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

The Ground Truth (Reality)

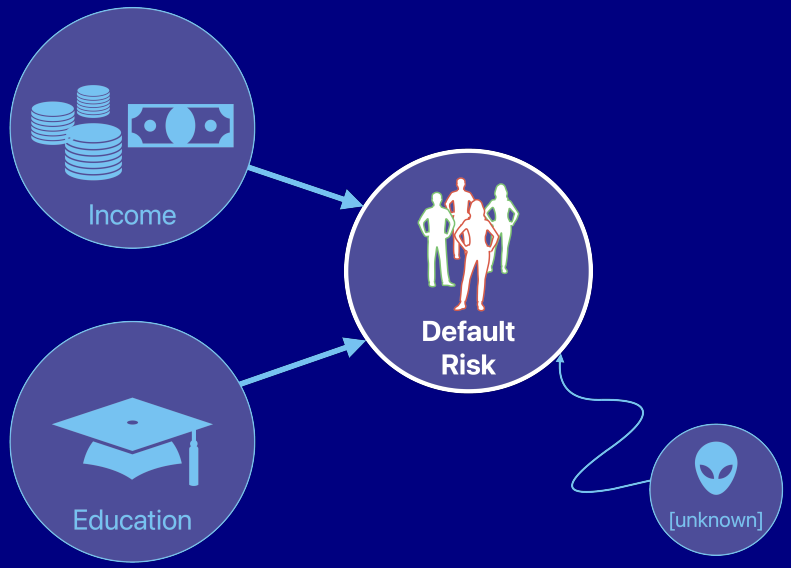

Figure 1: Predictors of default risk.

The Ground Truth (Reality)

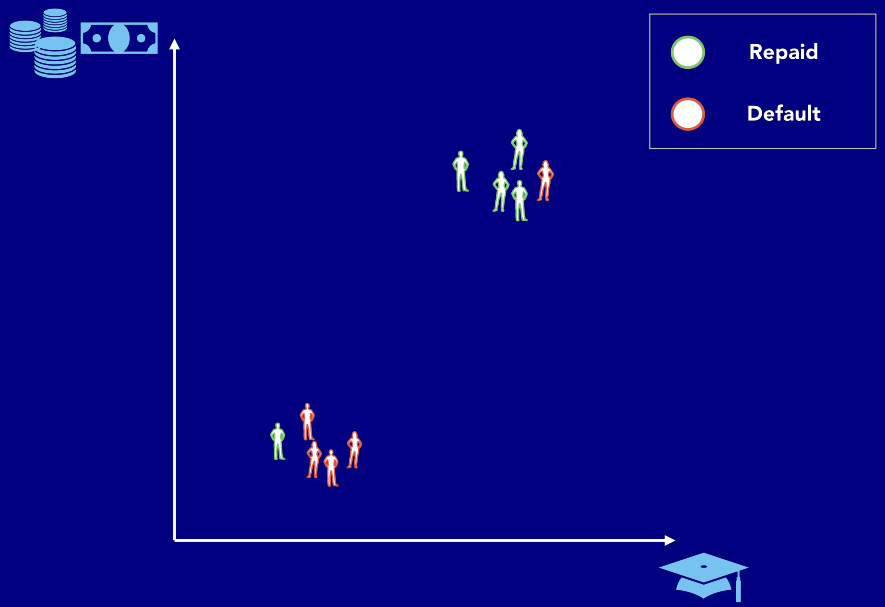

Figure 2: Ground truth outcomes across two predictors.

Black-Box AI

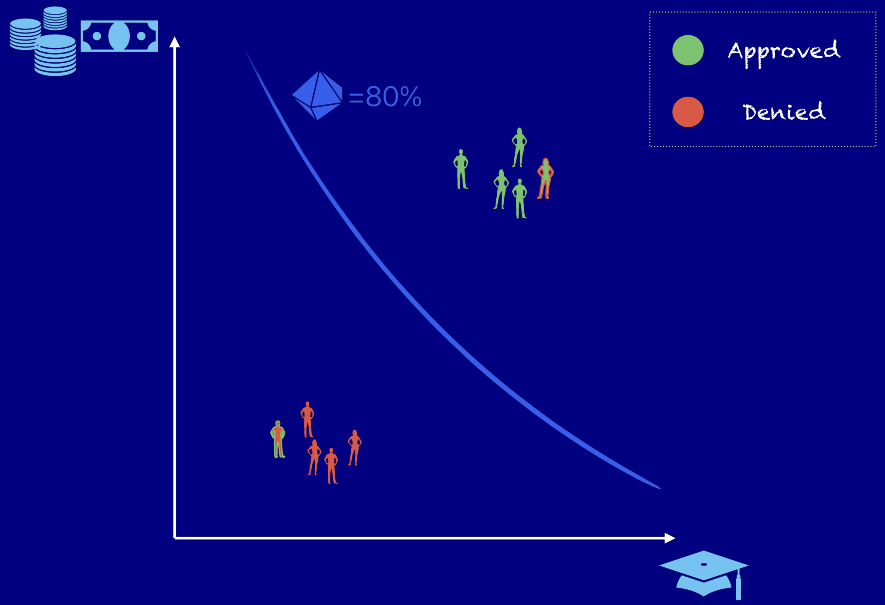

Figure 3: Classifier predicts correctly 8 out of 10 times.

Black-Box AI

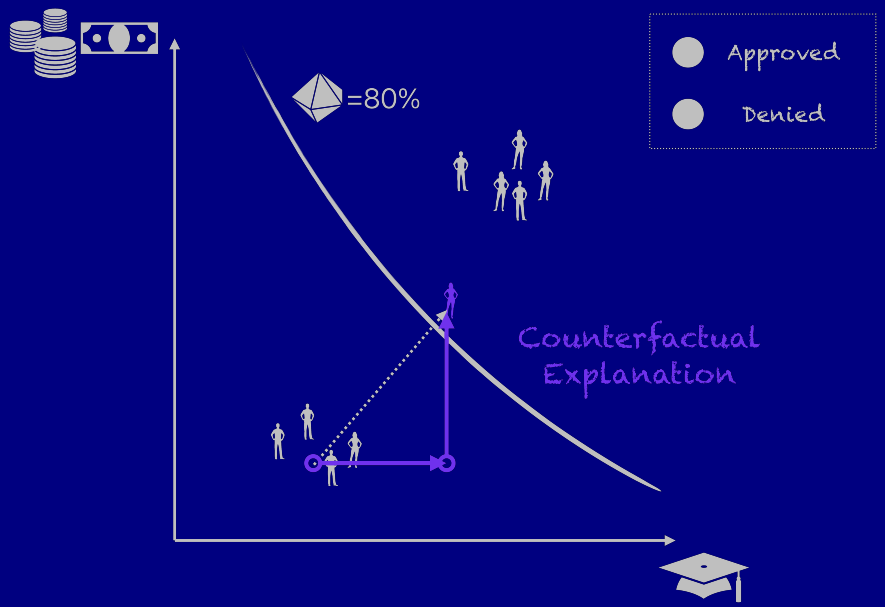

Figure 4: Simple counterfactual explanation for the black-box AI. (Chapter 2)

Black-Box AI

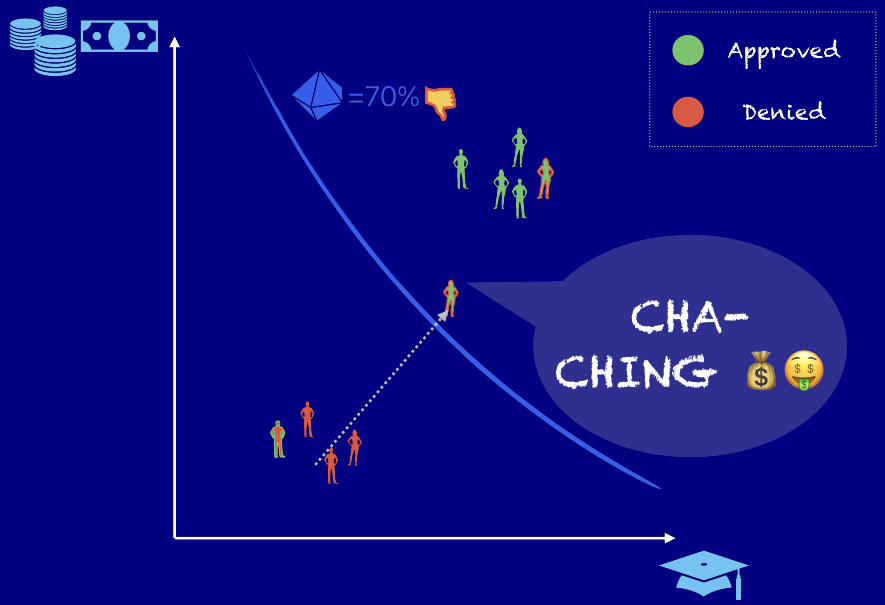

Figure 5: One happy recourse recipient, many losers. (Chapter 3)

Black-Box AI

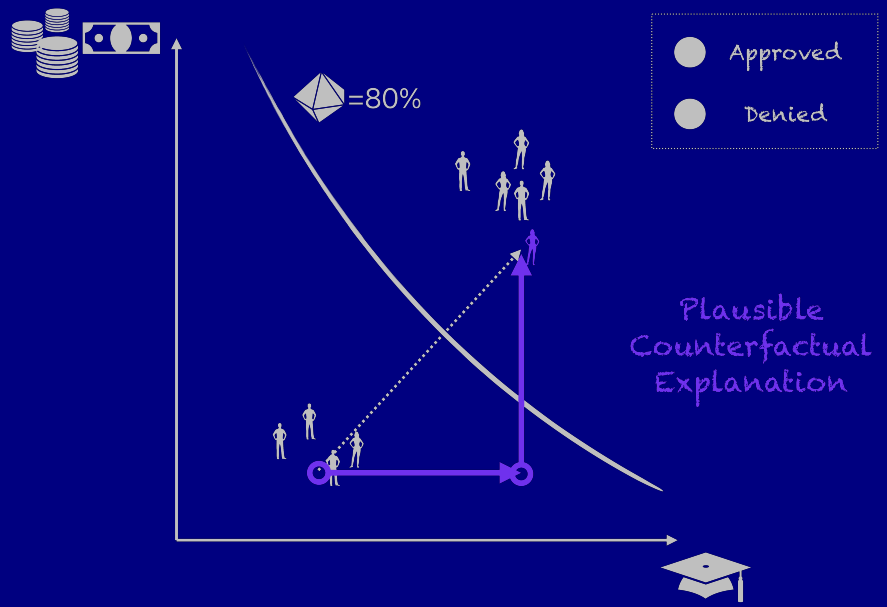

Figure 6: Plausible counterfactual explanations for the black-box AI. (Chapter 3)

Black-Box AI

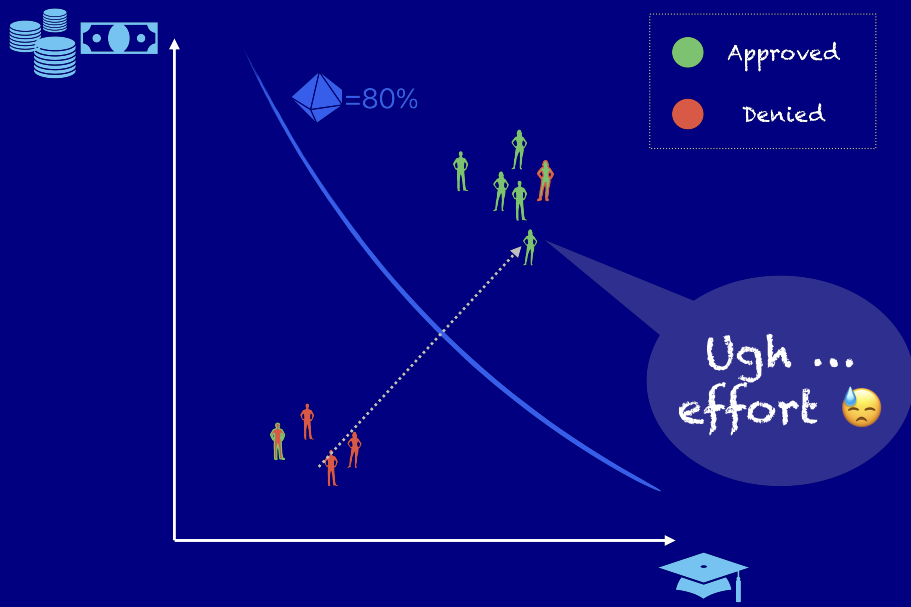

Figure 7: One somewhat happy recourse recipient, no losers. (Chapter 3)

Big, Beautiful Black-Box AI

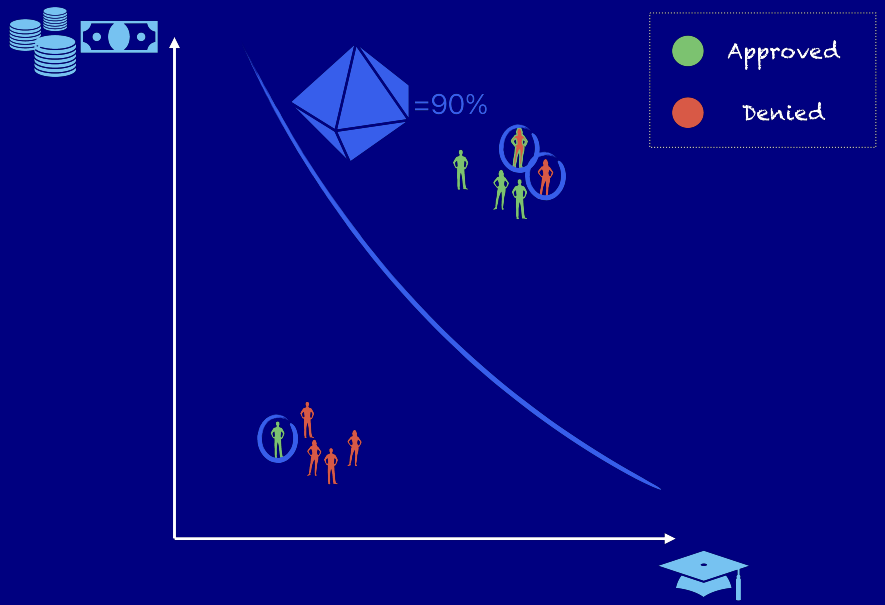

Figure 8: Classifier predicts correctly 9 out of 10 times. But ... (Chapter 4)

Big, Beautiful Black-Box AI

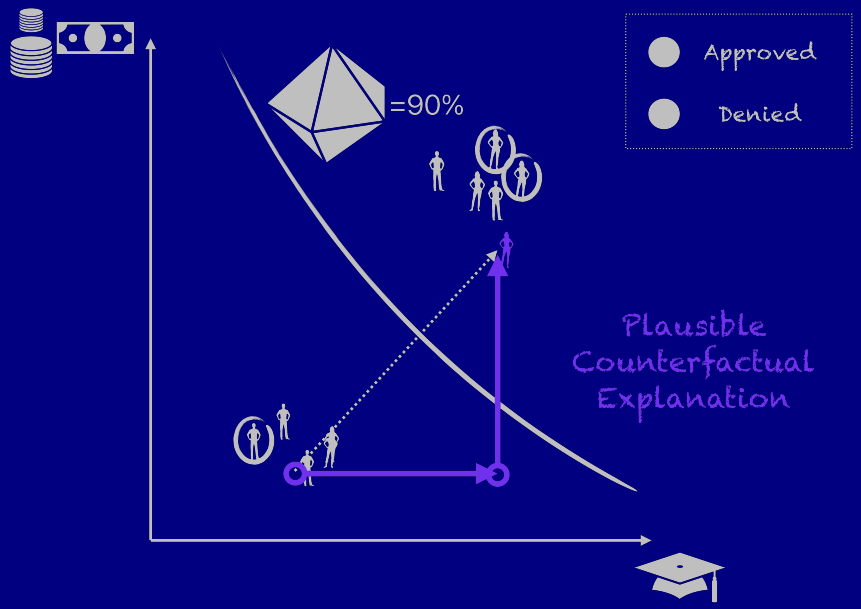

Figure 9: Plausible counterfactual explanations remains valid. Happy days? (Chapter 4)

Big, Beautiful Black-Box AI

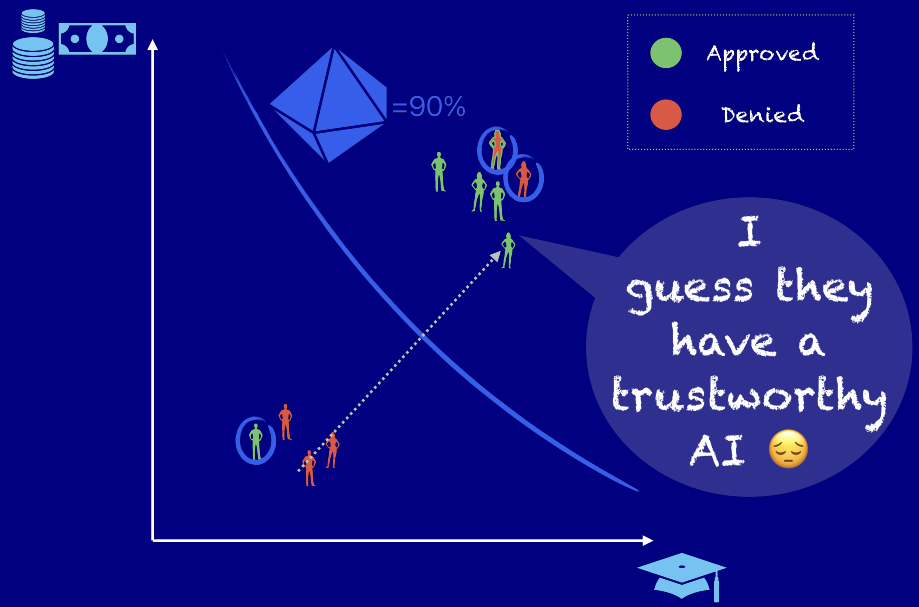

Figure 10: White-washed black-box: plausible CE hides bias. (Chapter 4)

Holding Models Accountable

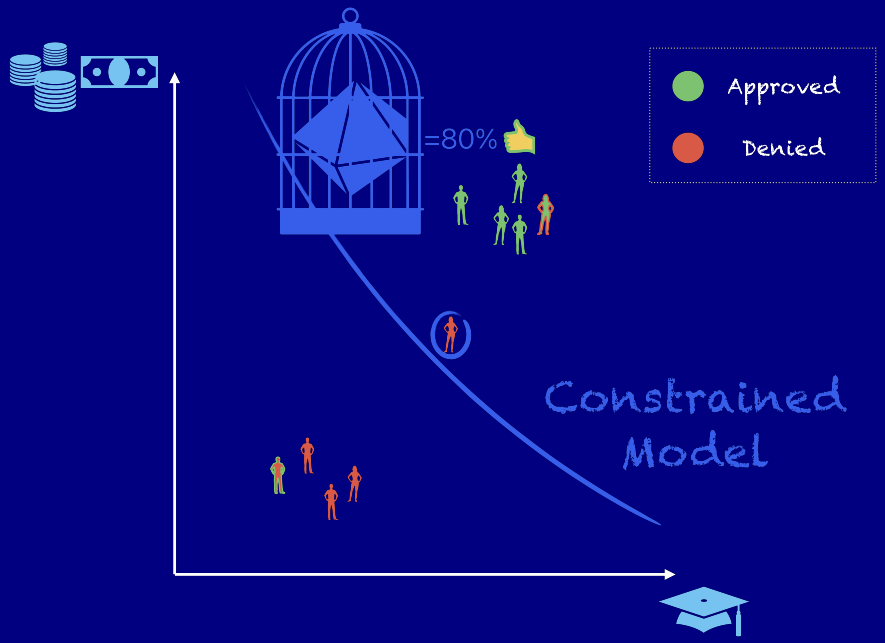

Figure 11: A model trained to use plausible explanations for predictions. (Chapter 5)

‘ok but agi bruh’

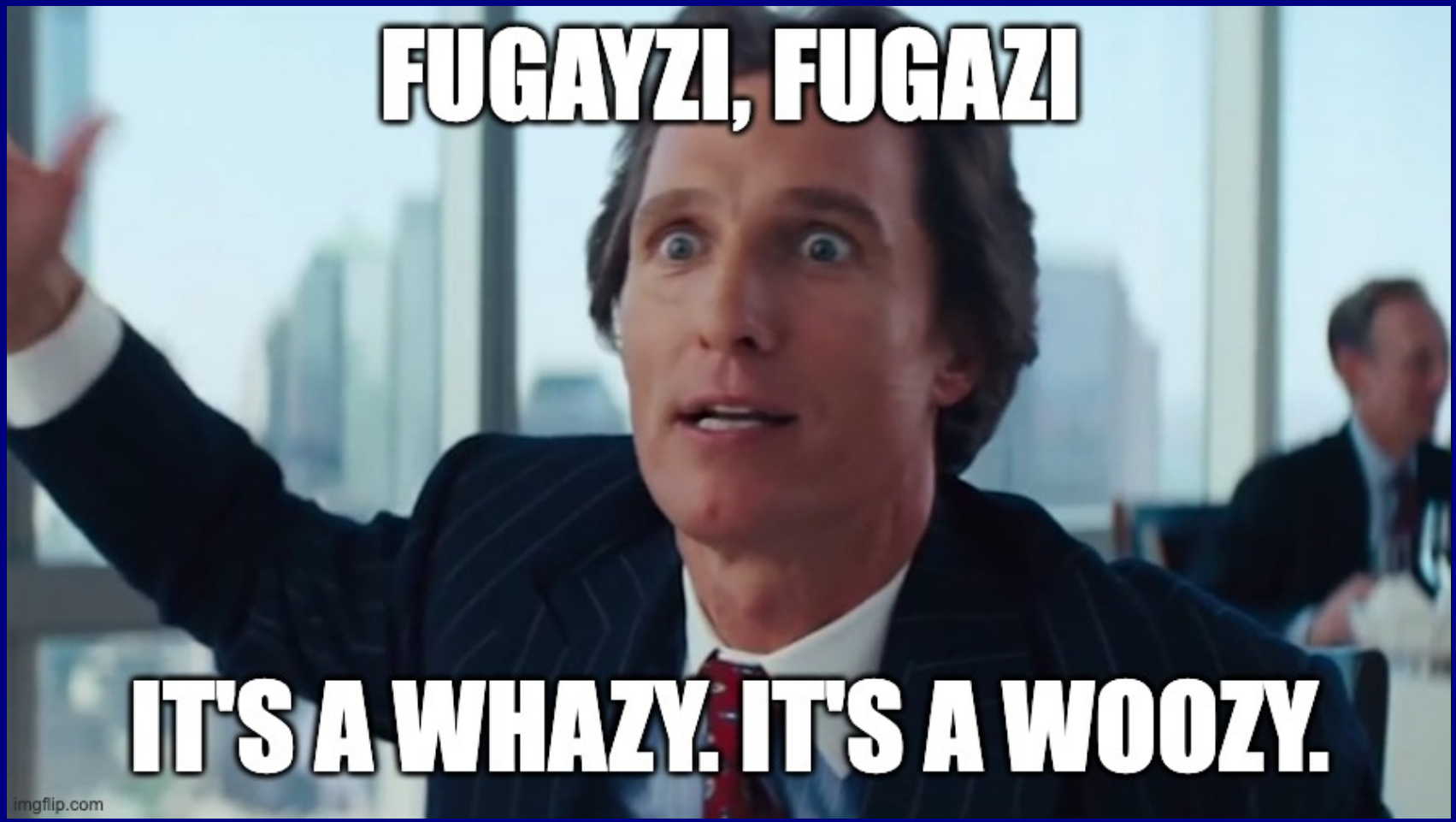

Figure 12: My personal take on “AGI by 2027”. (Chapter 6)