Explaining Decisions One Conversation at a Time: Opportunities and Risks of LLMs as Explainability Assistants

Filip Cano
Institute of Science and Technology Austria

In a nutshell

- Explainability methods are useful, but made by and for engineers
- “Right of an explanation” as a foundation of trust
  - But… What is an explanation? 📖
- LLMs can help us bridge this gap!

📖 Tim Miller, *Explanation in artificial intelligence: Insights from the social sciences*. Artificial Intelligence (vol 267). 2019

Explainable vs Interpretable AI

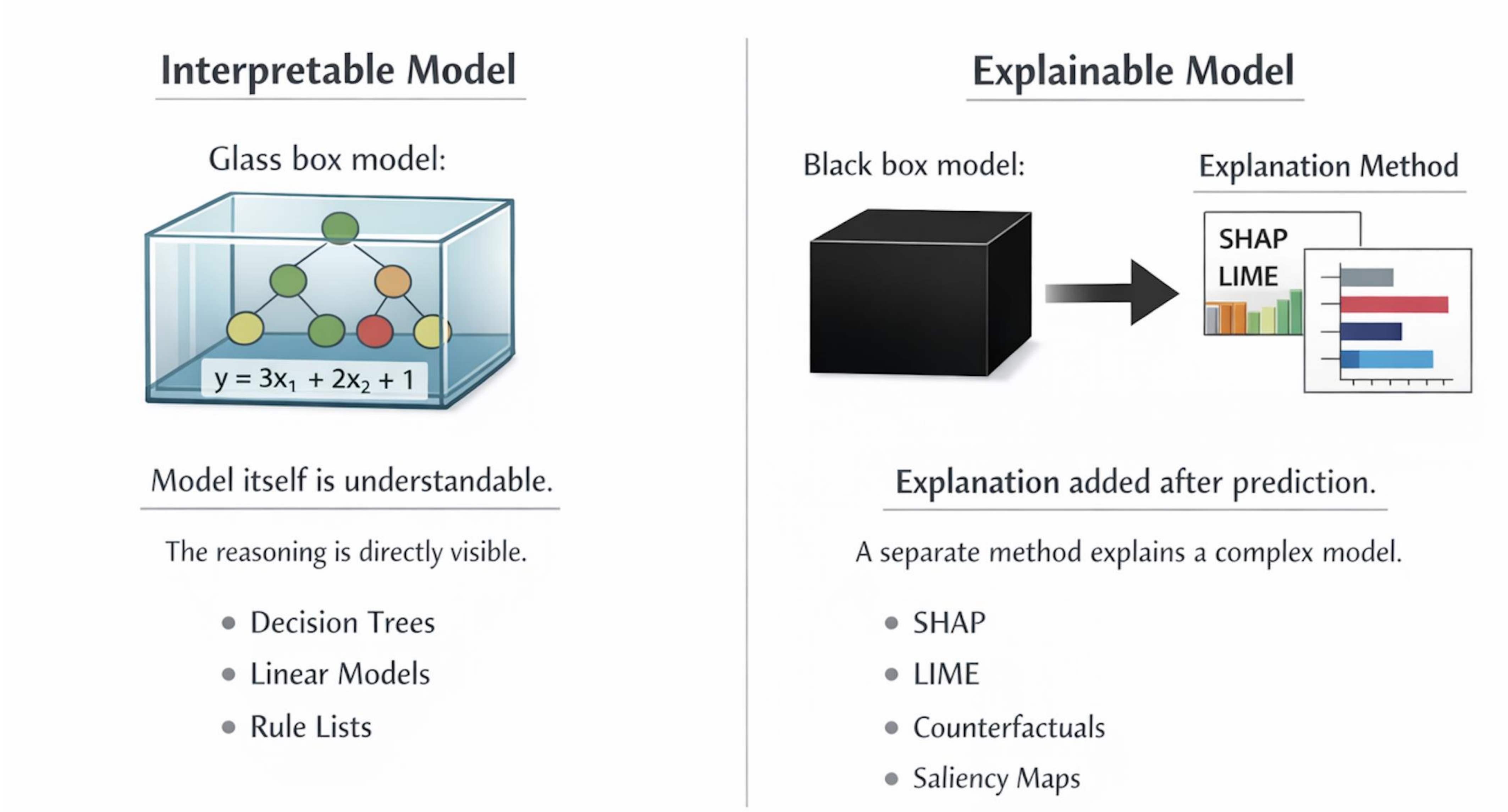


*Image generated with help of generative AI

Explainable vs Interpretable AI

Interpretable AI

- Transparency by design
- Examples: Decision trees, rule lists, linear models…
- You can directly inspect the decision logic…. If you can understand it

Explainable AI

- Post-hoc explanations of opaque methods.
- Examples: SHAP, LIME, saliency maps, counterfactuals…
- You can apply it to any method… if your explanation is faithful enough

Global vs Local Explainability

- Much of explainable AI is based on surrogate models that are interpretable
- Explanatory surrogate models can be *local* or *global*

A good explanation is a conversation

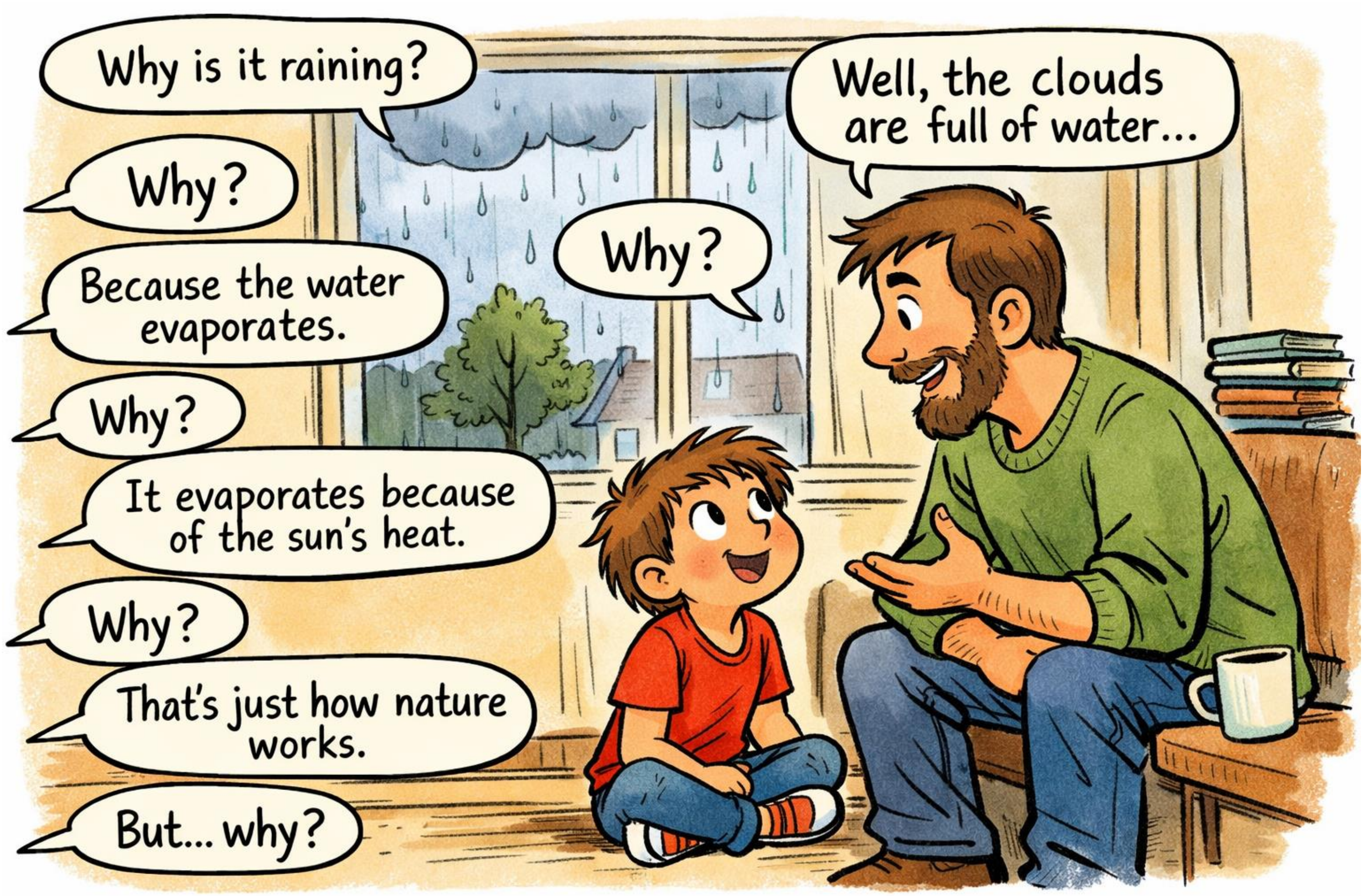
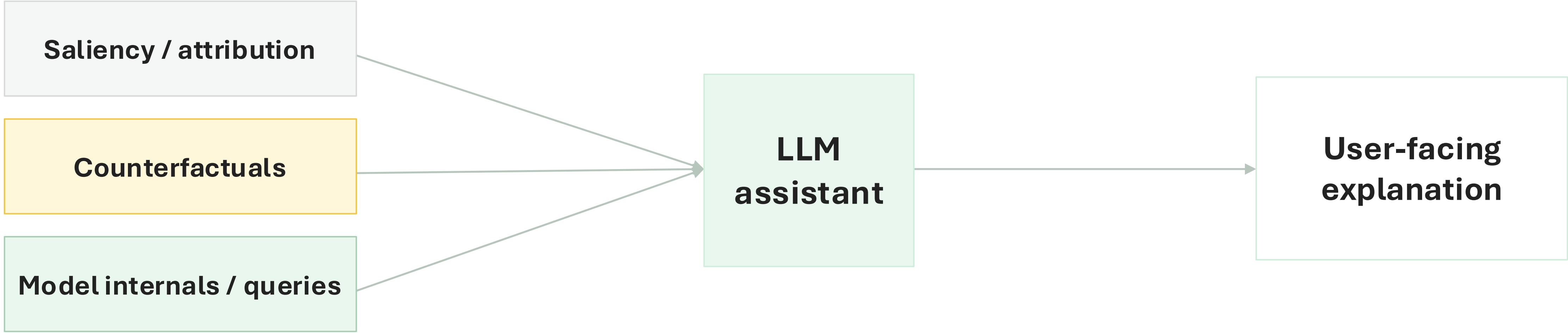


*Image generated with help of generative AI

LLMs as an Interface Between Models and Users



- translate technical signals
- adapt to user background
- support follow-up questions

The Two Roles of Explanations

AI Engineer's perspective

- Understanding the model helps with testing, validation, and debugging
- Explanation can lead to efficiency and performance improvements

User/subject's perspective

- "Right of an explanation"
- Accountability and fairness concerns
- *What do I need to change to obtain a different decision?*

The Danger: Explanation that Sound Just Right

Hallucination

Fluent but fabricated statements can be mistaken for faithful explanations.

Conflict avoidance

A polite assistant may suppress ambiguity or disagreement to keep the conversation smooth.

Oversimplification

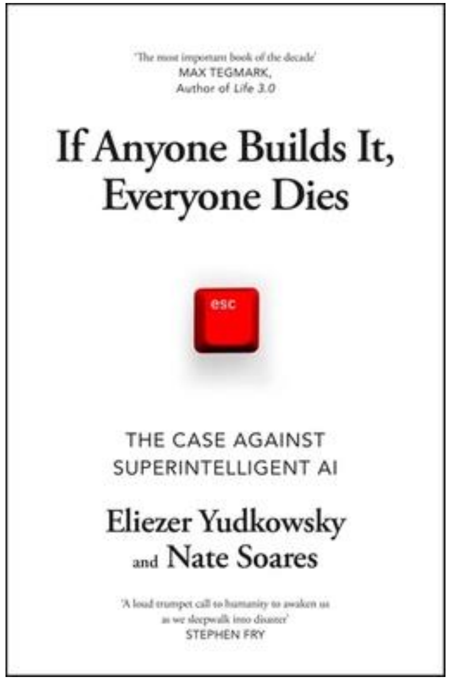
Important interactions and edge cases can disappear inside a tidy narrative.

Can We Verify Explanations?

- Anchor explanations to real model signals
- Test changes before providing counterfactual explanations
- Compare with database of know failures
- Test faithfulness of surrogate models

The Risk of Losing Build Public Trust

- Public trust in technology builds slowly and is destroyed quickly
- Cautionary tale: COMPAS 📖
- Current perception of LLMs is mixed
  - Comparison of current and future AI with nuclear weapons



📖 J. Angwin et al. *Machine Bias*. ProPublica. 2016

Main Uses of LLMs for Explanations

1. Assistive interpretation

• summarize model behavior
• interpret saliency / attribution outputs
• translate counterfactuals into plain language
• act as a conversational layer on top of existing XAI tools

2. Direct rationale generation

• generate justifications for their own predictions
• generate justifications for another model's prediction
• sound coherent and contextual
• but may fail to reflect the real reasoning process

Some Best Practices

1. Ground explanations in **verifiable model behavior**
2. Clear communication of **uncertainties and limitations**
3. **Adapt** the explanation to the **audience**
4. Use **multiple methods** to corroborate claims
5. Keep the whole explanation pipeline **transparent and auditable**

The Road Ahead

Hybrid systems

LLMs could connect more directly to model internals so faithfulness can be checked dynamically.

→

Better evaluation

Metrics based on fidelity and user impact.

→

Societal deployment

Stronger ethics guardrails and failure containment. Build-up of public trust

Conclusion

- LLMs can make **explanations** more **interactive** explanations for interpretable and explainable AI methods
- Main risks:
  - Explanations that sound right but are **not grounded** on real behaviour
  - Once you lose **public trust**, it's difficult to gain it back