Counterfactual Explanations and Algorithmic Recourse for Trustworthy AI

Research Seminar

Patrick Altmeyer

Delft University of Technology

2026-03-11

The Ground Truth (Reality)

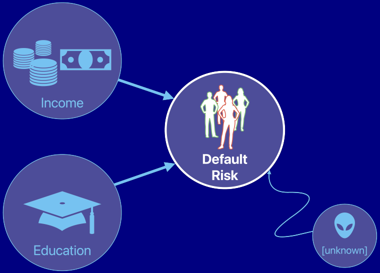

Figure 1: Predictors of default risk.

The Ground Truth (Reality)

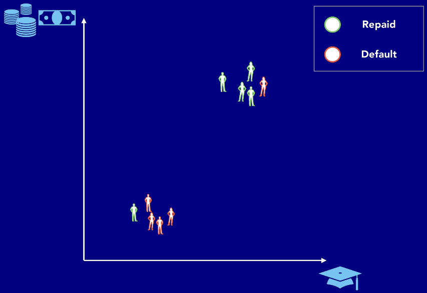

Figure 2: Ground truth outcomes across two predictors.

Black-Box AI

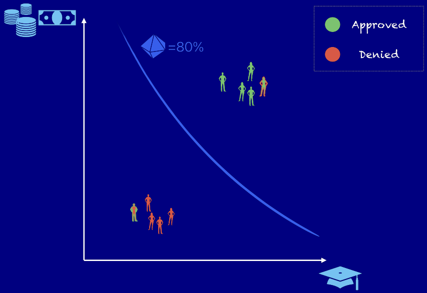

Figure 3: Classifier predicts correctly 8 out of 10 times.

Black-Box AI

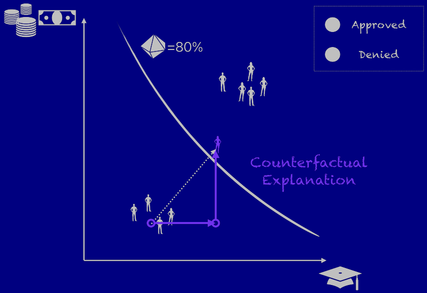

Figure 4: Simple counterfactual explanation for the black-box AI. JuliaCon Proceedings 2023

Black-Box AI

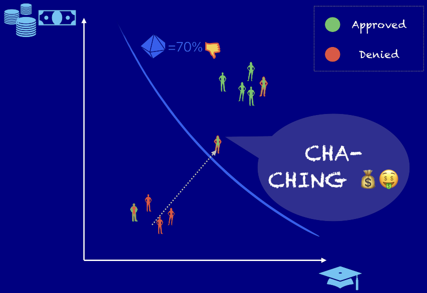

Figure 5: One happy recourse recipient, many losers. IEEE SaTML 2023

Black-Box AI

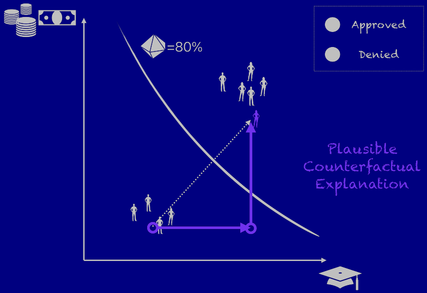

Figure 6: Plausible counterfactual explanations for the black-box AI. IEEE SaTML 2023

Black-Box AI

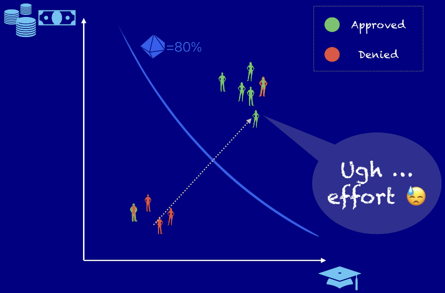

Figure 7: One somewhat happy recourse recipient, no losers. IEEE SaTML 2023

Big, Beautiful Black-Box AI

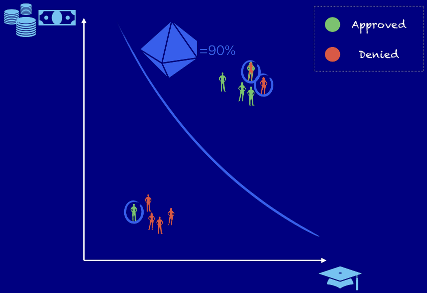

Figure 8: Classifier predicts correctly 9 out of 10 times. But … AAAI 2024

Big, Beautiful Black-Box AI

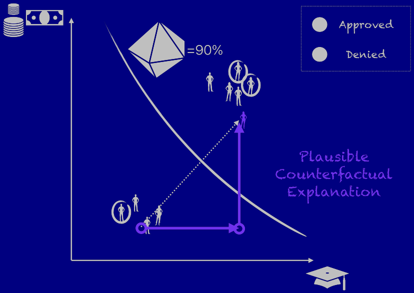

Figure 9: Plausible counterfactual explanations remains valid. Happy days?
AAAI 2024

Big, Beautiful Black-Box AI

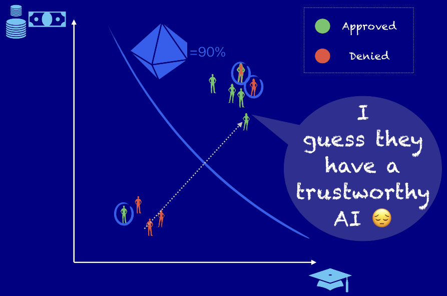

Figure 10: White-washed black-box: plausible CE hides bias. AAAI 2024

Holding Models Accountable

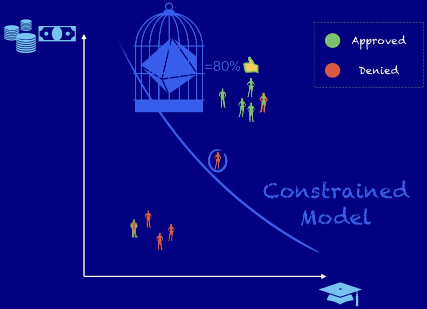

Figure 11: A model trained to use plausible explanations for predictions. IEEE SaTML 2026

‘ok but agi bruh’

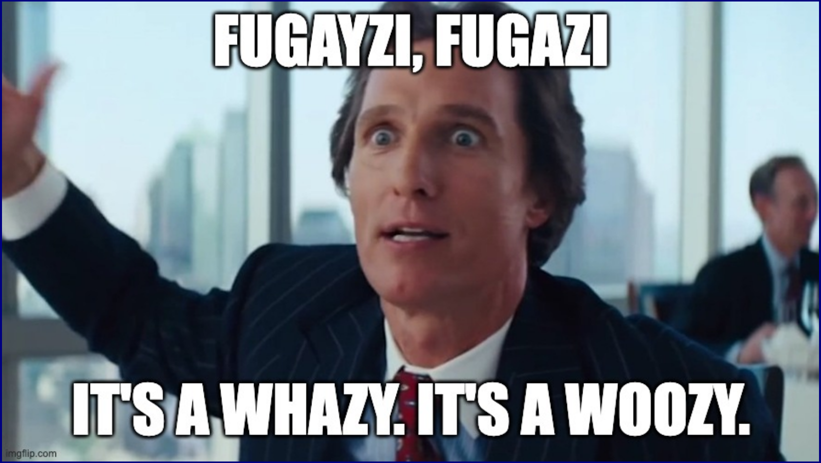

Figure 12: My personal take on “AGI by 2027”. ICML 2024

In all seriousness ...

- Useful? Absolutely

In all seriousness ...

- Useful? Absolutely
- AGI? Sentient? Conscious? No: ‘emergence’ in complex systems does not hint at any of this

In all seriousness ...

- Emergence broadly described as broad behaviour of complex systems that's different from its constituent parts:

In all seriousness ...

- Emergence broadly described as broad behaviour of complex systems that's different from its constituent parts:
    - Example 1: asset price bubbles in financial markets -> locally predictable, rational behaviour, but also market failure

In all seriousness ...

- Emergence broadly described as broad behaviour of complex systems that's different from its constituent parts:
  - Example 1: asset price bubbles in financial markets -> locally predictable, rational behaviour, but also market failure
  - Example 2: tornado -> just dust and debris, but also a possible disaster

In all seriousness ...

- Emergence broadly described as broad behaviour of complex systems that's different from its constituent parts:
    - Example 1: asset price bubbles in financial markets -> locally predictable, rational behaviour, but also market failure
    - Example 2: tornado -> just dust and debris, but also a possible disaster
- Does it matter?