

# Holding AI Accountable

Short answers to 'What are you actually doing in your *Ph.D.?*'

**Patrick Altmeyer    Arie van Deursen    Cynthia C. S. Liem**

Delft University of Technology

2026-02-25

# Toy Problem

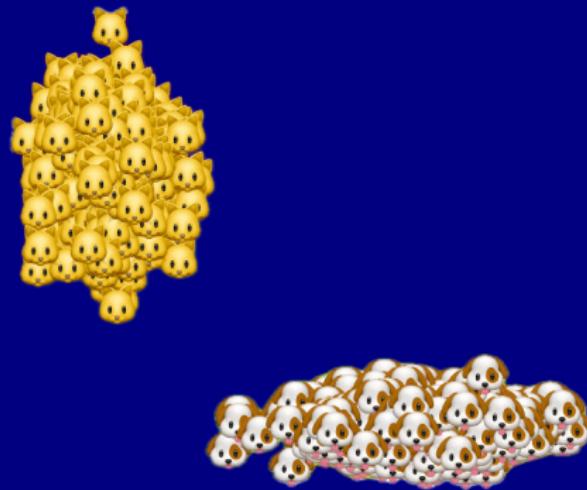


Figure 1: Cat or dog? A classification problem.

# Toy Problem

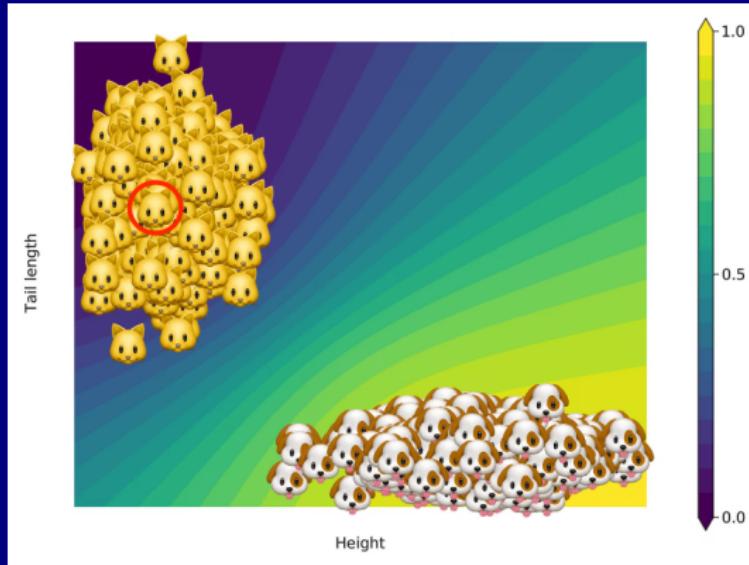


Figure 2: Solving the problem using AI, but model is opaque.

# Toy Problem

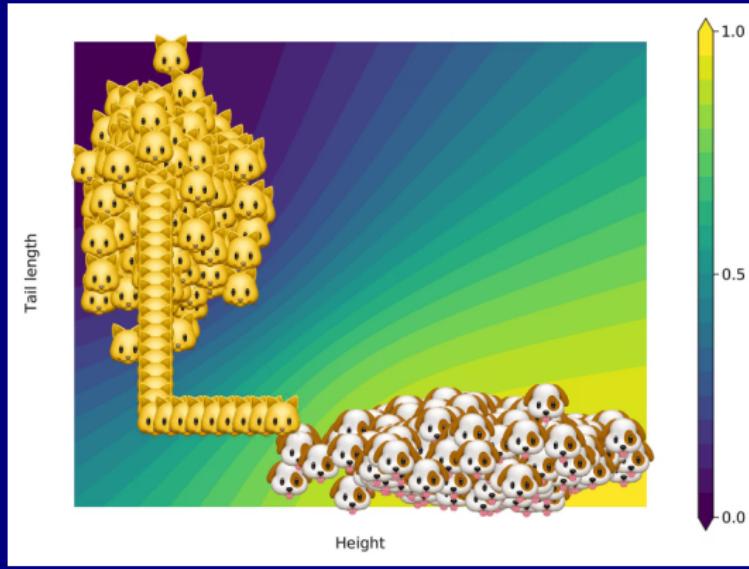


Figure 3

# The Ground Truth

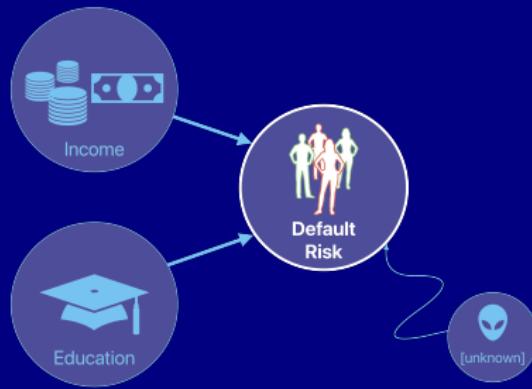


Figure 4: Predictors of default risk.

# The Ground Truth

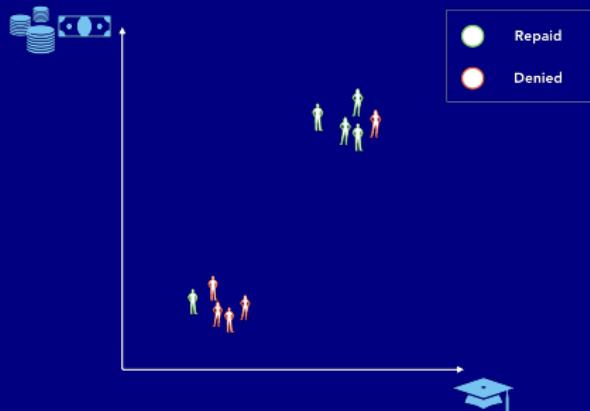


Figure 5: Ground truth outcomes across two predictors.

# Black-Box AI

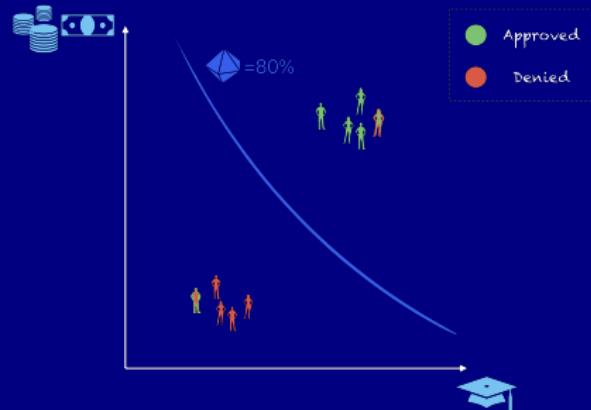


Figure 6: Classifier predicts correctly 8 out of 10 times.

# Black-Box AI

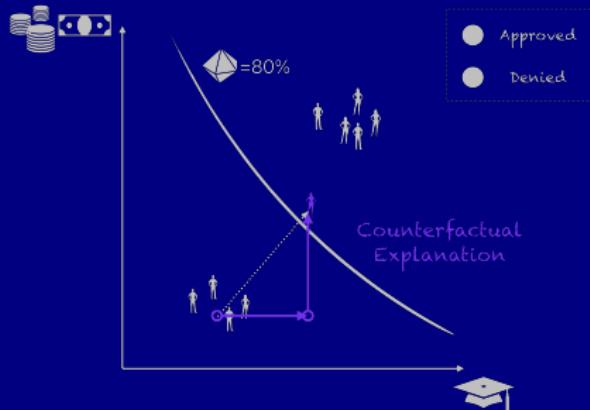


Figure 7: Simple counterfactual explanation for the black-box AI.

# Black-Box AI



Figure 8: One happy recourse recipient, many losers.

# Black-Box AI

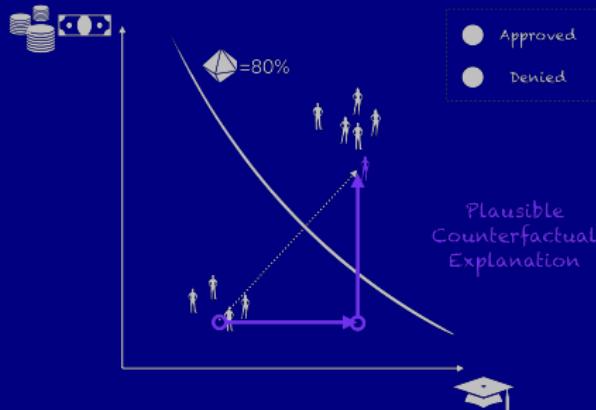


Figure 9: Plausible counterfactual explanations for the black-box AI.

# Black-Box AI

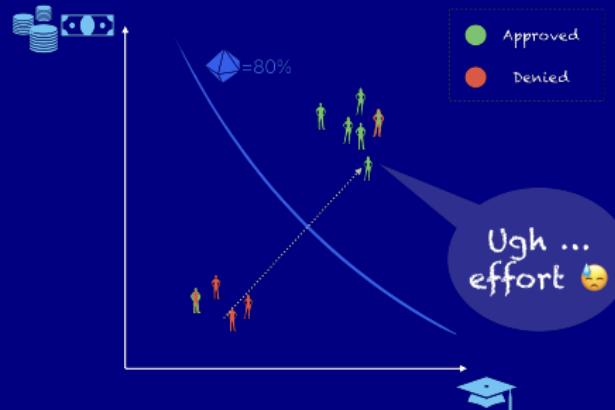


Figure 10: One somewhat happy recourse recipient, no losers.

# Big, Beautiful Black-Box AI

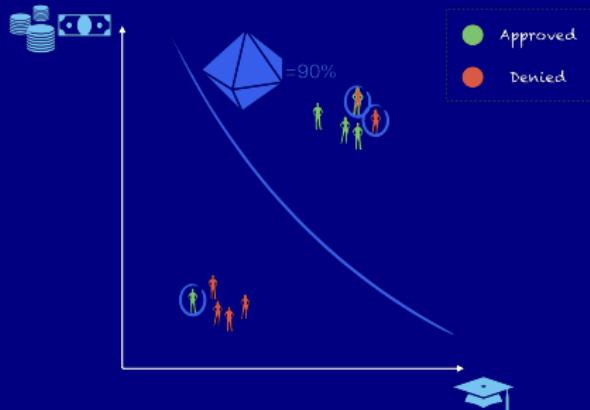


Figure 11: Classifier predicts correctly 9 out of 10 times. But ...

# Big, Beautiful Black-Box AI

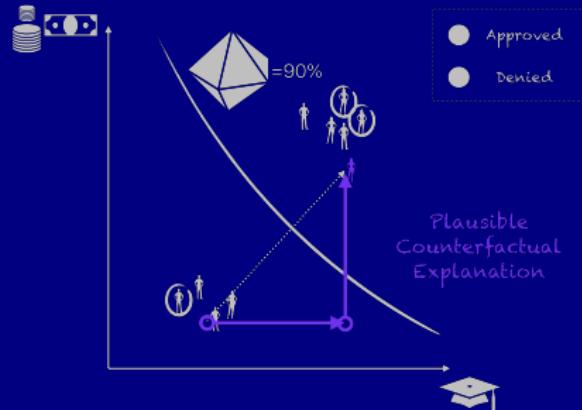


Figure 12: Plausible counterfactual explanations remains valid. Happy days?

# Big, Beautiful Black-Box AI

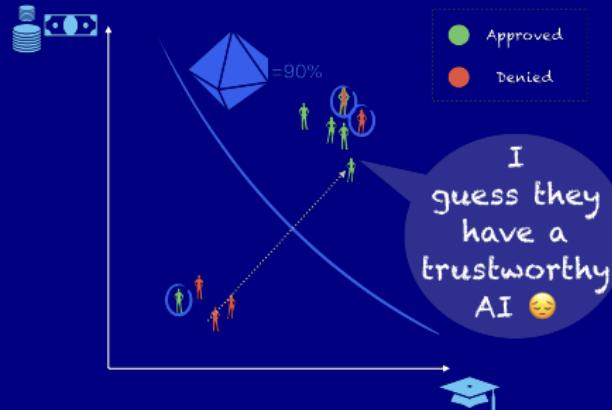


Figure 13: White-washed black-box: plausible CE hides bias.

# Holding Models Accountable

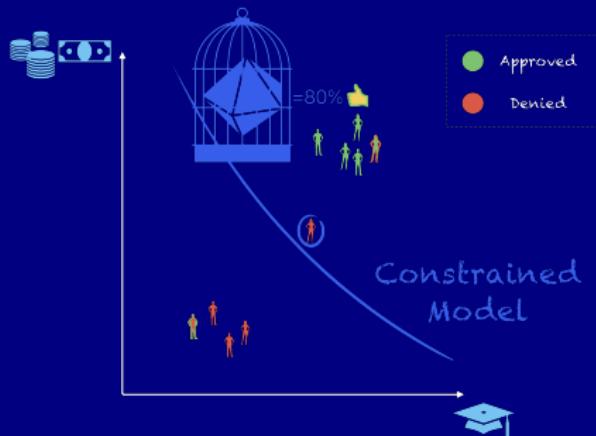


Figure 14: A model trained to use plausible explanations for predictions

'ok but agi bruh'

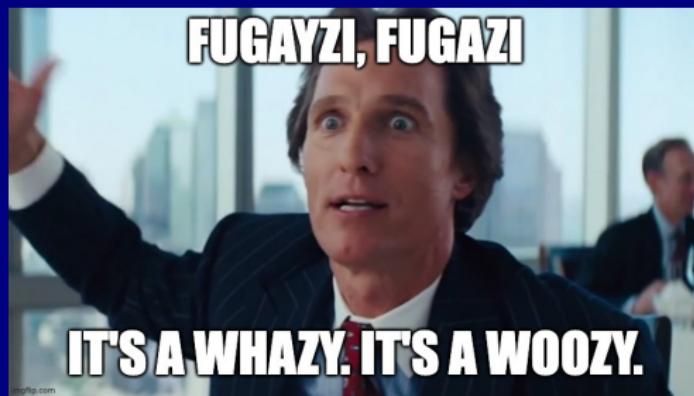


Figure 15: My personal take on “AGI by 2027”.

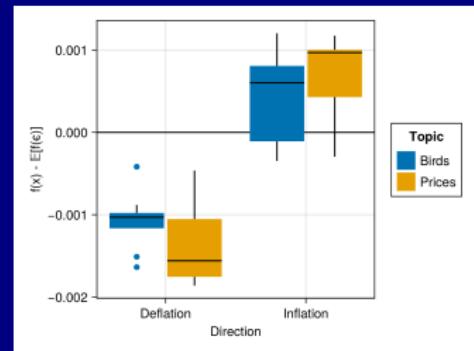


Figure 16: Birds or prices? It doesn't matter!