Unpacking the Black Box: Regulating Algorithmic Decisions

Laura Blattner[1] Scott Nelson[2] Jann Spiess[1]

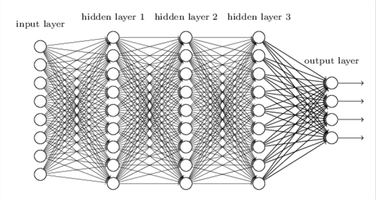
November 2021

[1]Stanford GSB and [2]Chicago Booth

Motivation

- Reliance on **prediction algorithms** in high-stakes screening decisions
- Incentive conflicts between agents building prediction functions and principals overseeing their use
  - *Medical testing*: Insurance company worries hospital over-predicts risk
  - *Hiring*: Employer worries about fairness of job offers by hiring agency
  - *Lending*: Financial regulator worries about model risk or disparate impact
- Move to automated rules allow for systematic (even ex-ante) review, but is complicated by complexity of algorithms

Brain illustration: Yunus Şahin



Neural network illustration: Michael Nielsen

- **This paper:** How can we effectively mitigate incentive conflicts if black-box algorithms are too complex to be fully described?

This Paper

- Automated rules allow for systematic scrutiny of screening decisions
- Complexity $\rightarrow$ face decision how to *restrict* and *explain* them
- How can we effectively mitigate incentive conflicts if black-box algorithms are too complex to be fully described?
  - ☹ Ex-ante restrictions to simple functions inefficient
  - 😐 Use an algorithmic audit based on a simpler representation of the algorithm ('explainer')
  - ☺ Design the audit to target the dimensions affected most by incentive conflict ('targeted explainer')
- Theoretically, make precise and justify explanations of complex ML models in a principal-agent model where explainability is means to an end
- Empirically, demonstrate that results matter for credit underwriting

Contribution

1. Nascent literature on *incentive conflicts and algorithmic design* (e.g. Rambachan et al. 2020; Gillis and Spiess 2019; Athey et al. 2020).
    - **We apply principal-agent toolbox to (realistic) case where algorithms are too complex to be described**
2. Finance literature on *disclosure and supervision* (Goldstein and Leitner, 2013; Parlatore and Phillipon, 2020)
    - **We study disclosure design when available information is limited and compare and contrast audit designs on real-world data**
3. Computer science literature on *algorithmic explainability* (e.g., Lakkaraju and Bastani, 2020; Slack et al., 2020; Lakkaraju et al., 2019)
    - **We derive optimal explainer design from economic theory and apply on real world data**

Regulator–Lender Example

1. *Rule-setting stage:* **Regulator** sets the rules of the game

2. *Training stage:* **Lender** learns relationship

$$s(X) = \alpha + \beta \underbrace{X_1}_{\text{past default}} + \gamma \overbrace{X_2}^{\text{high utilization}} + \delta X_1 \cdot X_2$$

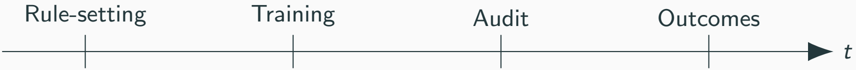
   between features X and default, chooses credit score

$$\hat{f}(X) = \alpha + \hat{\beta} X_1 + \hat{\gamma} X_2 + \hat{\delta} X_1 \cdot X_2$$

3. *Audit stage:* **Regulator** performs audit

4. *Outcome stage:* Consequences of deploying $\hat{f}$ and payoffs are realized

Misalignment



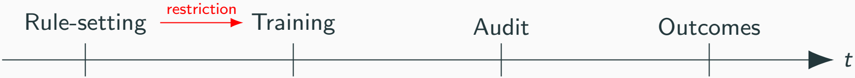
Regulator welfare:

$$W(f;d) = \underbrace{\text{prediction fit}}_{-\mathrm{E}[(f(X)-s(X))^2]} - \underbrace{\text{penalty for disparate impact}}_{\lambda\,(\mathrm{E}_d[f(X)|G=0]-\mathrm{E}_d[f(X)|G=1])^2}$$

Lender utility:

$$U(f;d) = \begin{cases} \overbrace{\text{prediction fit} + \text{profit from subprime loans}}^{-\mathrm{E}_d[(f(X)-s(X)-\Delta_{\text{overall}}-\Delta_{X_2}X_2)^2]}, & \text{audit passes} \\ -\infty, & \text{audit fails} \end{cases}$$

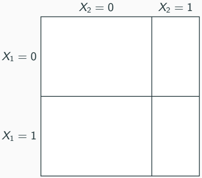
Policy Tool: Ex-Ante Function Restriction

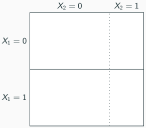


Policy	Alignment	Flexibility/efficiency
No restriction	☹	☺
Ex-ante restriction	☺	☹

Policy Tool: Restriction on Explanation

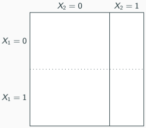
- **Information constraint:** Regulator cannot process fully complex $\hat{f}(X) = \hat{\alpha} + \hat{\beta}\, X_1 + \hat{\gamma}\, X_2 + \hat{\delta}\, X_1 \cdot X_2$ (or firm does not reveal)
- **Low-dim explainer:** Can process 2-dim linear projection $\mathrm{E} : \mathcal{F} \to \mathbb{R}^2, f \mapsto \mathrm{E}f$
- **Audit based on explainer:** Decide audit based on simple explanation





$$\mathrm{E}_0 f = \begin{pmatrix} \mathsf{E}_d[f(X)|X_1 = 1] \\ \mathsf{E}_d[f(X)|X_1 = 0] \end{pmatrix}$$

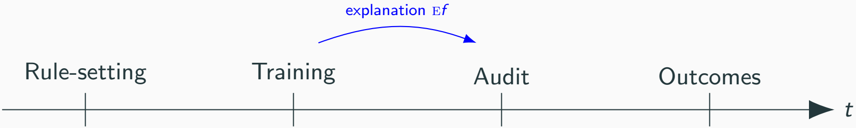
Best prediction explainer: max. overall information $\Rightarrow \mathrm{E}_0$: regress on const., X_1



$$\mathrm{E}^* f = \begin{pmatrix} \mathsf{E}_d[f(X)|X_2 = 1] \\ \mathsf{E}_d[f(X)|X_2 = 0] \end{pmatrix}$$

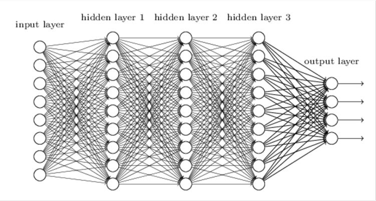
Targeted explainer: inspect misalignment $\Rightarrow \mathrm{E}^*$: regress on const., X_2

Policy Tool: Restriction on Explanation

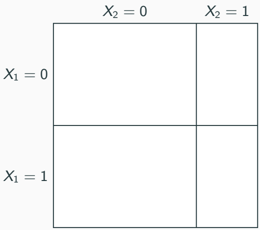


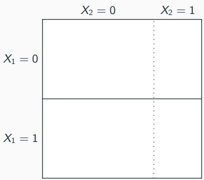
Policy	Alignment	Flexibility/efficiency
No restriction	☹	☺
Ex-ante restriction	☺	☹
Prediction explainer	😐	☺
Targeted explainer	☺	☺

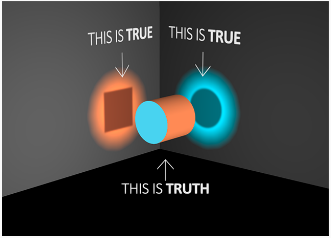
Complex Functions, Simple Explanations



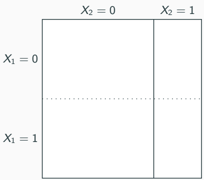
Neural network illustration: Michael Nielsen



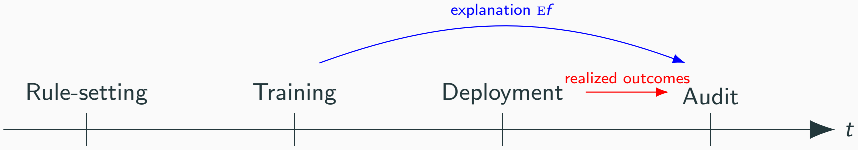




"This is Truth", viral3d.com

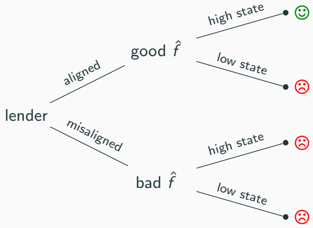


Ex-Post Audits



- So far have assumed that audit uses info from *before* deployment
  - Opportunity to avoid bad outcomes *before* they happen
  - Outcomes may be unobserved or only realized with delay
  - Limited liability or risk aversion may limit effectiveness of ex-post audits

- What is the role of explainers if outcomes are *also* available?



- Enforces conservative choice, inefficient if uncertainty is high
- Regulation should depend on the contribution of lender to outcome
- Optimal regulation can combine both

Taking Theory to the Data

- **Data**: TransUnion credit report data + Infutor semi-annual panel on 50m ppl from 2009–2017 (as in Blattner, Nelson 2020; here: use 50k subsample)

- Build **prediction function for credit card default** with custom loss function to model three cases:

Lender minimizes prediction loss	$\min_f \left[\underbrace{-\mathrm{E}[Y \log \hat{Y} + (1-Y)\log(1-\hat{Y})]}_{\text{prediction loss}}\right]$
Regulator minimizes pred. loss plus loss from social preference	$\min_f$ pred. loss+ $\lambda\Big(\underbrace{\mathrm{E}[\mathrm{logit}(\hat{Y})\vert M{=}1] - \mathrm{E}[\mathrm{logit}(\hat{Y})\vert M{=}0]}_{\text{social preference}}\Big)^2$
Lender subject to audit constraint	$\min_f$ pred. loss $+\, \varphi \underbrace{N_{J,\beta^*_{\text{regulator}}}(\hat{f})}_{\text{audit constraint}}$

- **Implementation**: Neural net w/ 2 hidden layers 40 neurons on 50 covariates; stochastic gradient decent with Adam in TensorFlow

Explainers in the Data

Neural network generates predictions $\hat{f}(X)$

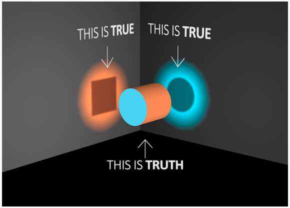
- **Best prediction explainer**: J from LASSO logit on credit card default (Y)

$$\begin{aligned}\hat{f}(X_i) =& \beta_0 + \beta_1 \# \text{ trades delinquent} \\ & + \beta_2 \text{ agg. credit line} + \cdots + \beta_{10} \# \text{ bankruptcies} + \epsilon_i\end{aligned}$$

- **Targeted explainer**: J from LASSO logit on group status (G)

$$\begin{aligned}\hat{f}(X_i) =& \beta_0 + \beta_1 \# \text{ trades delinquent} \\ & + \beta_2 \# \text{ unpaid collections} + \cdots + \beta_{10} \# \text{ collections} + \epsilon_i\end{aligned}$$

- **Audit constraint**: $\hat{\beta}_J = \beta_J^*$



"This is Truth", viral3d.com

Results: Disparate Impact

1. Complex model improves predictive performance relative to simple model;
2. Neural net allows for larger preference misalignment than (simpler) logit;
3. Targeted explainer better than prediction explainer at aligning incentives.

	AUC	Log loss	Δ log odds (disparate impact)
Neural network (two hidden layers) on 50 covariates			
Lender	0.842	0.327	0.935
Regulator (wants small Δ log odds)	0.834	0.337	0.450
Lender w/ prediction explainer	0.834	0.343	0.535
Lender w/ targeted explainer	0.828	0.351	0.457
Logistic regression on 20 covariates			
Lender	0.797	0.360	0.517
Regulator (wants small Δ log odds)	0.795	0.361	0.331
Prediction explainer	0.797	0.359	0.336
Targeted explainer	0.795	0.362	0.332

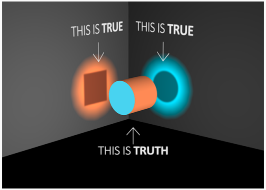
Conclusion

Opportunity and challenge: Move to automated rules allows for systematic scrutiny, but complexity means we face decision how to *restrict* and *explain* them

Broader context: Explainability, interpretability and transparency central to machine learning implementation, but often lack clear definition and motivation

This paper: How to regulate black-box algorithms that are too complex to be described completely? Answer from principal–agent model: targeted explainers!

Related Agenda: Evaluate explainer tools for financial regulation (with FinRegLab)



"This is Truth", viral3d.com

Thank you!
`jspiess@stanford.edu`