Dissenting Explanations: Leveraging Disagreement to Reduce Model Overreliance

Omer Reingold, Judy Hanwen Shen, Aditi Talati
Stanford University

Stanford Computer Science

Motivation

Trust in Explanations:

- While explainability is a desirable characteristic of increasingly complex black-box models, modern explanation methods have been shown to be inconsistent and contradictory (Krishna et al. 2022).
- Seemingly trivial choices in model architectures, random seeds, and hyperparameters may lead to inconsistent and contradicting explanations (Brunet et. al. 2022)
- The effect of explanations on model overreliance appears to depend on the task at hand (Bansal et al. 2021, Vasconcelos et al. 2022)

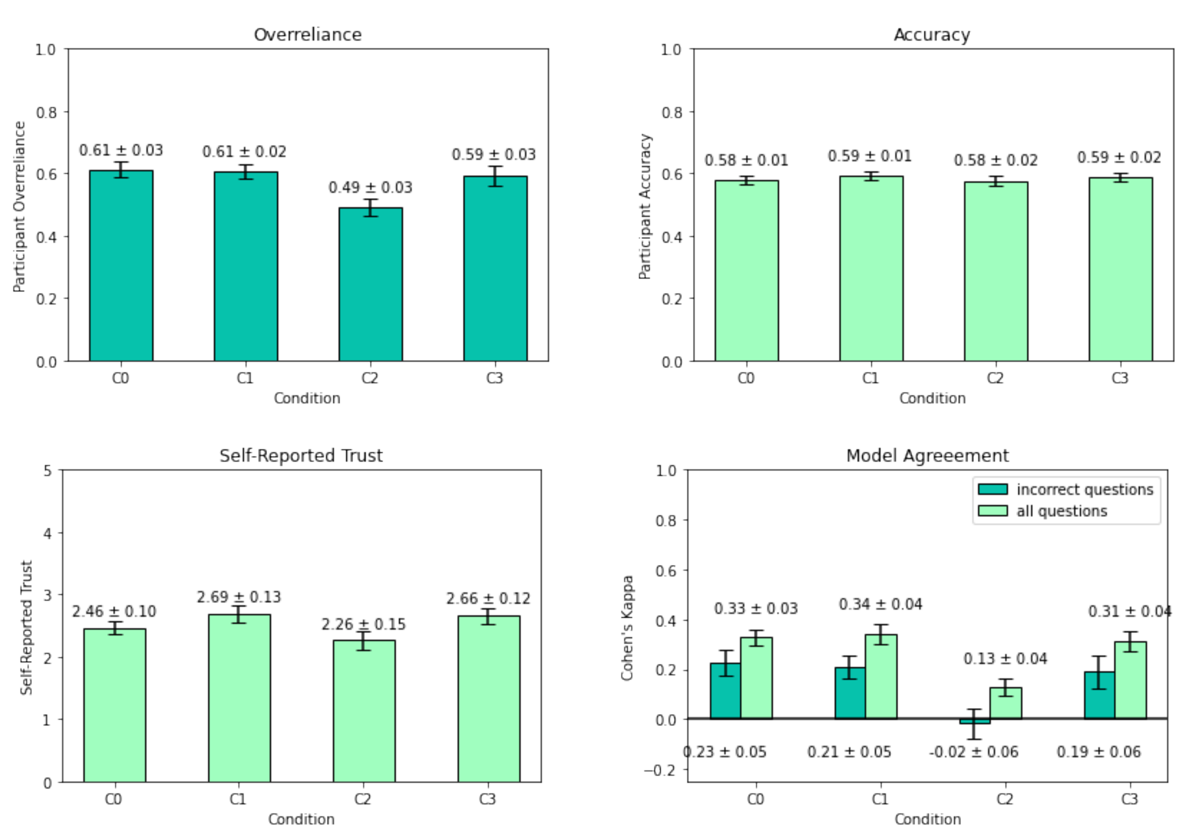
High Level Questions:

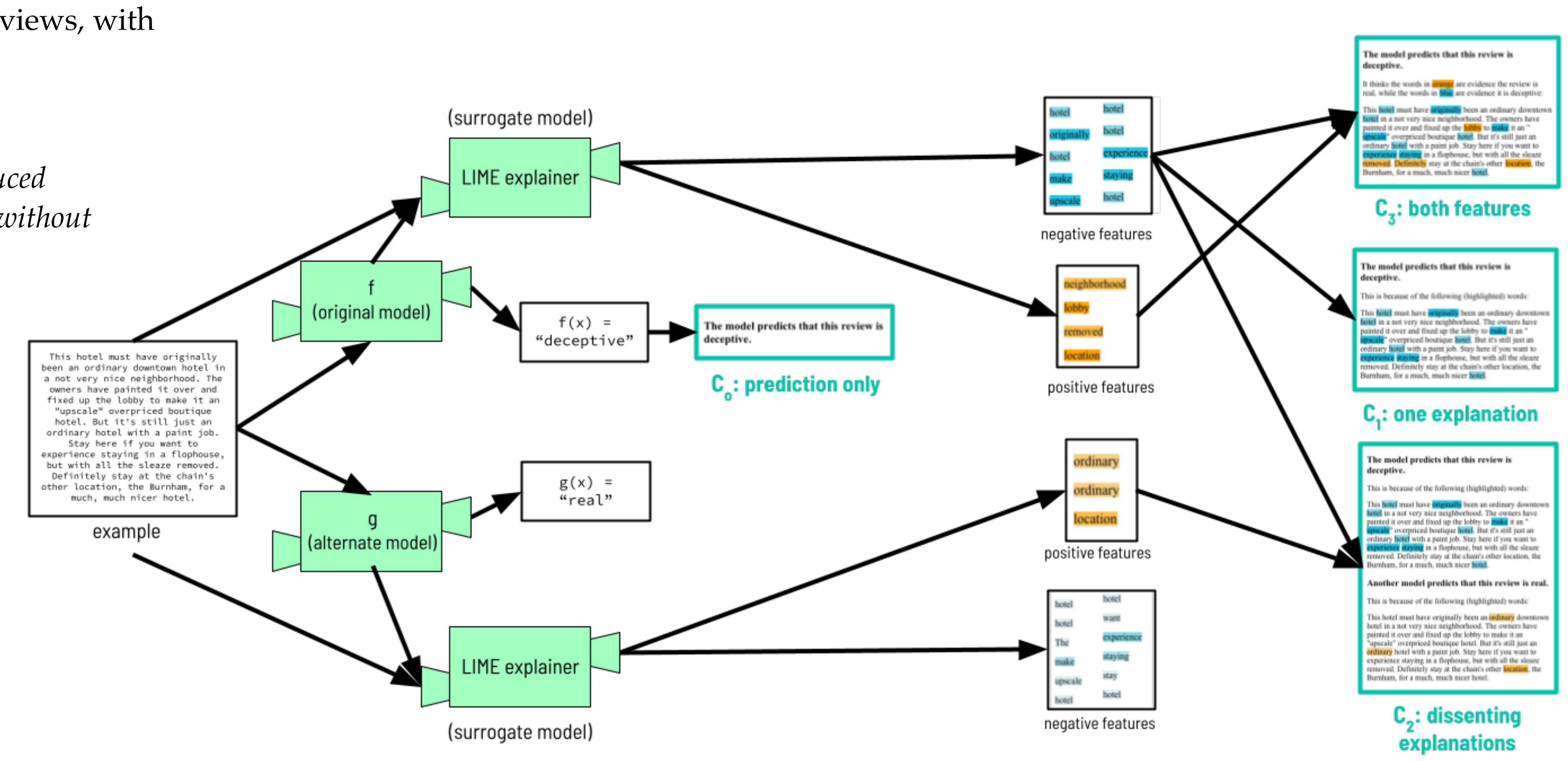
To what extent do explanations "explain" a decision and to what extent do they merely advocate for a decision?

Can we leverage disagreement in models and explanations to reduce overreliance on incorrect model outputs for general predictive tasks?

User Study: The Importance of Dissenting Explanations

- asked users to distinguish between real and made-up hotel reviews, with AI assistance
- 20 reviews, 8 of which the AI predicted incorrectly
- participants were given one of four types of explanations
- found that providing dissenting explanations *significantly reduced overreliance* as compared to a singular explanation (p = 0.001) *without reducing participant accuracy* (p = 0.210)





(a) the four different explanation types

Dissenting Explanations

Definition 3.1 (Dissenting Explanation). Let f, g be any two different classifiers and let $(x, y) \sim \mathcal{D}$ be any example. Then, $e(x, g)$ is a *dissenting explanation* for $e(x, f)$ if $f(x) \neq g(x)$.

f(x) original model → explanation e(x, f)

f(x) ≠ g(x)

g(x) alternate model → dissenting explanation e(x, g)

Techniques: Generating Dissenting Explanations

Global Model Disagreement

Definition 3.2 (Global predictive disagreement). Let f, g be any two different classifiers, the global disagreement between f and g on some set D is:

$$\delta_D(f, g) = \frac{1}{|D|} \sum_{x \in D} \mathbb{1}[f(x) \neq g(x)]$$

Problem 5.1. *Given reference model f and training data D, find some g such that $\delta_D(f, g)$ (Definition 3.2) is maximized while $\mathrm{Err}_{D_{test}}(f) \approx \mathrm{Err}_{D_{test}}(g)$.*

Method #1 (Regularization)

$$L(x, y, f) = \frac{1}{n}\sum_{i=1}^{n} l(g(x_i), y_i) + \frac{\lambda}{n}\sum_{i=1}^{n} l(g(x_i), \overline{f(x_i)})$$

λ	Accuracy	Disagreement	Corr.
0.0	0.889 ± .010	8.66 ± 0.6 %	40.1 %
0.1	0.883 ± .017	8.75 ± 0.5 %	38.9 %
0.25	0.859 ± .021	10.9 ± 3.4 %	34.2 %
0.5	**0.807 ± .017**	**16.6 ± 2.3 %**	**35.7 %**

(a) REG objective (batch size 10)

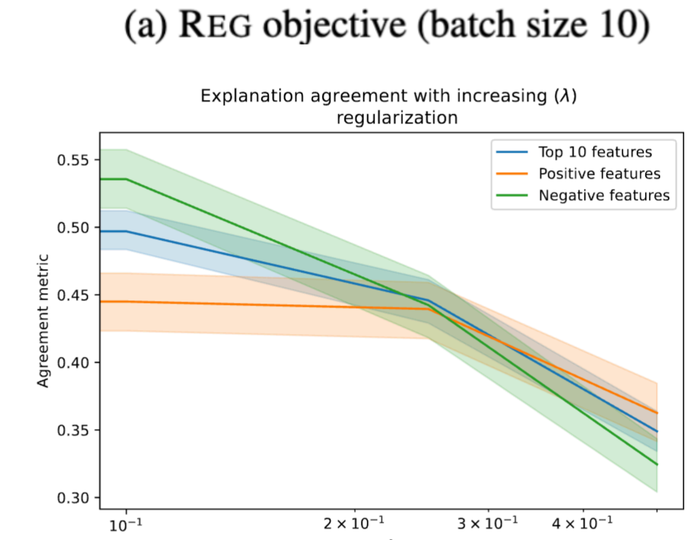
Method #2 (Reweighing)

$$L(x, y, f) = \frac{1}{n}\sum_{i=1}^{n} w_i l(g(x_i), y_i)$$

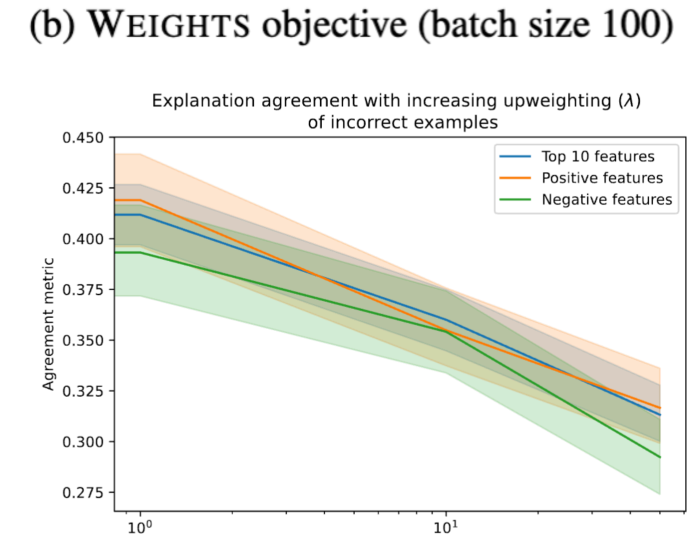
$$w_i = 1 + \lambda \mathbb{1}[f(x_i) \neq y_i)]$$

λ	Accuracy	Disagreement	Corr.
0	0.859 ± .019	8.68 ± 0.7 %	28.4%
1	0.865 ± .014	8.56 ± 1.2 %	30.5%
10	0.854 ± .008	10.8 ± 1.5 %	35.3%
50	**0.826 ± .018**	**14.9 ± 0.7 %**	**40.1%**

(b) WEIGHTS objective (batch size 100)



Explanation Agreement



Local Model Disagreement

Problem 5.3. *Given reference model f, training data D, and a test instance x find some g where $f(x) \neq g(x)$ where $\mathrm{Err}_{D_{test}}(f) \approx \mathrm{Err}_{D_{test}}(g)$.*

|D|	Success Rate	TOPK Agree.	Acc.
1280	0.543 ± .249	0.756 ± .131	0.880
640	0.723 ± .200	0.464 ± .122	0.889
320	0.910 ± .082	0.352 ± .111	0.844
160	0.987 ± .013	0.275 ± .115	0.780
80	1.000 ± .000	0.227 ± .103	0.675

(a) SVM

Iter.	Freq.	TOPK Agree.	Acc.
<5	19.7%	0.946 ± .091	0.902
5-10	20.9%	0.878 ± .113	0.892
10-15	18.1%	0.786 ± .117	0.886
15-20	19.1%	0.770 ± .159	0.883
>20	22.2%	0.782 ± .114	0.869

(b) Neural Network

We can generate both local and global model disagreement in order to create dissenting explanations even without relying on model multiplicity.

References

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and Weld, D. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp.1–16, 2021.

Brunet, M.-E., Anderson, A., and Zemel, R. Implications of model indeterminacy for explanations of automated decisions. In Advances in Neural Information Processing Systems

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., and Lakkaraju, H. The disagreement problem in explainable machine learning: A practitioner's perspective. arXiv preprint arXiv:2202.01602, 2022.

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M., and Krishna, R. Explanations can reduce overreliance on ai systems during decision-making. arXiv preprint arXiv:2212.06823, 2022.