Leveraging Contextual Counterfactuals Toward Belief Calibration

Qiuyi (Richard) Zhang[1] Michael S. Lee[2] Sherol Chen[1]

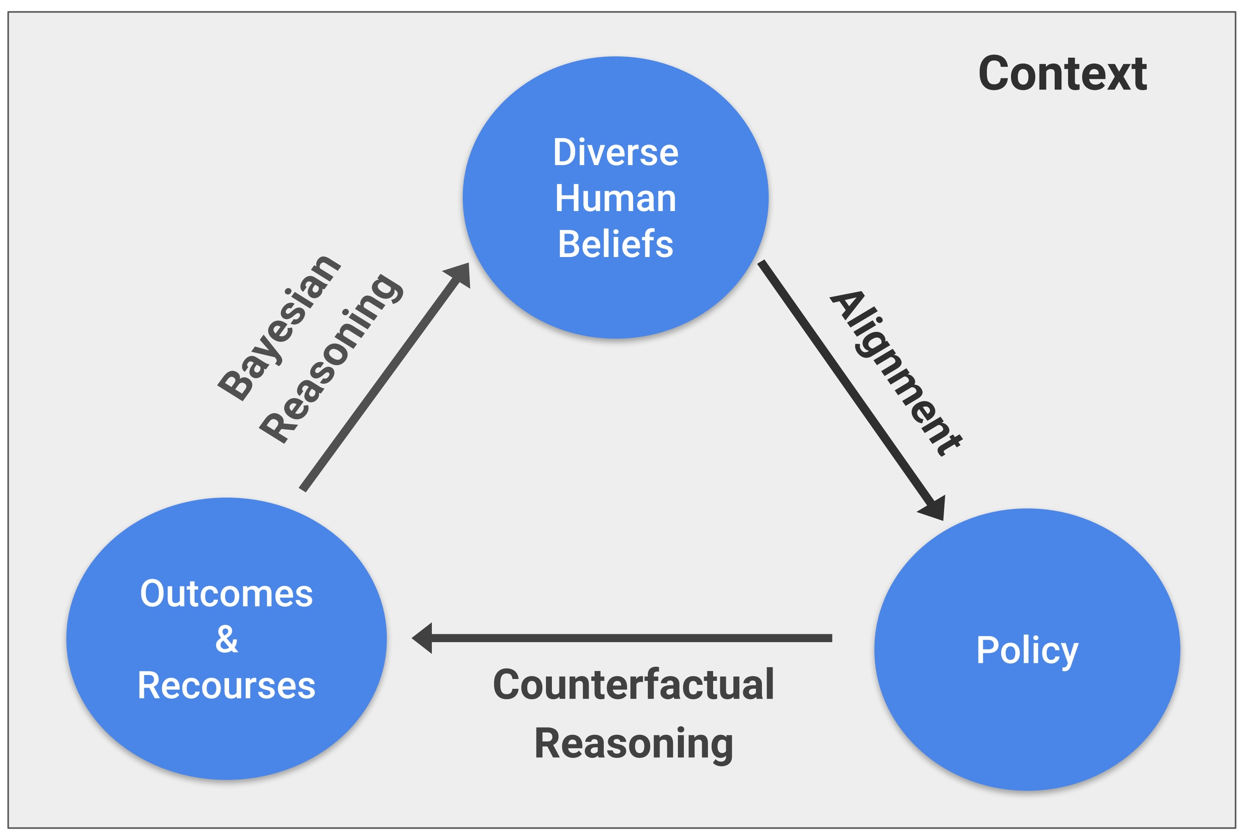
[1]Google Deepmind, [2]Carnegie Mellon University

Overview

Alignment between AI and humans is a difficult problem due to diversity of human beliefs. We identify the **meta-alignment problem** – even if a set of "alignment" beliefs are identified, how should the model calibrate the strength of each belief for beneficial societal impact?

We argue that **counterfactual reasoning** over possible outcomes and recourses are key to **identifying optimal belief strengths** that can generalize to different contexts.

We explore these ideas on credit default classification, and find **surprising results** through counterfactual analysis, such as increased leniency gives higher predictive and social alignment.



Human Studies on Counterfactuals

Counterfactual reasoning can shape beliefs prior to decision making through anticipated regret...

Attitude to Speeding by Type of Video Seen

	Type of video				
	Normative belief	Behavioral belief	Perceived behavioral control	Anticipated regret	Control
M	21.58	20.88	21.47	23.49	19.73
(*SD*)	(6.02)	(4.95)	(6.13)	(6.49)	(5.95)

Only videos priming participants of anticipated regret led to statistically significant changes in beliefs in speeding[3] (higher value → more negative attitudes).

...and can influence decision making itself through anticipated consequences.

People significantly differ in labeling toward a factual description (e.g. a dog looks aggressive) vs toward a normative judgment (e.g. a dog looks aggressive, and therefore violate an apartment's policy).[4]

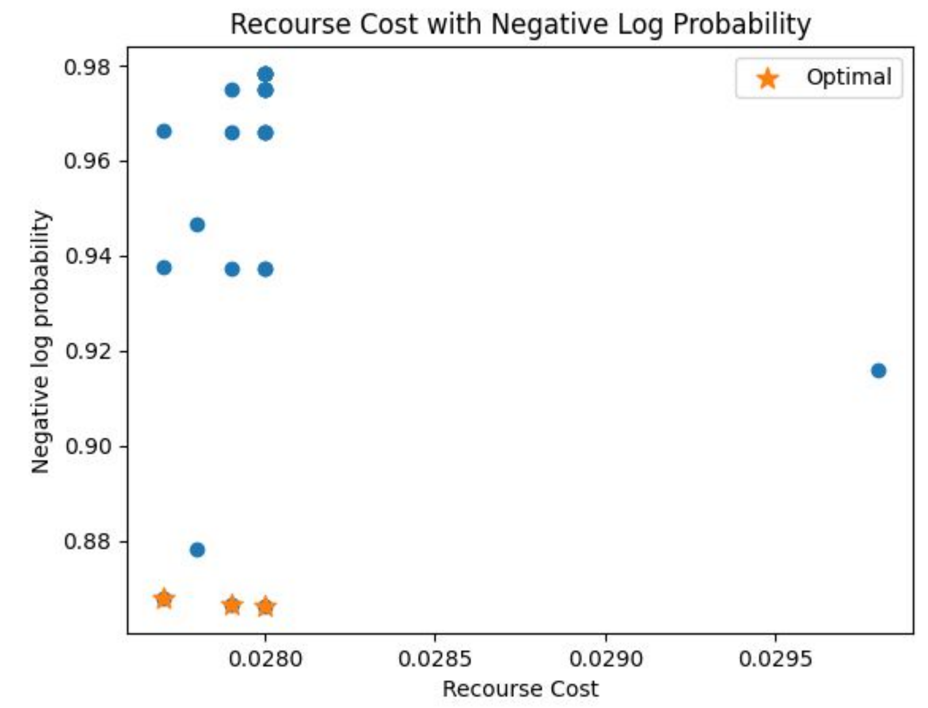
"Getting a decision wrong factually is just a matter of describing the world incorrectly. Getting it wrong normatively is a matter of potentially doing harm to another human."

Experimental results

Calibrating the strength of beliefs via contextual counterfactuals on on σ (noise in credit decisions), λ (feature regularization) $\in [0.001, 0.01, 0.1, 1, 10]$ during Bayesian linear regression **credit default classification**[5].

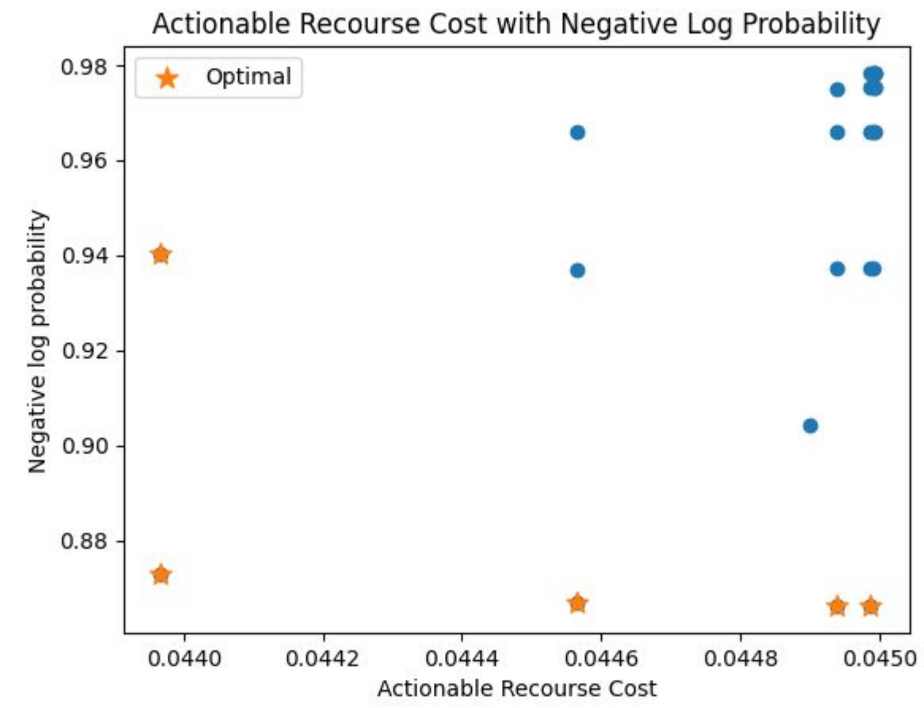
Context 1: All features are actionable in recourse calculation

σ	λ	Recourse Cost	Log-Prob
10	0.1	0.0277	0.878
10	1	0.0279	0.865
10	10	0.028	0.866



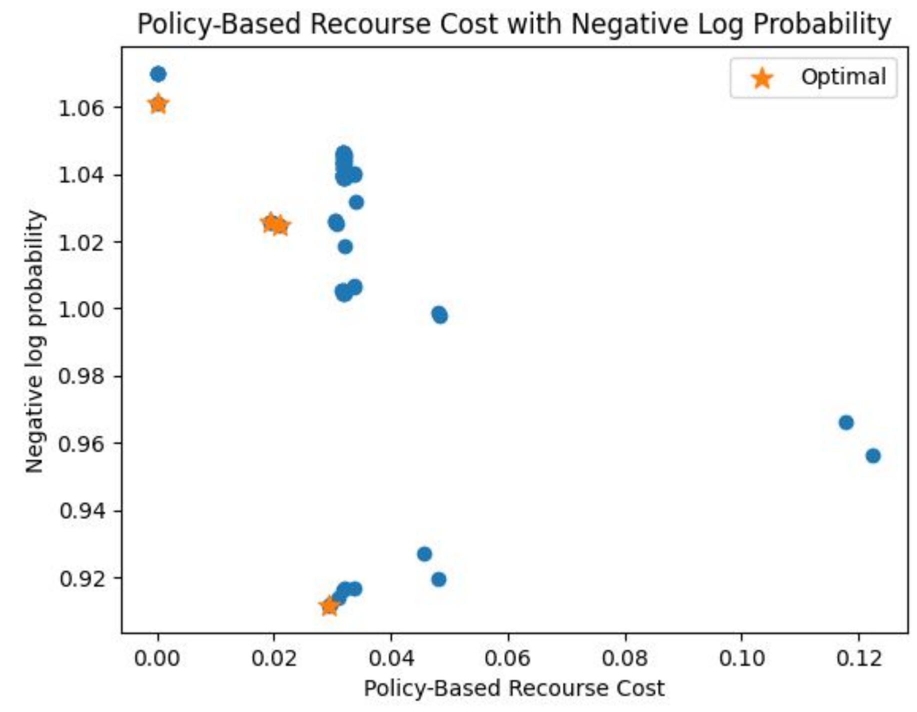
Context 2: Higher recourse cost for non-actionable features (e.g. age)

σ	λ	Actionable Cost	Log-Prob
1	0.001	0.0439	0.94
10	0.01	0.044	0.878
10	0.1	0.0446	0.877
10	1	0.0449	0.866
10	10	0.045	0.866



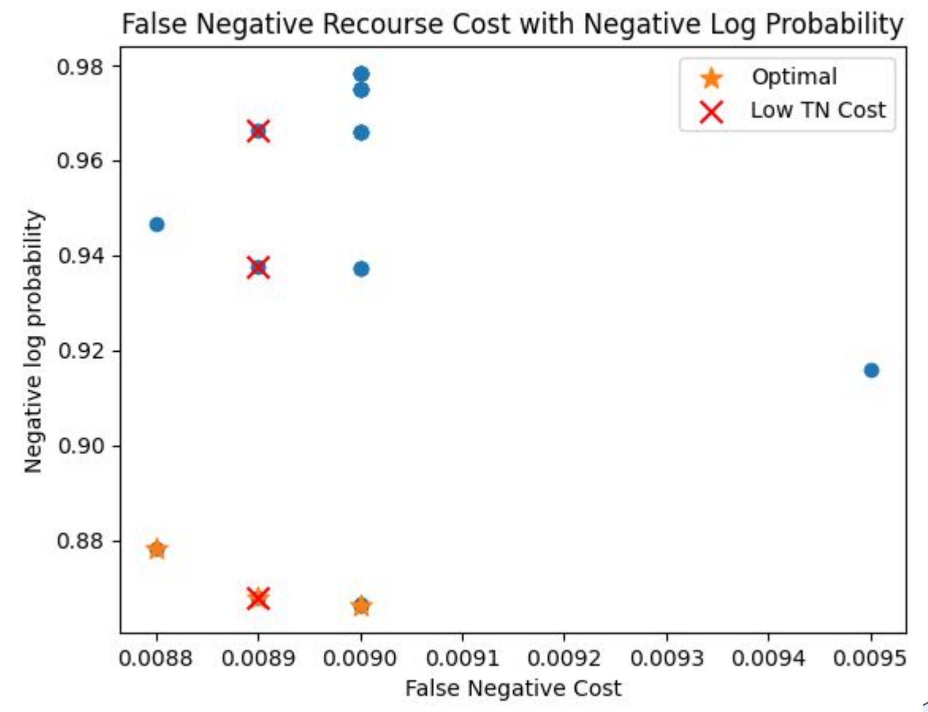
Context 3: Policy calibration (e.g. adding 'benefit of the doubt' via $\beta\sigma$)

σ	λ	β	Policy Cost	Log-Prob
10	0.001	10	0.0	1.06
1	10	10	0.019	1.025
1	0.01	10	0.021	1.024
10	10	1.0	0.029	0.911



Context 4: Balancing recourse costs among false and true negatives. To mitigate risk, we institute a high recourse cost for true negatives.

σ	λ	FN Cost	TN Cost	Log-Prob
10	0.01	0.0088	0.0189	0.878
10	10	0.009	0.0189	0.866



Conclusions

We explore the meta-alignment problem of calibrating the strength of beliefs on noise and feature regularization to different contexts in credit default classification, analyzing the subsequent distribution of outcomes and recourses to select the optimal belief strengths.

[3] D. Parker, S. G. Stradling, and A. S. Manstead. Modifying beliefs and attitudes to exceeding the speed limit: an intervention study based on the theory of planned behavior. Journal of Applied Social Psychology, 1996.
[4] A. Balagopalan, D. Madras, D. H. Yang, D. Hadfield-Menell, G. K. Hadfield, M. Ghassemi. Judging facts, judging norms: Training machine learning models to judge humans requires a modified approach to labeling data. Science Adv. 2023.
[5] I. Yeh, C. Lien. "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients" 2009.