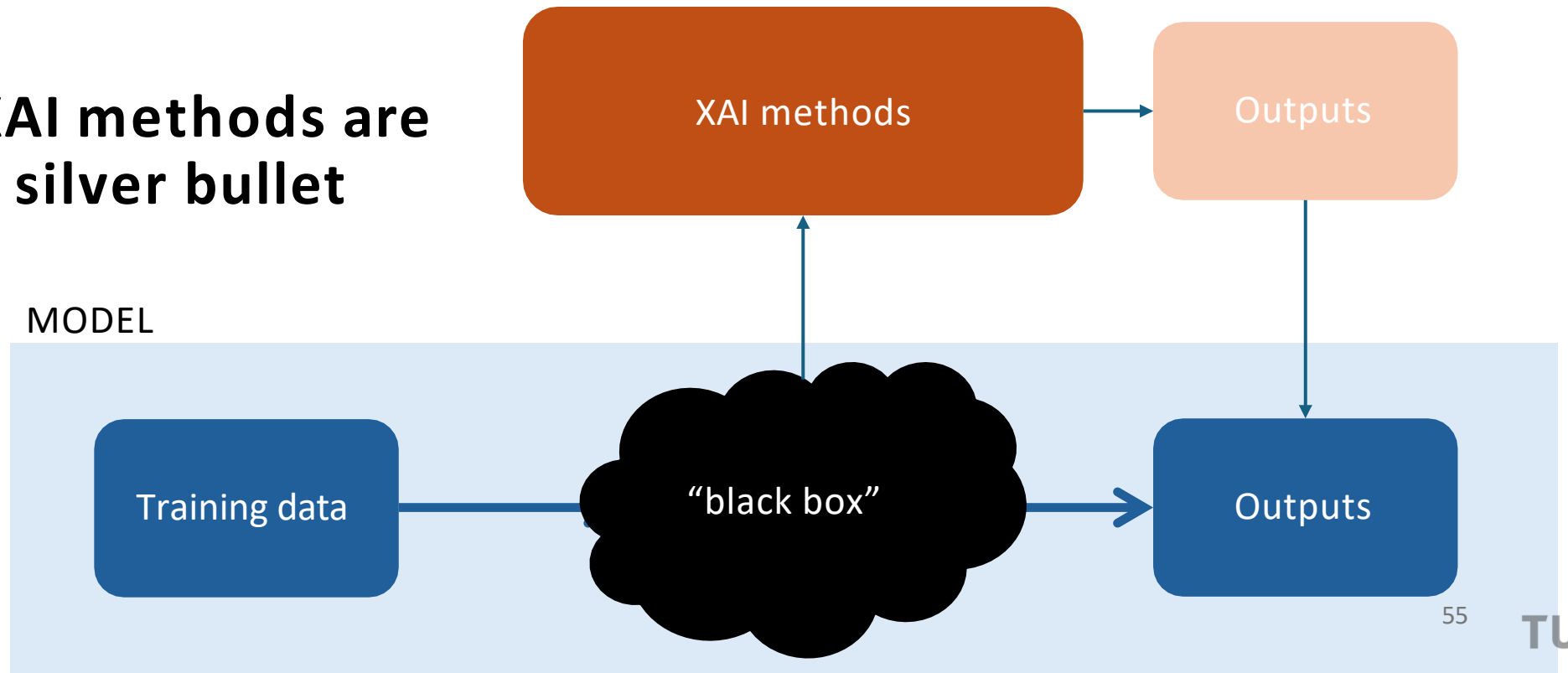
- The opacity problem is not just the “black box” system, it can occur in **the entire AI ecosystem**
- It's a **sociotechnical problem**
- **Stakeholder-dependent**
 - Different sources of opacity
 - Technical illiteracy
 - Intentional secrecy
 - System complexity
- The **general public** as a stakeholder is often worst off; lack of public AI understanding exacerbates social inequalities



**How to achieve
transparency?**

Technical solutions

- As we've seen, XAI methods are a way to alleviate the opacity of machine learning systems.
- **However, XAI methods are not (yet?) a silver bullet**



Issues with XAI methods

- **Fidelity** – post-hoc methods might not perfectly represent the black box model's true decision process, potentially leading to misleading explanations

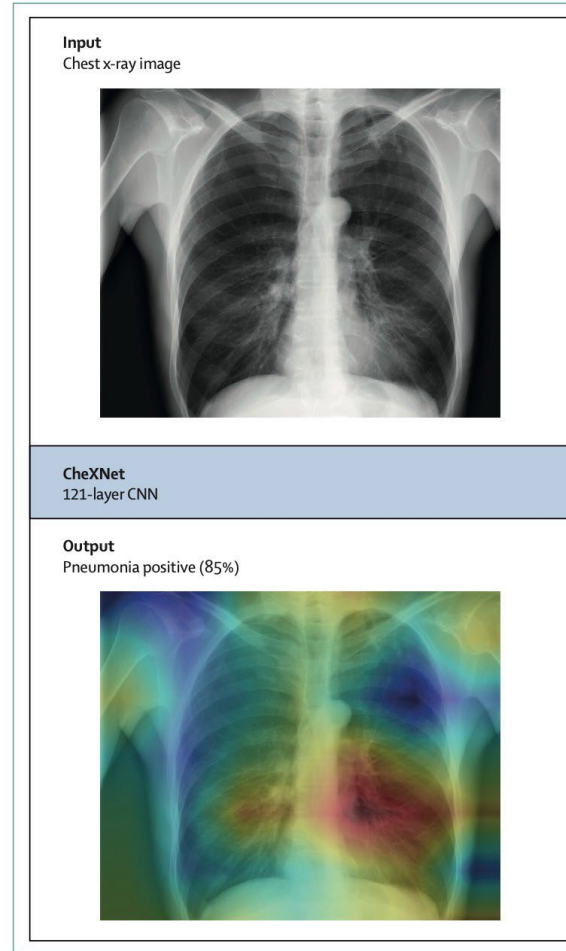
Evaluate quality / reliability of approximations:

- **Robustness** - many XAI methods are unreliable / not robust / prone to adversarial attacks
- **Inconsistency** - different methods give different explanations? → disagreement problem

- Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020, February). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 180-186).
- Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., & Lakkaraju, H. (2022). The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*.

Issues with XAI methods

- **Interpretability gap** in heatmaps
- Also **vulnerable to adversarial attacks**
- Therefore, Ghassemi et al. (2021) conclude that they shouldn't be used/trusted in the medical domain



“Even the hottest parts of the map contain both useful and non-useful information (from the perspective of a human expert), and simply localising the region does not reveal exactly what it was in that area that the model considered useful.”

Issues with XAI methods

- **Multiple valid counterfactuals** — For a given prediction, there may be multiple ways to change the input features to achieve a different outcome. Selecting the most meaningful or realistic counterfactual is not always straightforward.
- **Feasibility** — Not all counterfactuals may be actionable or realistic. For example, a counterfactual that suggests a person change their age or gender to receive a loan is not useful in practice.
- **Computational complexity** — Finding the minimal and most meaningful set of feature changes in high-dimensional data can be computationally expensive, especially for complex models.

Regulation

- A “**right to explanation**” can mandate (limited) insights into the relevant technologies
- **GDPR (Recital 71)**
 - “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.
- **AI Act (Art. 86)**
 - “Any affected person subject to a decision which is taken by the deployer on the basis of the output from a high-risk AI system (...) shall have the right to obtain from the deployer **clear and meaningful explanations** of the role of the AI system in the decision-making procedure and the main elements of the decision taken.”

Education

- **Education** in digital literacy, data science, and evidence-based reasoning may give non-expert stakeholders a greater appreciation of how their data can be used and will be used to reach data-driven decisions
- But of course, we cannot expect everyone to have the same level of access to such education.

Transparency

- **Recall why opacity is a problem**
 - if an automated decision support system is opaque, we cannot trust, challenge, improve, or intervene on its outputs and internal decision logic. It threatens control / oversight. This causes harm in high stakes contexts (injustices, unsafety, no accountability / recourse, etc).
- **Note that transparency is a means to an end**
 - We don't want it for its own sake, but because it is necessary to achieve other goals that we value intrinsically (justice; safety; understanding; accountability; harm prevention; etc.)

Conclusions / takeaways

- **Opacity in ML causes all types of harm in high stakes contexts**
 - Injustices; unsafety; lack of recourse, accountability, etc.
- **Opacity is not just a technical problem**
 - It is deeply social
 - It's much larger than just “black box systems”, occurs throughout entire ecosystem
- **Opacity is stakeholder-dependent**
 - Technical illiteracy
 - Intentional secrecy
 - System complexity
- **To increase transparency, we need:**
 - More robust XAI
 - Strong regulation
 - Public education
 - Understanding and continuous evaluation of further goals



Questions?