Explanations with a Purpose: Regulating Black-Box Algorithmic Decisions

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

May 2023

[1]Stanford GSB and [2]Chicago Booth

Delegation approach to econometric decisions

1. **Principal** (designer) observes η and chooses $\mathcal{C} \subseteq \mathbb{R}^{\mathcal{Z}}$ to minimize

$$\mathsf{E}_\eta \, \mathsf{E}_\pi L^P(\hat{\tau}(\mathcal{C}); \theta)$$

2. **Agent** (researcher) observes $\pi \sim \mathsf{P}_\eta$ and chooses $\hat{\tau} \in \mathcal{C}$ to minimize

$$\mathsf{E}_\pi L^A(\hat{\tau}; \theta)$$

Delegation approach to econometric decisions

1. **Principal** (designer) observes η and chooses $\mathcal{C} \subseteq \mathbb{R}^{\mathcal{Z}}$ to minimize

$$\mathsf{E}_\eta \, \mathsf{E}_\pi L^P(\hat{\tau}(\mathcal{C}); \theta)$$

2. **Agent** (researcher) observes $\pi \sim \mathsf{P}_\eta$ and chooses $\hat{\tau} \in \mathcal{C}$ to minimize

$$\mathsf{E}_\pi L^A(\hat{\tau}; \theta)$$

by specifiying function class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$, loss $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, mapping $T : \mathcal{F} \to \mathcal{C}, \hat{f} \mapsto \hat{\tau}$

3. **Algorithm** observes data $z \sim \mathsf{P}_\theta$, chooses $\hat{f}$ to minimize (optimistically)

$$\mathsf{E}_\pi[\mathsf{E}_\theta[\ell(\hat{f}(x), y)]|z]$$

or (practically)

$$\mathbb{E}_z[\ell(\hat{f}(x), y)]$$

- **Robustness:** $T(\hat{f}) \in \mathcal{C}$ for all $\hat{f} \in \mathcal{F}$
- **Efficiency:** $\hat{\tau} = T(\hat{f})$ good solution to original goal

Data-driven decisions with multiple objectives across domains

- Robust integration of machine learning into causal inference
- Design of pre-analysis plans
- Strategic classification (Hardt et al., 2016)
- AI alignment (Hadfield-Menell and Hadfield, 2019)
- Manipulation-proof machine learning (Björkegren et al., 2020)
- Regulation of AI (Rambachan et al., 2020)
- Prediction-powered inference (Angelopoulos et al., 2023)

Claim: Integration econometrics, ML, data-driven decision making with mechanism design

- can be good frame to diagnose and address misalignment, and
- allows leveraging formal tools from mechanism design

Explanations with a Purpose: Regulating Black-Box Algorithmic Decisions

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

May 2023

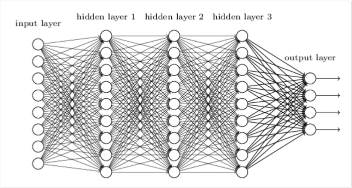
[1]Stanford GSB and [2]Chicago Booth

Motivation

- **Prediction algorithms** in high-stakes screening decisions (medical testing, hiring, lending)
- Incentive conflicts between agents building prediction functions and principals overseeing their use
  - *Medical testing*: Insurance company worries hospital over-predicts risk
  - *Hiring*: Employer concerned about fairness of interview invites by manager
  - *Lending*: Financial regulator worries about disparate impact or model risk
- Move to automated rules allows for systematic (even ex-ante) review,
  but is complicated by complexity of algorithms, leading for calls around simplicity and transparency

Brain illustration: Yunus Şahin

$\longrightarrow$

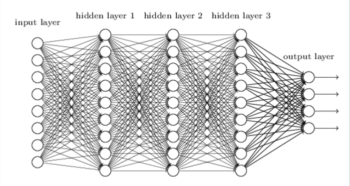


Neural network illustration: Michael Nielsen

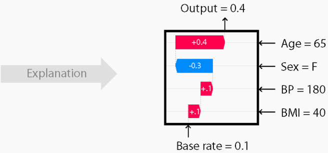
- **This project:** Study in principal–agent model how can effectively mitigate incentive conflicts if black-box algorithms are too complex to be fully described, apply to credit data

Complexity and explanations in a principal–agent model

- **Starting point:** Complexity of algorithms means agent cannot fully describe algorithm to principal
- **First policy option:** Limit agent to simple/transparent algorithms that can be fully described
- **Second policy option:** Principal requires agent to provide a simple description/explanation of algorithm behavior in terms of key drivers or limited data



Neural network illustration: Michael Nielsen



This Project

- **Theoretically**, make precise and justify explanations of complex ML models in a principal-agent model where explainability is *means to an end*
  - ☹ Ex-ante restrictions to simple, fully transparent functions
  - 😐 Oversight based on a simpler representation of the algorithm (‘explainer’)
  - ☺ Design the explainer to target the dimensions affected most by incentive conflict (‘targeted explainer’)
- **Empirically**, demonstrate that results matter in two substantial applications to credit underwriting

Contribution

1. Law and economics literature on *fairness and discrimination oversight of algorithms* (e.g. Kleinberg et al., 2018; Gillis and Spiess 2019; Hellman, 2019; Yang and Dobbie, 2020)
   - **We derive optimal restrictions in a principal-agent model with explicit misaligned preferences**
2. Nascent literatures on *data analysis with conflicts of interest and replication concerns* (e.g. Glaeser, 2006; Di Tillio et al., 2017; Spiess, 2018) as well as *incentive conflicts and algorithmic design* (e.g. Rambachan et al. 2020; Athey et al. 2020)
   - **We apply principal-agent toolbox to (realistic) case where algorithms too complex to be described**
3. Finance literature on *disclosure and supervision* (e.g. Goldstein and Leitner, 2013; Parlatore and Phillipon, 2020)
   - **We study disclosure design when available information is limited, evaluate on real-world data**
4. 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**
5. Mechanism-design literature on *optimal delegation* (including Holmstrom, 1977, 1984; Melumad and Shibano, 1991; Alonso and Matouschek, 2008; Frankel, 2014)
   - **We consider delegation with a complexity constraint**

Overview

Overview

Delegation Setup

- An **agent** chooses a **prediction function** $f : \mathcal{X} \to \mathbb{R}$ to maximize utility $U^A(f;\theta)$
- The choice is overseen by a **principal** with utility $U^P(f;\theta)$

1. Principal chooses restriction $\widehat{\mathcal{F}} \subseteq \mathcal{F}$ based on prior π
2. Agent chooses $\hat{f} \in \widehat{\mathcal{F}}$ based on training signal $\theta \sim \mathrm{P}_\pi$

Delegation Setup

- An **agent** chooses a **prediction function** $f : \mathcal{X} \to \mathbb{R}$ to maximize utility $U^A(f;\theta)$
- The choice is overseen by a **principal** with utility $U^P(f;\theta)$

0. Principal sets rules
   - Ex-ante restrict lender to simple functions $\mathcal{F} \cong \mathcal{E}$ that can be fully explained *or*
   - Leave functions ex-ante unrestricted ($\mathcal{F} = \mathbb{R}^{\mathcal{X}}$), and choose explanation mapping $\mathrm{E} : \mathcal{F} \to \mathcal{E}$
1. Principal chooses restriction $\widehat{\mathcal{F}} \subseteq \mathcal{F}$ based on training signal θ

   Principal cannot observe complex $f \in \mathbb{R}^{\mathcal{X}}$, only lossy "explanation" $\mathrm{E}f \in \mathcal{E}$, so

   $$\widehat{\mathcal{F}} = \{f \in \mathcal{F}; \mathrm{E}f \in \widehat{\mathcal{E}}\}$$

   - Simple proxy models, e.g. linear projection on a few covariates
   - Variable-importance measures, such as SHAP for complex machine-learning models
   - Evaluation at a limited number of data points $x \in \mathcal{X}$
2. Agent chooses $\hat{f} \in \widehat{\mathcal{F}}$ based on training signal θ

Overview

Lending Example

- An **agent** chooses a **prediction function** $f \in \mathbb{R}^{\mathcal{X}}$ to maximize utility $U^A(f;\theta)$
- The choice is overseen by a **principal** with utility $U^P(f;\theta)$

Lending Example

- A **lender** chooses a **credit score** $f \in \mathbb{R}^{\mathcal{X}}$ for data (Y, X), where $Y \in \{0, 1\}$ repayment and $X \in \mathcal{X}$ credit file, to maximize $U^A(f; \theta) = \mathrm{E}_\theta[u(f(X), Y)]$
  - Credit scoring utility: $\boxed{u(f(X), Y) = -(Y - f(X))^2}$
  - Loan profit: $u(f(X), Y) = r\, \mathbb{1}(f(X) \geq p^*)\, Y - c\, \mathbb{1}(f(X) \geq p^*)\, (1 - Y)$
- Choice is **overseen by a regulator** maximizing utility $U^P(f; \theta)$
  - Risk preference (different $Y|X$, same X):

$$\boxed{U^P(f; \theta) = \mathrm{E}_\theta[u(f(X), Y)|S{=}\text{low}]} \qquad S \in \{\text{high}, \text{low}\}$$

  - Target population (same $Y|X$, different X):

$$U^P(f; \theta) = \mathrm{E}_\theta[u(f(X), Y)|D{=}\text{new customers}] \qquad D \in \{\text{new customers}, \text{existing customers}\}$$

  - Disparate impact (majority indicator G):

$$U^P(f; \theta) = \mathrm{E}_\theta[u(f(X), Y)] - \lambda(\mathrm{E}_\theta[f(X)|G = 1] - \mathrm{E}_\theta[f(X)|G = 0])$$

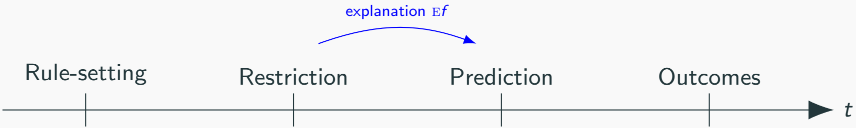
Timing and Choices

$$P(Y{=}1|X,S) = \alpha(S) + \beta \underbrace{X_1}_{\text{past default}} + \gamma(S) \overbrace{X_2}^{\text{HELOC}} + \delta X_1 \cdot X_2 \qquad S \in \{\text{high}, \text{low}\}$$

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

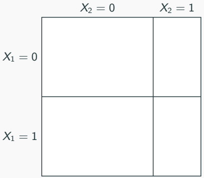
0. *Rule-setting stage:* **Regulator** sets the rules of the game
1. *Restriction stage:* **Regulator** sets restrictions based on limited information about $\hat{f}$
2. *Prediction stage:* **Lender** learns relationship (here: two covariates, binary) between features X and repayment Y, chooses credit score $\hat{f}(X)$

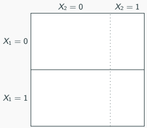
Complex Functions, Simple Explanations



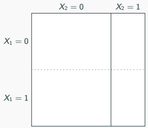
- **Information constraint:** Regulator cannot process fully complex $\hat{f}(X) = \hat{\alpha} + \hat{\beta}\, X_1 + \hat{\gamma}\, X_2 + \hat{\delta}\, X_1 \cdot X_2$ (or lender does not reveal)

- **Low-dim explainer:** Projection $\mathrm{E} : \mathcal{F} \to \mathcal{E}, f \mapsto \mathrm{E}f$ on one of covariates





$$\mathrm{E}_1 f = \begin{pmatrix} \mathsf{E}[f(X)|X_1 = 1] \\ \mathsf{E}[f(X)|X_1 = 0] \end{pmatrix}$$



$$\mathrm{E}_2 f = \begin{pmatrix} \mathsf{E}[f(X)|X_2 = 1] \\ \mathsf{E}[f(X)|X_2 = 0] \end{pmatrix}$$

Baseline Policy Choices: No Regulation and Function Restrictions

Lender **learns** the distribution of repayment probabilities

$$\mathrm{P}_\theta(Y=1|X,S) = \alpha(S) + \beta \underbrace{\tilde{X}_1}_{\text{past default}} + \gamma(S) \overbrace{\tilde{X}_2}^{\text{HELOC}} + \delta \tilde{X}_1 \cdot \tilde{X}_2$$

where centered and reparametrized so that $\mathrm{E}[\tilde{X}_1] = 0 = \mathrm{E}[\tilde{X}_2], \tilde{X}_1 \perp \tilde{X}_2$

Lender and regulator **maximize**

$$U^A(f;\theta) = \mathrm{E}_\theta[-(Y - f(X))^2] \qquad U^P(f;\theta) = \mathrm{E}_\theta[-(Y - f(X))^2 | S{=}\text{low}]$$

Lender prefers:	Regulator prefers:	Both agree on:
$\hat{\alpha} = \bar{\alpha} = \mathrm{E}_\theta[\alpha]$	$\hat{\alpha} = \alpha(\text{low})$	$\hat{\beta} = \beta$
$\hat{\gamma} = \bar{\gamma} = \mathrm{E}_\theta[\gamma]$	$\hat{\gamma} = \gamma(\text{low})$	$\hat{\delta} = \delta$

1. **No function restriction, no audit**. Get maximal distortion

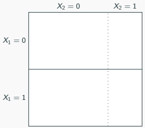
$$\hat{f}(X) = \bar{\alpha} + \beta \tilde{X}_1 + \bar{\gamma} \tilde{X}_2 + \delta \tilde{X}_1 \cdot \tilde{X}_2$$

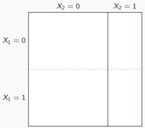
2. **Ex-ante restriction to explainable function**. Eliminates misalignment at large cost

$$\hat{f}(X) = \alpha(\text{low}) + \beta \tilde{X}_1$$

Policy Choices: Explainer Audits

Information constraint: Regulator cannot process fully complex $\hat{f}(X) = \hat{\alpha} + \hat{\beta}\,\tilde{X}_1 + \hat{\gamma}\,\tilde{X}_2 + \hat{\delta}\,\tilde{X}_1 \cdot \tilde{X}_2$





Agnostic explainer: max. overall information
$\Rightarrow \mathrm{E}_0$: regress $\hat{f}(X)$ on constant and $\tilde{X}_1$

Targeted explainer: inspect misalignment
$\Rightarrow \mathrm{E}^*$: regress $\hat{f}(X)$ on constant and $\tilde{X}_2$

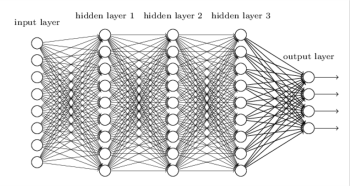
3. **No restriction, audit w/ agnostic explainer E_0.** Partially aligns choices

$$\hat{f}(X) = \alpha(\text{low}) + \beta\,\tilde{X}_1 + \underbrace{\bar{\gamma}\,\tilde{X}_2}_{\text{not detectable by } \mathrm{E}_0} + \delta\,\tilde{X}_1 \cdot \tilde{X}_2$$

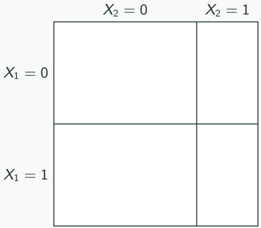
4. **No restriction, audit w/ targeted explainer E^*.** Can achieve first best

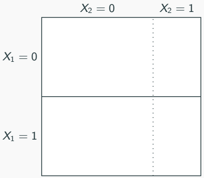
$$\hat{f}(X) = \alpha(\text{low}) + \beta\,\tilde{X}_1 + \gamma(\text{low})\,\tilde{X}_2 + \delta\,\tilde{X}_1 \cdot \tilde{X}_2$$

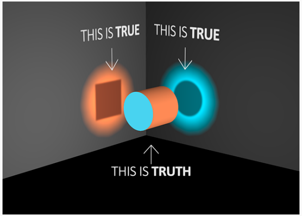
Complex Functions, Simple Explanations



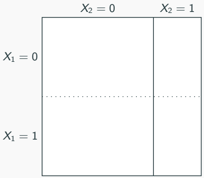
Neural network illustration: Michael Nielsen







“This is Truth”, viral3d.com



Overview

General Delegation Model

- Misaligned preferences over choice $f \in \mathbb{R}^{\mathcal{X}}$

$$U^A(f;\theta) = \int_{\mathcal{X}} u^A(f(x), x; \theta)\ \mathrm{d}\mu^A(x;\theta) \qquad U^P(f;\theta) = \int_{\mathcal{X}} u^P(f(x), x; \theta)\ \mathrm{d}\mu^P(x;\theta)$$

- Delegation game
  1. Principal chooses $\hat{\mathcal{F}} \subseteq \mathcal{F}$
  2. Agent chooses $\hat{f} \in \hat{\mathcal{F}}$

- Explanation constraint

$$\hat{\mathcal{F}} = \{f \in \mathcal{F}; \mathrm{E}f \in \hat{\mathcal{E}}\} \qquad \mathrm{E} : \mathcal{F} \to \mathcal{E}$$

- Consider two **policy design choices**
  - Restrict functions from $\mathcal{F} = \mathbb{R}^{\mathcal{X}}$ to $\mathcal{F} = \mathcal{E}$ to achieve perfect alignment
  - Otherwise, design of explainer E

General Results

Covariate shifts: $U^A(f;\theta) = \int_{\mathcal{X}} u(f(x), x; \theta)\ \mathrm{d}\mu^A(x;\theta) \quad U^P(f;\theta) = \int_{\mathcal{X}} u(f(x), x; \theta)\ \mathrm{d}\mu^P(x;\theta)$

Assume that $\mu^P(\cdot;\theta) \ll \mu^A(\cdot;\theta)$ then choices from $\mathcal{F} = \mathbb{R}^{\mathcal{X}}$ are aligned

Model shift: $U^A(f;\theta) = \int_{\mathcal{X}} u^A(f(x), x; \theta)\ \mathrm{d}\mu(x) \qquad U^P(f;\theta) = \int_{\mathcal{X}} u^P(f(x), x; \theta)\ \mathrm{d}\mu(x)$

$$u^A(f(x), x; \theta) = -(f(x) - f^A(x;\theta))^2 \qquad u^P(f(x), x; \theta) = -(f(x) - f^P(x;\theta))^2$$

When $\min_S \mathrm{E}_\eta \min_\beta \int_{\mathcal{X}} (f^A(x) - f^P(x) - x_S'\beta)^2\ \mathrm{d}\mu(x) < \min_S \mathrm{E}_\eta \min_\beta \int_{\mathcal{X}} (f^P(x) - x_S'\beta)^2\ \mathrm{d}\mu(x)$ then optimal regulation = no ex-ante constraint + targeted explainer

Distributional preference: $U^P(f;\theta) = U^A(f;\theta) - \lambda \left(\int_{\mathcal{X}} f(x)\ \mathrm{d}\mu_1(x) - \int_{\mathcal{X}} f(x)\ \mathrm{d}\mu_0(x) \right)$

Equivalent to model shift with $u^P(f(x), x; \theta) = u(f(x), x; \theta) - \lambda \left(\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu} - \frac{\mathrm{d}\mu_0}{\mathrm{d}\mu} \right)$;
optimal targeted explainer is best prediction of group identity

Overview

Input-Based Restrictions vs Outcome-Based Tests

- **Input-based prohibitions:** do not allow use of/access to specific covariates
  - Often inefficient
  - Sometimes even counterproductive
- **Model-based simplicity/transparency restrictions:** limit structure of models
  - Comes at cost by shifting Pareto frontier
  - In our data cost larger than gain
- **Model-based explainability restriction:** inspect key model properties
  - Practical constraints on processing, IP often mean that information limited
  - Well-designed model summary can close the gap to first-best
- **Outcome-based audits:** use realized properties of algorithmic decisions
  - Does not fully leverage ability to describe and intervene before
  - May not be enough for counterfactual evaluation

Conclusion

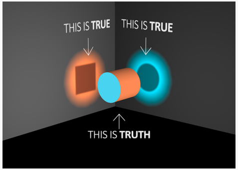
Opportunity and challenge: Move to automated rules allows for systematic scrutiny, but complexity means we face decision how to *restrict* and *explain* them

Broader context: Explainability, interpretability, transparency central to machine learning implementation and called for in policy debates, but often lack clear economic definition and motivation

This project: How to regulate black-box algorithms that are too complex to be described completely?

- *Answer from principal–agent model:* complexity–oversight trade-off leads to targeted explainers
- *Calibration in data:* excess cost of full transparency/simplicity, targeted explainers second best

Comments/new draft: `jspiess@stanford.edu`



"This is Truth", viral3d.com


Machine Learning Explainability & Fairness: Insights from Consumer Lending

Empirical White Paper

FinRegLab in collaboration with Laura Blattner and Jann Spiess

`finreglab.org`