# Unpacking the Black Box: Regulating Algorithmic Decisions

Laura Blattner Scott Nelson Jann Spiess

May 2024

**Abstract**

What should regulators of complex algorithms regulate? We propose a model of oversight over 'black-box' algorithms used in high-stakes applications such as lending, medical testing, or hiring. In our model, a regulator is limited in how much she can learn about a black-box model deployed by an agent with misaligned preferences. The regulator faces two choices: first, whether to allow for the use of complex algorithms; and second, which key properties of algorithms to regulate. We show that limiting agents to algorithms that are simple enough to be fully transparent is inefficient as long as the misalignment is limited and complex algorithms have sufficiently better performance than simple ones. Allowing for complex algorithms can improve welfare, but the gains depend on how the regulator regulates them. Regulation that focuses on the overall average behavior of algorithms, for example based on standard explainer tools, will generally be inefficient. Targeted regulation that focuses on the source of incentive misalignment, e.g., excess false positives or racial disparities, can provide second-best solutions. We provide empirical support for our theoretical findings using an application in consumer lending, where we document that complex models regulated based on context-specific explanation tools outperform simple, fully transparent models. This gain from complex models represents a Pareto improvement across our empirical applications that is preferred both by the lender and from the perspective of the financial regulator.

Authors are listed in alphabetical order.

We thank Mario Curiki, Georgy Kalashnov, and Ruying Gao for outstanding research assistance. We thank Susan Athey, Simon Freyaldenhoven, Talia Gillis, Paul Goldsmith-Pinkham, Damian Kozbur, Danielle Li, Sendhil Mullainathan, Ashesh Rambachan, Amit Seru, Ken Singleton, PR Stark, Chenzi Xu, Louis Kaplow, Kathryn Spier, the FinRegLab team, and seminar and conference participants at Stanford, Yale, Harvard, Zurich, Oxford, the NBER Summer Institute, Stanford SITE, the New Perspectives on Consumer Behavior in Credit and Payments Markets Conference, the AEA Annual Meeting, the FTC, ENSAI, the OCC, EC, the Artificial Intelligence and Big Data in Finance Research (ABFR) Forum, NASMES, the Machine Learning in Economics Summer Institute, and the 2nd Zurich AI + Economics Workshop for helpful discussions and comments. We thank the Stanford Institute for Human-Centered Artificial Intelligence (HAI) and Amazon Web Services for generous support. Any errors or omissions are the responsibility of the authors.

This manuscript supersedes an earlier version that was accepted and presented at EC'22, with an extended abstract published as: Blattner, Laura, Scott Nelson, and Jann Spiess (2022). Unpacking the Black Box: Regulating Algorithmic Decisions. In *Proceedings of the 23rd ACM Conference on Economics and Computation (EC'22)*, page 559.

# I. Introduction

The increasing adoption of complex prediction algorithms in sensitive applications such as hiring, lending, medical testing, college admissions, and pre-trial detention raises the questions how a regulator should regulate them. On the one hand, when algorithms replace human decision-makers, a regulator can now analyze and intervene in decision processes more systematically. On the other hand, the increasing complexity of artifical-intelligence algorithms makes this goal elusive, and regulation often has to rely on an imperfect understanding of complex black-box algorithmic decisions. These concerns have led to calls for restricting the complexity of algorithms and relying on simpler, fully transparent decision rules.

This article captures trade-offs between complexity and oversight of algorithms in a principal–agent model with misaligned preferences. We consider a delegation game between a principal who regulates a machine-learning algorithm and an agent who has the technology to build it. There is an incentive conflict between agent and principal, but the principal is limited in how much she can learn about the agent's black-box prediction model, and she has to regulate based on limited information about the model instead. Within this framework, we ask what a regulator of algorithms should regulate. We show that a restriction to fully transparent, simple algorithms that fully aligns choices comes at a large cost. As an alternative, we consider the regulation of complex models based on simple model explanations, and show that appropriately designed explanations that capture the misalignment between principal and agent provide a second-best solution. We then show the applicability of these results in an empirical application to consumer lending in a large credit bureau dataset.

Decision-makers increasingly rely on complex prediction algorithms to make high-stakes decisions. In many such settings, incentive conflicts arise between those building the prediction tools and the entities tasked with overseeing their use. An employer might worry about a hiring manager using a prediction model that produces low job offer rates for minority job applicants. A financial regulator might worry about lenders' risk models under-predicting credit risk to enable increased leverage. An insurance company might worry about a hospital's prediction model over-predicting the risk of heart attack leading to costly over-testing. A key challenge for algorithmic oversight is to determine when such complex algorithmic prediction functions represent the principal's preferred choices, and when they reflect misaligned incentives.

We capture the incentive conflicts between an agent who designs a prediction algorithm and a principal who regulates its use in a principal–agent model. Our setup mirrors that of a classical delegation problem, where the agent chooses a prediction function according to his preferences, but is subject to constraints set by the principal. For example, a company may restrict which algorithms can be used for hiring, a financial regulator may put constraints on risk-scoring models used by banks, and an insurance company may set standards for the screening algorithms used in a hospital. Classical delegation models going back to Holmström (1984) and further developed in e.g. Alonso and Matouschek (2008) and Frankel (2014) suggest solutions that partially restrict agent choices in a way that increases alignment, while still leaving enough flexibility to leverage the agent's technology and private information.

In this article, we depart from standard delegation models by assuming that the agent's choice of prediction function may be too complex to be fully available to the principal. Specifically, we assume that the prediction function chosen by the agent comes from a black-box machine learning model with so many parameters that fully communicating it may not be feasible. This restriction is motivated by the increasing use of very complex AI models in high-stakes decisions, such as deep neural networks and boosted tree models. The constraint may also be motivated by concerns around releasing proprietary algorithms that the maker of the algorithm may want to protect, precluding the communication of the full prediction function to the principal.

Instead of accessing the full algorithm, we assume that the principal can reglate an algorithm only based on a simple description of the complex prediction function. Building upon work in computer science, we call this simple description an "explanation" of the complex model. For example, the behavior of a complex neural network may be described in terms of a few key variables that explain a good fraction of the model's behavior, such as those identified by tools like SHAP (Lundberg and Lee, 2017). Similarly, a complex tree model may be described in terms of those covariates deemed most important according to their contribution to splitting the data. In our formal model, we capture this idea by assuming that the principal can only capture a simple (linear) projection of the complex prediction model on a few key variables. Consequently, the principal misses some of the behavior of the complex model.

A first response to the limited ability of the principal to capture complex models could be to restrict the agent to fully transparent, simple models that can be fully regulated. Within our framework, we argue that restrictions to fully explainable models come at a potentially large cost. For example, an employer could require a hiring manager to rely on simple decision trees, a financial regulator may require a bank to score creditworthiness only using a relatively interpretable logistic regression, or an insurance company may require a hospital to use a transparent decision rule for assessing the risk of heart attack. While fully aligning choices, such restrictions also reduce the efficiency of the prediction functions. As a result, we show formally that simple, fully transparent algorithms can only be optimal when the principal's first-best solution is easier to describe in simple terms than is the difference between the agent's and the principal's preferred prediction functions. In the above examples, this would usually require that the misalignment in preferences is very large, or that the loss from simple algorithms is small.

As an alternative, we consider second-best regulation that allows for complex predictions and only regulates based on their simple explanations. For example, a lender may still be allowed to use fully complex deep-learning algorithms to predict repayment, but has to choose among models for which a simple explanation indicates alignment with the regulator's risk preferences. While the principal will not be able to align preference optimally in this case, this solution can provide a more efficient trade-off between efficiency and alignment than simple algorithms, unless the misalignment is large or hard to detect by the explainer.

We also show that it matters which specific model explanations regulation is based on. Regulation based in an explaination that focuses on preserving the most information about the average behavior of the prediction function – we term this the "agnostic explainer" – is generally inefficient. This approach is the focus of many available explainer tools, and we argue that we can improve over it in our context. Instead, optimal regulation regulates based on features related to preference misalignment. Intuitively, this "targeted explainer" inspects parts of the algorithm function that are most likely to reflect the preference misalignment. This targeted regulation can achieve high utility for the principal as long as the difference in agent and principal goals can be captured well by some low-dimensional representation. Improved explainer tools for the regulation of algorithms in critical applications should therefore be specific to the context and nature of preference misalignment. For example, if our concern is unfair hiring decisions, then regulation should capture features that are particularly related to differences across protected groups. If group membership itself is observable, then the optimal explainer in this case focuses on group membership itself as in Kleinberg et al. (2018), but our characterization extends to cases when only correlates of membership are observable. If we are concerned with a bank taking on loans that are especially high-risk in a stress-test scenario, then our description of complex credit scores should focus on features that are particularly likely to identify risk in that scenario. And an insurance company may want to audit a hospital's prediction model specifically based on features that correlate with over-testing.

We apply our framework for overseeing complex algorithms to the regulation of unsecured consumer lending in an empirical case study using large-scale credit bureau data. We consider a lender who builds a credit-scoring function using boosted trees from over 500 features on a training dataset of over 200,000 borrowers. We then consider two types of regulators overseeing the lender. A first regulator is concerned with the fairness of credit scores, and has a preference for low disparate impact across protected groups. A second regulator has different risk preferences from the lender, and wants the lender to deploy a credit score that predicts default well in a bad state of the economy. In both cases, the regulator has to rely on simple descriptions of the complex credit score in terms of a few key variables whose role is measured by a linear regression. We then compare different policy options and explainers on a hold-out dataset.

Across both applications in our empirical exercise, we document that well-regulated complex models improve over simple, fully transparent solutions. Restricting the lender to a simple linear regression with a few key variables leads to suboptimal fit and moderate disparate impact in the first application. Here, moving to complex models with simple explainer constraints can improve fit, while also reducing disparities across groups. Similarly, complex models can improve efficiency across all states of the economy in the risk application. Our empirical results therefore provide an example for our theoretical finding that the cost of full transparency may be too large when effective alternatives in the form of simple model descriptions are available.

Our empirical example also documents the value of targeted regulation over agnostic explainers. While either approach offers a Pareto improvement over restrictions to very simple models, the best outcomes for both regulator and lender are achieved with a targeted explainer, which focuses on those aspects of the credit score that are particularly relevant for aligning preferences, rather than with an agnostic explainer, which focuses on those covariates that best summarize the credit scoring function on average. In the case of disparate impact, the targeted explainer focuses particularly on covariates that are most related to differences across groups. With different risk preferences, the targeted explainer considers those variables that are indicative of changes in repayment behavior across states of the economy. The benefit of improved regulation is best illustrated by its impact on disparate impact in approval rates. While an unconstrained lender would be 4.5 percentage points more likely to approve non-minority applicants, this difference is reduced to 0.9 percentage points with an agnostic explainer, and to close to zero when the explainer is targeted to disparate impact. Despite the explanation constraint, the performance of the credit score as a risk predictor remains good. This stands in contrast with restricting the lender to a very simple model, which has considerably worse prediction fit while reducing disparate impact by only 1.0 percentage points.

**Related literature and contributions.** Our work contributes to a nascent literature that studies algorithmic decision-making (e.g. Athey, Bryan, and Gans, 2020) and how to regulate it. Most work in this area has focused on trade-offs between algorithmic fairness and performance: Gillis and Spiess (2019), Coston, Rambachan, and Chouldechova (2021), and Gillis (2021) examine the design and limits of algorithmic fairness audits; Dwork et al.

(2012), Corbett-Davies et al. (2017), Corbett-Davies and Goel (2018), and Yang and Dobbie (2020) consider ex-ante (i.e., pre-deployment) fairness constraints for algorithms; Liang, Lu, and Mu (2023) characterize how varying the information available to an algorithm can trade off between fairness and model performance. These trade-offs are sometimes understood as tracing out a fairness-vs.-accuracy Pareto frontier (Menon and Williamson, 2018; Little, Weylandt, and Allen, 2022; Liang, Lu, and Mu, 2023; Meursault et al., 2022), such that the role of social planner can be seen as choosing a preferred outcome on the frontier (Kearns and Roth, 2019).[1] Most related to our approach, Rambachan et al. (2020) study the regulation of algorithmic fairness in a principal-agent framework, where a social planner (principal) may not be able to achieve outcomes on a first-best frontier, and where second-best feasibility becomes relevant. Related concerns about the ability of a planner to interpret an algorithm have prompted calls for lower model complexity (Rudin, 2019), debates about model interprability (Doshi-Velez and Kim, 2017; Lipton, 2018), and analyses of optimal algorithmic transparency (Sun, 2021).

A parallel literature examines limits to the efficiency of algorithmic decisions, for example due to algorithmic bias (e.g. Lambrecht and Tucker, 2019; Arnold, Dobbie, and Hull, 2022; Fuster et al., 2022). This work has asked whether such biases are worse than their analogs from simpler or non-algorithmic decisionmaking (e.g. Cowgill and Tucker, 2017; Agrawal, Gans, and Goldfarb, 2019; Chan, Gentzkow, and Yu, 2022; Arnold, Dobbie, and Hull, 2022), whether algorithms necessarily inherit biases from human-generated training data (e.g. Rambachan and Roth, 2019; Cowgill and Tucker, 2019; Barocas and Selbst, 2016; Angelova, Dobbie, and Yang, 2023), and whether algorithms are able to learn over time about disadvantaged groups that the algorithm initially evaluates poorly (Li, Raymond, and Bergman, 2020). While often not expressed in a principal-agent framework, these concerns about algorithmic bias closely parallel the agency conflicts we study: an algorithm designer concerned about bias in her algorithm can use the same regulatory approach we provide, here viewing her algorithm as her agent.

Relative to these literatures, we make three contributions. First, we offer a framework that nests many types of potential incentive misalignment between a developer of an algorithm and social planner, or between any algorithmic agent and a principal: these include a broad set of distributional objectives and fairness concerns, as well as, for example, diverging risk preferences. Second, existing analyses of algorithmic audits and other algorithmic regulation often assume that disclosure of all underlying algorithmic inputs (data, training procedure, and decision rule) is possible. We study a world in which regulators will have access only to parts of this information, for example a simplified representation of the credit scoring model. Given the complexity of machine learning and artificial intelligence tools, and potential limitations on the technical or legal reach of regulators, it is important to study optimal algorithmic regulation under informational constraints. Third, we provide empirical validation for our theoretical results in a real-world, economically important setting, where we examine the regulation of consumer credit scoring. As in our Theorem 1, this empirical application shows that both a regulator and a lender can achieve privately preferred outcomes when the lender is permitted to use complex algorithmic credit scoring, rather than being constrained to simple and explainable models.

From an empirical perspective, we also contribute to a body of work on the role of default prediction models in US consumer finance. Most of this work explores properties of these models and their benefits, for example through overcoming adverse selection (Einav, Jenkins, and Levin, 2013; Adams, Einav, and Levin, 2009), deterring moral hazard (Chatterjee et al., 2020), and facilitating loan securitization (Keys, Seru, and Vig, 2012; Keys et al., 2010). Related work also warns that algorithmic underwriting can shape disparities in credit misallocation (Blattner and

[1] See also critical discussions of existing approaches to fairness in Kasy and Abebe (2021) and Kasy (2023), and the broader discussion of whether algorithms and artificial intelligence are aligned with societal goals in Korinek and Balwit (2022).

Nelson, 2022), reduce loan approval rates for disadvantaged groups (Fuster et al., 2022), and perpetuate cross-group disparities in loan terms (Bartlett et al., 2019). These concerns motivate our work to study optimal algorithmic regulation and highlight some of the sources of preference misalignment we study in our theoretical framework.

We differ from the literature on optimal *public* disclosure in the financial system (e.g. Goldstein and Leitner, 2017; Faria-e Castro, Martinez, and Philippon, 2017; Williams, 2017; Judge, 2020), and the literature on firm disclosure more broadly (e.g. Leuz and Verrecchia, 2000; Greenstone, Oyer, and Vissing-Jorgensen, 2006), as we focus on private disclosure to a regulator when decisions are automated and based on complex risk prediction algorithms. We assume there are (technical, practical, or legal) limitations on the amount of information the regulator can obtain about the algorithms and ask what optimal regulation looks like given these constraints. This approach differs from the existing literature which assumes that regulators can exercise choice over how much information to request.

More broadly, we add to a growing literature in computer science that studies algorithmic audits and derives specific explainability techniques from axioms about their deployment-agnostic properties (e.g. Bhatt et al., 2020; Carvalho, Pereira, and Cardoso, 2019; Chen et al., 2018; Doshi-Velez and Kim, 2017; Guidotti et al., 2018; Hashemi and Fathi, 2020; Lundberg and Lee, 2017; Murdoch et al., 2019; Ribeiro, Singh, and Guestrin, 2016). In particular, Lakkaraju and Bastani (2020), Slack et al. (2020), and Lakkaraju et al. (2019) study the limitations of post-hoc explanation tools in providing useful and accurate descriptions of the underlying models, and show that simple explanations can be inadequate in distinguishing relevant model behavior. Relative to these contributions, we show that the optimal regulatory design for algorithms with partial information depends on the nature of preference misalignment that motivates regulation. In other words, we highlight that explaining or interpreting a model inherently requires an understanding of the objectives of that explanation or interpretation, while purely technical or axiomatic approaches may miss important welfare-relevant consequences of model behavior. We also highlight some limitations of recent debates around the interpretability and explainability of prediction models. Embracing our utility optimization framework, we show that requiring a model to be fully explainable or interpretable can be misguided since it may force an agent to sacrifice model flexibility in ways that reduce rather than increase welfare.

In the long-standing literature on delegation under moral hazard, our setup can be viewed as a multi-tasking problem where each dimension of the agent's scoring rule is a separate hidden action. In contrast with the literature's traditional focus on how to use heterogeneously noisy signals of these hidden actions in an incentive scheme (Holmstrom and Milgrom, 1991, 1994), we study the principal's choice over which dimensions of these hidden actions – or which low-dimensional representation of them – she wants to observe in an audit. One closely related finding to ours is Baker (1992), which studies optimal incentive schemes under incentive misalignment over multiple actions, though Baker (1992) takes the information structure as given rather than chosen by the principal. Alternatively, our setting can also be viewed as a delegation problem (Holmström, 1984) that considers constaints on the set of actions allowed to the agent (Melumad and Shibano, 1991; Alonso and Matouschek, 2008; Frankel, 2014). Related work on principals designing the information structure is also found in the Bayesian persuasion literature (Kamenica and Gentzkow, 2011), though our setting is one of monitoring by the principal rather than of persuasion.

**Structure of this article.** The remaining article is organized as follows: Section II sets up our model and presents the main theoretical results. Section III lays out our empirical implementation. Section IV concludes.

# II. A Model of Regulation with Complexity Constraints

In this section, we model the regulation of algorithms as a two-player game between a principal and an agent. The principal delegates the choice of a prediction function to the agent. The agent receives a training signal and chooses the prediction function subject to constraints set by the principal. The principal can not fully observe the potentially complex prediction function chosen by the agent, and instead has to rely on a simple description. The preferences of agent and principal are not necessarily aligned.

Based on this model, we consider trade-offs between complexity and oversight. We study two types of policies and show how they affect equilibrium outcomes. First, we consider the case where there are ex-ante restrictions on algorithms, which force the agent to use only simple prediction functions that the principal can understand completely. Alternatively, we consider the case in which the agent chooses among complex algorithms and the principal restricts the agent's choice based on simple descriptions only. We characterize the trade-off between these two policy tools – ex-ante restrictions vs. ex-post explanations. We then describe the optimal design of simple ex-post explanations of complex prediction functions.

## II.1 Model setup

A principal oversees the choice by an agent of prediction function $f \in \mathcal{F}$. The agent receives a training signal $\theta \in \Theta$, and chooses a prediction function $f$ to maximize utility $U^A_\theta(f)$. The principal can constrain the choice of prediction function $f$, and has a preference over prediction functions expressed by utility $U^P_\theta(f)$.

For concreteness, we assume that the prediction function represents a mapping $f : \mathcal{X} \to \mathbb{R}$ and that the training signal $\theta$ parametrizes a distribution $\mathrm{P}_\theta$ that implies a joint distribution over a response variable $y \in \mathcal{Y}$ and a covariate vector $x \in \mathcal{X}$. Throughout, we assume that the agent's preference can be expressed as an average gain $U^A_\theta(f) = \mathrm{E}_\theta[u(f(x), y)]$ from deploying the function $f$ on data $(y, x)$ that follows the distribution $\mathrm{P}_\theta$. As examples of the utility function $u(f(x), y)$, for continuous $y$ the agent may care about minimizing the mean-squared prediction error of $f(x)$, $u(f(x), y) = -(y - f(x))^2$. Or the agent may want to maximize profit $u(f(x), y) = \mathbb{1}(f(x) \geq \bar{f})(r\, y - c\,(1 - y))$ from classifying instances $x$ with binary outcome $y \in \{1, 0\}$ correctly according to a threshold $\bar{f}$, where $r$ is the return for a correctly classified $y = 1$ instance and $c$ the cost of misclassifying a $y = 0$ instance. As we lay out in the following applications, such a prediction $f(x)$ can correspond to a score that measures the aptitude of job applicants and the classification $\mathbb{1}(f(x) \geq \bar{f})$ to the decision whether to invite the applicant for an interview, where $r$ is the return to a successful interview and $c$ the cost of an unsuccessful one. Similarly, $f(x)$ can be a credit score that estimates the repayment probability of loan applicants used to determine who gets a loan, or a predicted incidence of a disease used to make testing decisions. In these cases, $u(f(x), y)$ would be the gain of the agent in classifying an individual instance with true response $y$ as $f(x)$, and $U^A_\theta(f)$ the overall benefit across instances. The principal's preference $U^P_\theta(f)$ may, for example, differ from the maximization of expected utility by distributional considerations, different risk preferences, or an emphasis on different target populations. In order to accommodate such cases, we assume that $\mathrm{P}_\theta$ may imply a distribution over additional random variables or the membership in a subgroup of the population.

**Application 1** (Disparate impact in job screening)**.** *A company delegates the screening of applicants to a manager. The manager aims to score applicants in a way that is most reflective of future work performance, while the company also cares about ensuring that the screening process does not discriminate against protected groups and that interviewers get to see a diverse group of applicants. Specifically, we assume that the manager chooses scores* $f(x)$ *to maximize expected utility* $U^A_\theta(f) = \mathrm{E}_\theta[u(f(x), y)]$*, while the company attaches a penalty to average differences between male and female applicants,* $U^P_\theta(f) = \mathrm{E}_\theta[u(f(x), y)] - \lambda(\mathrm{E}_\theta[f(x)|g{=}1] - \mathrm{E}_\theta[f(x)|g{=}0])$*, where* $g$ *is a binary indicator for group membership and* $g = 1$ *denotes male applicants.*[2]

**Application 2** (Varying risk preferences in credit provision)**.** *A financial regulator oversees the credit provision of a lender. The lender aims to maximize overall expected utility* $U^A_\theta(f) = \mathrm{E}_\theta[u(f(x), y)]$ *from credit scoring, while the regulator cares about the potential downside in a bad (or "low") state of the economy,* $U^P_\theta(f) = \mathrm{E}_\theta[u(f(x), y)|s{=}\mathit{low}]$*. In that rare low state, the regulator believes that the probability* $\mathrm{E}_\theta[y|x, s{=}\mathit{low}]$ *of repayment is lower than their repayment probability* $\mathrm{E}_\theta[y|x, s{=}\mathit{high}]$ *in the (more common) high state, leading to excess risk taken on by the lender. Here, we assume that the future state* $s$ *is not known at the time of credit scoring, so the different preferences of the lender and the regulator represent different trade-offs between the high and low states of the economy.*

**Application 3** (Changing target populations in medical testing)**.** *A hospital considers deploying an algorithmic screening tool to support testing decisions for its patients. The algorithm is optimized to maximize predictive power based on the vendor's baseline sample of patients,* $U^A_\theta(f) = \mathrm{E}_\theta[u(f(x), y)|d{=}\mathit{vendor}]$*. The hospital, however, cares about the performance of the screening tool on its specific group of patients,* $U^P_\theta(f) = \mathrm{E}_\theta[u(f(x), y)|d{=}\mathit{hospital}]$*. While the hospital's patients' health outcomes still follow the same conditional distribution* $y|x$*, they may have different demographics and medical histories* $x$ *than the baseline sample, and the hospital would therefore trade off net benefits across different patient groups differently than the vendor.*

In all three examples, the preferences $U^P_\theta$ for the principal and $U^A_\theta$ for the agent are only partially aligned, since the principal has distributional preferences (over disparate impact, or "fairness" in our discussion to follow), cares about a different distribution of outcomes $y$ (due to risk preferences, or more generally what we refer to below as "model shift"), or puts weight on different parts of the covariates $x$ (changes in the target population, or what we refer to below as "covariate shift").

Having set up and motivated misaligned preferences, we now model the oversight over algorithms as a standard delegation problem where we allow for the principal to impose restrictions on the agent's choice of prediction functions. Specifically, the principal first receives a signal $\pi$ in the form of a prior over the training signal $\theta \in \Theta$, chooses a restriction $\mathcal{F}_\pi \subseteq \mathcal{F}$ of functions $f$ the agent can choose from, and then the agent chooses a function $f \in \mathcal{F}_\pi$ upon learning the realization $\theta \sim \mathrm{P}_\pi$. This setup follows the classical delegation-set approach without transfers of Holmström (1984) as further developed in Melumad and Shibano (1991), Alonso and Matouschek (2008), and Frankel (2014).

To this standard model of delegation, we add a central concern in overseeing complex algorithms, namely that the functions $f \in \mathcal{F}$ may be too complex to be fully described to (or understood by) the principal. Instead of

[2] Disparate impact, as expressed here, is only one of multiple ways in which we could capture a concern about discrimination (e.g. Kleinberg, Mullainathan, and Raghavan, 2016; Chouldechova, 2017). Even for disparate impact, we could alternatively consider measures of conditional parity, use a squared penalty $\lambda(\mathrm{E}_\theta[f(x)|g{=}1] - \mathrm{E}_\theta[f(x)|g{=}0])^2$ to express a concern with larger average differences, and/or formulate disparities in terms of the implied classification decision. We opt for a simple linear formulation that allows seamless integration into our later theory, which can be understood as the Lagrange version of a restriction on $\mathrm{E}_\theta[f(x)|g{=}1] - \mathrm{E}_\theta[f(x)|g{=}0] \leq c$ that expresses a concern with the scoring of the protected group.

formulating restrictions on $f$, we therefore assume that the principal can only restrict functions based on simpler descriptions (or "explanations") $\mathrm{E}f \in \mathcal{E}$, so that the restrictions available to the principal take the form

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

for some $\mathcal{E}_\pi \subseteq \mathcal{E}$. Crucially, the space $\mathcal{E}$ of model descriptions is typically smaller than the full space of functions $\mathcal{F}$, in which case the mapping $\mathrm{E} : \mathcal{F} \to \mathcal{E}$ loses information because $f$ cannot be fully described. For now, we take the set $\mathcal{F}$ of ex-ante permissible prediction functions as well as the explanation technology $\mathrm{E}$ as given and later return to the implications of different choices of $\mathcal{F}$ and $\mathrm{E}$.

Simple explanations of complex algorithms may in practice correspond to simple proxy models that project a prediction function onto a few key variables, variable importance scores, or a few interpretable key statistics. For example, in our empirical example in Section III we approximate a complex credit score based on boosted trees with over 500 covariates by a simple linear regression on a few key regressors. Similarly, tools like SHAP (Lundberg and Lee, 2017) represent the main drivers of complex machine-learning predictions in terms of Shapley values associated with the contribution of individual covariates. For random forests, variable importance measures similarly provide a simple description of each covariate's role in improving predictions. In all of these cases, the simple descriptions represent an informative but incomplete summary of the output of complex algorithms.

We note that we use the term "explanation" and "algorithm" somewhat liberally here. Specifically, what we call "explanations" are ex-post descriptions of the (potentially complex) prediction function $f$ chosen by the agent, which we interpret as the output of a machine-learning algorithm. But we do not directly consider descriptions or explanations that extend from the function itself to how it was created by the underlying algorithm. We discuss such extensions in Section II.5 below.

Embedding the complexity constraint into our delegation model, we thus consider the following game:

1. The principal receives a signal $\pi$ in the form of a prior over the agent's training signal $\theta \in \Theta$, and chooses a restriction $\mathcal{E}_\pi$.
2. The agent receives a training signal $\theta \sim \mathrm{P}_\pi$, and chooses a function $f_\theta \in \mathcal{F}$ subject to $\mathrm{E}f \in \mathcal{E}_\pi$.

In this model, the agent chooses $f_\theta \in \mathcal{F}_\pi = \{f \in \mathcal{F}; \mathrm{E}f \in \mathcal{E}_\pi\}$ to maximize $U^A_\theta(f)$, while the principal chooses $\mathcal{E}_\pi$ to maximize the average expected utility $\mathrm{E}_\pi U^P_\theta(f_\theta)$. If the explanation $\mathrm{E}$ is exhaustive, such as in the case where $\mathcal{E} = \mathcal{F}$ and $\mathrm{E}$ is the identity, we recover a standard delegation game. Our main focus is instead on the case where $\mathrm{E} : \mathcal{F} \to \mathcal{E}$ maps complex functions to simple explanations, leading to a loss of information because multiple prediction functions $f$ are described by the same simple explanation $\mathrm{E}f$.

Before discussing concrete implementations of the simplicity constraint, we note an alternative framing of the delegation model. Instead of being motivated by an information asymmetry between principal and agent, the model can describe a world where the agent's technology chooses functions so complex the principal cannot fully discern them, or the information constraint may reflect concerns about revealing proprietary technology. In this case, the delegation problem remains applicable even when the principal has full information about $\theta$, and the principal chooses $\mathcal{E}_\theta$ to maximize $U^P_\theta(f_\theta)$. By focusing on this case of known $\theta$, we aim to separate the implications of constraints on the complexity from those of private information. Our general model remains applicable, however, to cases where both private information and asymmetries in the ability to comprehend fully complex prediction functions play a role. Such cases may be particularly relevant when not only computational technology but also training data are controlled by the agent.

## II.2 Linear explanations with known state and quadratic loss

Having set up a general model of delegation with complexity constraints, we now solve the game for a concrete implementation. Specifically, we consider the case of quadratic loss, simple linear explanations on a subset $S$ of covariates, and state $\theta$ known to the principal. For this case, we illustrate concrete solutions in an example. In Section III, we then apply the same setup in a large-scale empirical exercise.

First, we assume that the agent strives to minimize mean-squared error $\mathrm{E}_\theta[(y - f(x))^2]$ (that is, has utility $u(f(x), y) = -(y - f(x))^2$). This criterion represents a natural and frequently employed benchmark, which we see as a general-purpose proxy for prediction quality that allows us to derive explicit solutions.

**Assumption 1.** *The agent minimizes squared-error loss,* $u(f(x), y) = -(y - f(x))^2$.

Second, we assume that simple explanations are given by linear regressions of the predictions $f(x)$ on a set of a few key variables $S$. For now, we take that set as given, and discuss its optimal choice in Section II.4 below.

**Assumption 2.** *The explanation mapping* $\mathrm{E} : \mathbb{R}^{\mathcal{X}} \supseteq \mathcal{F} \to \mathcal{E} = \mathbb{R}^S$ *is a linear projection onto a set of* $|S| < \infty$ *covariates* $x_S \in \mathbb{R}^S$ *that are not collinear,* $\mathrm{E}f = \arg\min_{\beta \in \mathbb{R}^S} \mathrm{E}_\theta[(f(x) - x_S'\beta)^2]$.

Third, we assume that the agent chooses from (potentially complex) functions that include at least linear regression. Analogous to the explainers, we take this set $\mathcal{F}$ as given for now.

**Assumption 3.** *The function space* $\mathcal{F}$ *forms a real vector space that includes at least the linear functions* $f(x) = x_S'\beta$ *for* $\beta \in \mathbb{R}^S$.

Finally, we focus on the case where there is no information asymmetry about the data distribution. This assumption puts our attention instead on asymmetric information about complex prediction functions.

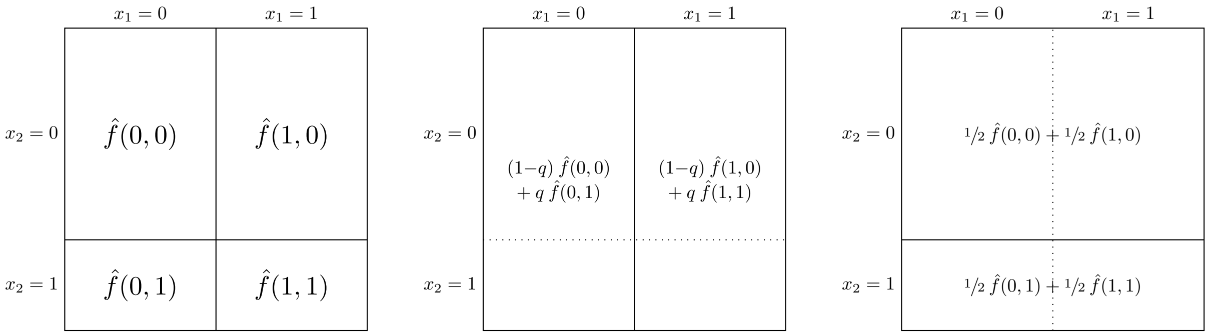
**Assumption 4.** *The principal learns the state* $\theta$*, that is,* $\pi$ *puts point mass on a specific* $\theta$.

These choices simplify the analysis substantially and allow us to make the role and specific structure of complexity constraints explicit. Concretely, the principal now applies constraints based on a simple linear proxy model with covariates $x_S$, subject to which the agent makes his choice. We can explicitly write the best choice of principal and agent in terms of linear projections.

**Proposition 1** (Constrained agent choice)**.** *Write* $f_\theta^A = \arg\min_{f \in \mathcal{F}} \mathrm{E}_\theta[(y - f(x))^2]$ *for the agent's first-best choice from* $\mathcal{F}$, $\beta_\theta^A = \mathrm{E}f_\theta^A$ *for its projection onto* $x_S$*, and* $r_\theta^A(x) = f_\theta^A(x) - x_S'\beta_\theta^A$ *for the remainder. Then an equilibrium of the game is given by the principal imposing the constraint* $\mathrm{E}f = \beta_\theta^P \in \mathbb{R}^S$ *and the agent choosing* $\hat{f}_\theta(x) = x_S'\beta_\theta^P + r_\theta^A(x)$*, where* $\beta_\theta^P$ *maximizes* $U_\theta^P(x \mapsto x_S'\beta^P + r_\theta^A(x))$ *over* $\beta^P$.

In other words, under the above assumptions, the agent follows the principal's instructions for the linear part of the prediction function corresponding to covariates $x_S$, and chooses the remaining part according to his first-best choice.

Before discussing the optimal choice of explainer variables, we apply the above result to a simple linear-regression example. In this example, the role of fully complex prediction functions is played by fully interacted linear regression on two binary covariates, while the role of simple explanations is represented by their projection onto only one of the covariates.


$x_1 = 0$
$x_1 = 1$
$x_2 = 0$
$x_2 = 1$
$\hat{f}(0,0)$
$\hat{f}(1,0)$
$\hat{f}(0,1)$
$\hat{f}(1,1)$
$x_1 = 0$
$x_1 = 1$
$x_2 = 0$
$x_2 = 1$
$(1-q)\,\hat{f}(0,0)$ $+\,q\,\hat{f}(0,1)$
$(1-q)\,\hat{f}(1,0)$ $+\,q\,\hat{f}(1,1)$
$x_1 = 0$
$x_1 = 1$
$x_2 = 0$
$x_2 = 1$
${}^1\!/_2\,\hat{f}(0,0) + {}^1\!/_2\,\hat{f}(1,0)$
${}^1\!/_2\,\hat{f}(0,1) + {}^1\!/_2\,\hat{f}(1,1)$

**(a)** Full function $\hat{f}$ chosen by the agent.

**(b)** Information retained by a simple explaination $\mathrm{E}_1\hat{f}$ that projects the complex function $\hat{f}$ linearly onto $x_1$ (and a constant).

**(c)** Information retained by a simple explaination $\mathrm{E}_2\hat{f}$ that projects the complex function $\hat{f}$ linearly onto $x_2$ (and a constant).

**Figure 1:** Illustration of the structure of a complex function $\hat{f}$ (left panel) as well as the information retained in simple explainers $\mathrm{E}\hat{f}$ (center and right panels) from the example. Each cell corresponds to a combination of the values of the two binary covariates $x_1$ and $x_2$, and the values in the cells or across two cells represent the information retained in each case.

**Example** (Fully interacted linear regression)**.** *For concreteness, we consider the case of predicting an outcome $y \in \mathbb{R}$ (which can be binary) from a pair of binary covariates $x_1, x_2$, with the distribution being independent across $x_1$ and $x_2$ with $\mathrm{E}_\theta[x_1] = {}^1\!/_2, \mathrm{E}_\theta[x_2] = q \in (0, {}^1\!/_2)$. In this world, we can without loss of generality write any (arbitrarily complex) conditional expectation of the outcomes as a fully interacted linear regression*

$$f_\theta(x) = \mathrm{E}_\theta[y|x] = \theta_0 + (x_1 - {}^1\!/_2)\,\theta_1 + (x_2 - q)\,\theta_2 + (x_1 - {}^1\!/_2)(x_2 - q)\,\theta_{12}$$

*since the covariates are both binary. However, we assume that the regulator can only understand projections of such functions onto one of the two covariates $x_1, x_2$ (and a constant $x_0$), $S = \{0,1\}$ or $S = \{0,2\}$ for the covariate space $\mathcal{X} = \{1\} \times \{1,0\}^2$. For $S = \{0,1\}$ as an example, for a function $\hat{f}(x) = \hat{\theta}_0 + (x_1 - {}^1\!/_2)\hat{\theta}_1 + (x_2 - q)\hat{\theta}_2 + (x_1 - {}^1\!/_2)(x_2 - q)\hat{\theta}_{12}$ the associated explainer $\mathrm{E}_1 : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^2$ yields $\mathrm{E}_1\hat{f} = \begin{pmatrix}\hat{\theta}_0 \\ \hat{\theta}_1\end{pmatrix}$. Figure 1 illustrates both the information content of the full function $\hat{f}$ (left panel) as well as the information retained in the simple explainers $\mathrm{E}_1\hat{f}$ (center panel) for the case $q = .3$. Assuming that the agent's choices are otherwise unconstrained ($\mathcal{F} = \mathbb{R}^{\mathcal{X}}$), this means that the principal determines $\hat{\theta}_0, \hat{\theta}_1$ according to her preference (say, $\beta^P_\theta$), and the agent chooses $\hat{\theta}_2, \hat{\theta}_{12}$ according to his (that is, $\theta_2, \theta_{12}$), to obtain*

$$\hat{f}_\theta(x) = \underbrace{\beta^P_{\theta,0} + (x_1 - {}^1\!/_2)\beta^P_{\theta,1}}_{\text{controlled by principal}} + \overbrace{(x_2 - q)\theta_2 + (x_1 - {}^1\!/_2)(x_2 - q)\theta_{12}}^{\text{controlled by agent}} = x'_S\beta^P_\theta + r^A_\theta(x).$$

*As in the general case of Proposition 1, the choice of the residual $r^A_\theta(x) = (x_2 - q)\theta_2 + (x_1 - {}^1\!/_2)(x_2 - q)\theta_{12}$ does not depend on the principal's choice.*

In the remaining part of this section, we now apply the result from the example to the specific applications introduced in Section II.1. In each application, we consider concrete choices of $x_1$ and $x_2$ in order to illustrate the structure of constrained solutions.

**Application 1** (Disparate impact in job screening, continuing from p. 8)**.** *We assume for concreteness that job* *applicants are scored based on whether they have a college degree ($x_1$) and whether they have any gaps in their resume ($x_2$) in order to predict future on-the-job performance $y$. If the manager only reports a simple description of the screening score in terms of college degree, then in equilibrium the company has full control over how a college degree affects screening, but the manager's algorithm has full flexibility over the marginal influence of gaps in the resume on applicant scoring. If gaps in the resume are correlated with gender $g$, then the resulting scores may still have excess disparate impact.*

**Application 2** (Varying risk preferences in credit provision, continuing from p. 8)**.** *We assume that the probability of repayment is estimated from two binary variables, past default ($x_1$) and whether the applicant has a home-equity line of credit (HELOC, $x_2$). For our illustration, we let having a HELOC be a positive indication of creditworthiness in the high state of the economy, but a negative indication in the low state ($\theta_2(high) > 0 > \theta_2(low)$). We also allow the overall level of risk (intercept) to vary with the state of the economy, so that we can write*

$$\mathrm{E}_\theta[y|x,s] = \theta_0(s) + (x_1 - {}^1\!/_2)\,\theta_1 + (x_2 - q)\,\theta_2(s) + (x_1 - {}^1\!/_2)(x_2 - q)\,\theta_{12}.$$

*We assume that the future state $s$ of the economy at the time of potential repayment is unknown at the time credit scoring happens, but that the distribution over $\theta = (\theta(high), \theta(low))$ is learned in the training stage. In this case, the lender wants to evaluate HELOC in terms of its average effect, while the regulator prefers that HELOC is counted only as a negative. An explanation in terms of past default helps align the overall level of risk, but does not affect how the lender's algorithm leverages information about HELOCs.*

**Application 3** (Changing target populations in medical testing, continuing from p. 8)**.** *The vendor's algorithm scores patients based on whether they have high blood pressure ($x_1$) and whether they have been diagnosed with a heart condition ($x_2$) to predict the incidence $y$ of a disease. A simple explanation in terms of high blood pressure does not restrict how the vendor's algorithm scores the incremental information from a heart condition.*

Across all applications, the simple explanation based on one of the two covariates gives the principal only partial control over the agent's algorithm. Depending on the nature of misalignment and the choice of explainer covariates, the choices are only partially aligned. While resulting prediction functions lead to better outcomes for the principal than not constraining the agent at all, they may still reflect substantial misalignment.

## II.3 Policy choice of simple models vs. simple explanations

Having laid out a simple delegation model as well as described its solution for quadratic loss and linear explainers, we now consider different implications of the rules of the game itself, and how these restrictions affect the equilibrium outcome of the game. We then provide formal results about when ex-ante restrictions to simple models can improve outcomes, and when oversight based on ex-post explanations dominates.

So far, we have taken the function class $\mathcal{F}$ and explainer mapping $\mathrm{E} : \mathcal{F} \to \mathcal{E}$ as given. In this section, we instead ask the ex-ante policy design question of how the function class $\mathcal{F}$ affects outcomes. Under the assumptions of Section II.2, we consider the difference in equilibrium outcomes between restricting $\mathcal{F}$ to be fully transparent, by which we mean an ex-ante restriction of $\mathcal{F}$ to simple linear functions on a subset of covariates, and leaving $\mathcal{F}$ unconstrained. In Section II.4 below, we then ask how the design of the explainer mapping $\mathrm{E} : \mathcal{F} \to \mathcal{E}$ affects equilibrium outcomes in the latter case, and solve for optimal explainers from the perspective of the principal.

Taken together, these two results allow us to characterize the optimal design of algorithmic regulation within our framework.

For this part of our analysis, we assume that principal and agent utilities take a similar form. To this end, we introduce notation that is helpful to express important sources of misalignment between principal and agent. We write $\mu_\theta$ for the probability measure over covariates $x$ implied by $\mathrm{P}_\theta$ and denote by $f_\theta(x) = \mathrm{E}_\theta[y|x]$ the (possibly infeasible) prediction function that minimizes mean-squared error $\mathrm{E}_\theta[(y - f(x))^2]$ if there are no constraints on the functions $f$. Then

$$U^A_\theta(f) = -\int_{\mathcal{X}} (f(x) - f_\theta(x))^2 \; \mathrm{d}\mu_\theta(x) + \text{const.}$$

with a constant part that does not depend on $f$, so the agent's utility is equivalent to minimizing $\int_{\mathcal{X}}(f(x) - f_\theta(x))^2 \; \mathrm{d}\mu_\theta(x)$. This formulation allows us to express misaligned utility in terms of deviations in target functions and covariate distributions.

**Assumption 5.** *The principal's preference is equivalent to minimizing* $\int_{\mathcal{X}}(f(x) - f^P_\theta(x))^2 \; \mathrm{d}\mu^P_\theta(x)$ *for some target* $f^P_\theta \in \mathbb{R}^{\mathcal{X}}$ *and some probability measure* $\mu^P_\theta$ *over* $\mathcal{X}$*.*

Hence, principal and agent utility may differ in the distribution over the target covariates and/or the target function. A first version of misalignment is given by different target populations, holding the target function fixed across principal and agent. Using terminology from machine learning, we call this variation across objectives "covariate shift."

**Definition 1** (Covariate shift)**.** *Principal and agent utility differ by* covariate shift *if* $f^P_\theta = f_\theta$ *but* $\mu^P_\theta \neq \mu_\theta$*.*

A second form of misalignment arises when populations are the same, but the agent cares about a different target function from that of the principal. We call this change "model shift", as only the preferred prediction function changes.

**Definition 2** (Model shift)**.** *Principal and agent utility differ by* model shift *if* $\mu^P_\theta = \mu_\theta$ *but* $f^P_\theta \neq f_\theta$*.*

In Proposition 2 below, we also consider distributional preferences, for which the principal cares about average differences in predictions across groups as in Application 1. There, we show that it can be interpreted as a special case of model shift.

We now contrast how two ex-ante policy options for the function class $\mathcal{F}$ and the explainer mapping $\mathrm{E}$ can help align choices through setting the rules of the game laid out in Section II.1. On the one extreme, we consider an ex-ante policy that restricts the function class $\mathcal{F}$ to functions that can be fully explained, so that $\mathcal{F} = \{x \mapsto x'_{S_1}\beta; \beta \in \mathbb{R}^{S_1}\} \cong \mathcal{E}$ for some set $S_1$ of covariates. In that case, the principal has full control over the agent's choice, ensuring full alignment. This policy option is one interpretation within our model of calls for fully transparent or explainable models.

On the other extreme, we consider an ex-ante policy that leaves the agent's choice fully unconstrained ($\mathcal{F} = \mathbb{R}^{\mathcal{X}}$) and only imposes constraints on the agent's choice via the explainer based on the covariates $x_{S_2}$. (Here, the explainer variables $x_{S_2}$ do not necessarily have to be the same as those in the ex-ante constrained case $x_{S_1}$.) In this case, alignment is only partial. This policy option connects to proposals that allow for fully complex functions, but subject these scores to targeted exams or require specific explanations for certification.

The following result describes the relative performance of full transparency and oversight based on ex-ante explainers for the specific cases of misalignment introduced above in terms of the principal's expected utility.

**Theorem 1** (Alignment through simplicity vs. alignment despite complexity)**.** *If misalignment stems from* covariate shift *with $\mu_\theta^P \ll \mu_\theta$, then choices from $\mathcal{F} = \mathbb{R}^{\mathcal{X}}$ are fully aligned, leading to higher principal and agent utility than any non-trivial ex-ante complexity constraint. If misalignment stems from* model shift *then choices from $\mathcal{F} = \mathbb{R}^{\mathcal{X}}$ subject to an explanation constraint with explainers $S_2$ lead to a higher expected principal utility outcome than ex-ante constraints with explainer covariates $S_1$ if and only if*

$$\min_{\beta \in \mathbb{R}^{S_2}} \int_{\mathcal{X}} (f_\theta(x) - f_\theta^P(x) - x'_{S_2}\beta)^2 \ \mathrm{d}\mu_\theta(x) < \min_{\beta \in \mathbb{R}^{S_1}} \int_{\mathcal{X}} (f_\theta^P(x) - x'_{S_1}\beta)^2 \ \mathrm{d}\mu_\theta(x),$$

*that is, whenever the difference between targets is easier to explain than the principal's target itself.*

In particular, considerations around different target populations only lead to misalignment for simple functions, but not when the agent is allowed to choose from a flexible function class. (Here, we ignore additional statistical considerations around resolving bias–variance trade-offs efficiently from limited data, in which case differential weighting of subpopulations may matter.) With model shift, on the other hand, ex-ante simplicity restrictions can be beneficial to the principal when the difference between preferences is harder to explain than the first-best choice of the principal, but are also inefficient otherwise.

Covariate shift and model shift represent two specific sources of misalignment. We note that the latter of these comprises another type of misalignment between principal and agent where the principal cares about disparate impact, as in the hiring case of Application 1 (and elaborated on in Footnote 2). Such a preference is equivalent to model shift with a specific target. As a consequence, we can apply our previous result to the case of distributional preferences of the principal.

**Proposition 2** (Distributional preferences)**.** *Assume that the principal's utility differs by the agent's utility by a cost associated with lower average predictions of the minority group,*

$$U_\theta^P(f) = \mathrm{E}_\theta[u(f(x), y)] - \lambda(\mathrm{E}_\theta[f(x)|g{=}1] - \mathrm{E}_\theta[f(x)|g{=}0])$$

*for $g$ a binary indicator of group status (and $g = 1$ the majority group). Writing $g_\theta(x) = \mathrm{E}_\theta[g|x]$ for the majority fraction at $x$ and $\bar{g}_\theta = \mathrm{E}_\theta[g]$ for the overall fraction of $g = 1$, then the principal's utility is equivalent to model shift with $f_\theta^P(x) = f_\theta(x) - \frac{\lambda}{2\bar{g}_\theta(1-\bar{g}_\theta)}(g_\theta(x) - \bar{g}_\theta)$. In particular, assuming that the explainer variables $x_{S_2}$ include an intercept, choices from $\mathcal{F} = \mathbb{R}^{\mathcal{X}}$ subject to an explanation constraint with explainers $S_2$ lead to a higher expected principal utility outcome than ex-ante constraints to only using covariates $S_1$ if and only if*

$$\frac{\lambda}{2\bar{g}_\theta(1-\bar{g}_\theta)} \min_{\beta \in \mathbb{R}^{S_2}} \int_{\mathcal{X}} (g_\theta(x) - x'_{S_2}\beta)^2 \ \mathrm{d}\mu_\theta(x) < \min_{\beta \in \mathbb{R}^{S_1}} \int_{\mathcal{X}} (f_\theta^P(x) - x'_{S_1}\beta)^2 \ \mathrm{d}\mu_\theta(x),$$

*that is, whenever a $\frac{\lambda}{2\bar{g}_\theta(1-\bar{g}_\theta)}$ fraction of the error in explaining group membership probabilities is smaller than the error in explaining the principal's target itself.*

The result shows that distributional preferences are equivalent to model shift where the difference between principal and agent targets are given by an offset related to the relative fraction of majority and minority instances. Equivalently, the principal prefers that predictions for the minority group are higher by a given offset than those of comparable majority instances. The size of the offset depends on the relative size of groups as well as the strength of the disparate-impact preference.

This result is a direct generalization of a result in Kleinberg et al. (2019) that considers optimal oversight over algorithms in a world where group identity $g$ is available to the agent. In this case, their result shows that optimal regulation enforces averages $\mathrm{E}[f(x)|g]$ across groups, while leaving predictions otherwise unconstrained. If $g$ is available as an explainer variable in $S_2$, then this result directly follows from the proposition. Indeed, in this case, the expected loss on the left is zero, and ex-post explanations generically lead to higher expected utility than any ex-ante restriction. Note, however, that our result is more general and does not require that the agent have access to group indicators, or that the final rule be allowed to depend on them.

Before discussing optimal choices of the covariate sets $S_1$ and $S_2$ (i.e., included regressors for the simple model and explainer variables for the complex model) we study the general result in the illustration from our example in Section II.2, and apply it to the three cases laid out in Section II.1. We start by comparing simple to complex solutions in the example.

**Example** (Fully interacted linear regression, continuing from p. 11). *We compare choices made by the agent from all functions $\mathbb{R}^{\mathcal{X}}$ subject to an explanation fixed by the principal to simple functions that are fully transparent. For $S_2 = \{0, 1\}$ as above, in the former case the resulting function is $\hat{f}_\theta(x) = \beta^P_{\theta,0} + (x_1 - 1/2)\beta^P_{\theta,1} + (x_2 - q)\theta_2 + (x_1 - 1/2)(x_2 - q)\theta_{12}$. If the agent is instead constrained to simple functions (here, $S_1 = \{0, 1\}$, meaning a linear function regression on $x_1$ with intercept), the resulting function is $\hat{f}_\theta(x) = \beta^P_{\theta,0} + (x_1 - 1/2)\beta^P_{\theta,1}$, which the principal can now fully control.*

We next discuss the implications of the example across the three applications, which represent the three sources of misalignment we discuss above: distributional preferences in the job screening case, model shift for lending, and covariate shift in the medical application.

**Application 1** (Disparate impact in job screening, continuing from p. 8). *We consider the choice of the company whether to restrict the hiring manager to score applicants only based on having a college degree ($x_1$). In that case, choices are aligned, but scoring may be inefficienct since it ignores the predictive information of gaps in the resume ($x_2$). If the misalignment over gaps in the resume is limited and the company's dislike of disparate impact is not too high, then a manager's algorithm that is allowed to be complex, but constrained in terms of its intercept and effect of college degrees, may still be preferred by the company.*

**Application 2** (Varying risk preferences in credit provision, continuing from p. 8). *The regulator considers whether to allow the lender's algorithm to use HELOC ($x_2$), given the misalignment over that variable. If the regulator allows for complex algorithms that are only constrained in terms of overall risk level and how they use past default ($x_1$) via the explainer $\mathrm{E}_1$, then choices over HELOC would be misaligned. From the regulator's perspective, this additional flexibility would only make sense if the benefit of including the interaction term outweighs the cost of the misaligned coefficient on HELOC.*

**Application 3** (Changing target populations in medical testing, continuing from p. 8). *If hospital and vendor only differ in the distribution of patients they care about, but not in their view on the incidence of the medical condition given available patient information, then unconstrained choices are not misaligned, no matter whether an explanation constraint is enforced or not. If, however, the vendor is barred from leveraging diagnosed heart conditions ($x_2$) in his algorithm, then choices are not only inefficient, but also potentially misaligned. This is because different joint distributions of high blood pressure ($x_1$) and heart condition across target populations imply* *different adjustments for the included coefficients $\theta_0, \theta_1$ to account for the excluded coefficients $\theta_2, \theta_{12}$ (equivalent to the dependence of omitted variable bias on the covariate distribution).*

For the choice between ex-ante restrictions and flexible prediction functions with ex-post explanations, this result makes a stark prediction: unless misalignment in targets is of at least the same order of magnitude and complexity as the principal's target function itself, complexity restrictions are not optimal for the principal, and will hurt both agent and principal utility when we compare restrictions and explanations using the same covariates. In the specific case of a preference for parity across groups, complexity restrictions are suboptimal unless the preferences $\lambda$ against disparity is severe *and* group membership is hard to explain. In principle, a combination of both policy levers can further improve outcomes, but may be unrealistic in practice.

## II.4 Optimal model explanations

In the previous sections, we provided a characterization of when oversight of flexible functions based on ex-post explanations can improve over ex-ante restrictions that limit the agent to explainable models. We now discuss the optimal choice of explainers that ensure second-best outcomes from the perspective of the principal. This addition can be interpreted as extending our model in Section II.1 by an initial stage at which the principal decides on the form of the required explanation. Practically, this means that the principal also specifies the set of covariates to include in the simple description of the complex model.

A standard approach to explaining prediction choices by the agent would be to leverage an explanation mapping that can capture the maximal amount of information about the agent's prediction function, which we call the *misalignment-agnostic explainer*. In our specific setting, this would be the mapping $\mathrm{E}_0 f = \arg\min_\beta \int_{\mathcal{X}} (f(x) - x'_{S_0}\beta)^2 \, \mathrm{d}\mu_\theta(x)$ with $S_0 = \min_{S \in \mathcal{S}} \min_\beta \int_{\mathcal{X}} (f_\theta(x) - x'_S\beta)^2 \, \mathrm{d}\mu_\theta(x)$ chosen to minimize the discrepancy between explained and unexplained part of the agent's first-best prediction function $f_\theta$, subject to the complexity constraint $|S| \leq k$.

However, this agnostic explainer is not generally optimal. Instead of relying on a set of covariates that best explain the average behavior of the prediction function, the following result shows that the principal can do better by requiring an explanation optimized for the specific source of preference misalignment.

**Theorem 2** (Targeted explainer)**.** *If misalignment stems from* model shift *and the function class is unconstrained, $\mathcal{F} = \mathbb{R}^{\mathcal{X}}$, then the optimal explainer solves*

$$S^* = \min_{S \in \mathcal{S}} \min_\beta \int_{\mathcal{X}} (f_\theta(x) - f_\theta^P(x) - x'_S\beta)^2 \, \mathrm{d}\mu_\theta(x),$$

*that is, the optimal explainer is* targeted *to the misalignment between principal and agent.*

Here, we assume that the principal knows the true distribution $\mathrm{P}_\theta$ of the data when deciding on explainer variables. The results in this section generalize to the case where the principal only has some belief about $\theta$ when specifying required explainer variables, such as the prior $\pi$ from Section II.1 or a belief coming from a less informative hyper-prior. In this case, the minimization is over the expectation $\mathrm{E} \min_\beta \int_{\mathcal{X}} (f_\theta(x) - f_\theta^P(x) - x'_S\beta)^2 \, \mathrm{d}\mu_\theta(x)$ with respect to that belief.

The targeted explainer takes a particularly simple form when the source of model shift stems from different distributional preferences as in Proposition 2. Specifically, the optimal explainer can be obtained from a simple prediction problem.

**Corollary 1** (Targeted explainer for disparate impact)**.** *If misalignment stems from* distributional preferences *and the function class is unconstrained, $\mathcal{F} = \mathbb{R}^{\mathcal{X}}$, then the optimal explainer solves*

$$S^* = \min_{S \in \mathcal{S}} \min_{\beta} \mathrm{E}_\theta[(g - x'_S \beta)^2],$$

*that is, the optimal explainer is a best explainer of the majority indicator.*

In other words, the covariates that best align choices with the principal's distributional preferences are those indicative of group membership, rather than those most predictive of the outcome itself. If group membership itself is available to the agent, then we recover the result from Kleinberg et al. (2019): in this case, the principal should regulate the algorithm based on group-specific averages $\mathrm{E}[f(x)|g]$, corresponding to an explainer that describes predictions by a regression on $g$, and leave the algorithm otherwise unconstrained. However, our result also extends to cases where the agent does not have access to protected group characteristics or is barred from using them, in which case the optimal explainer amounts to describing the algorithm in terms of features of the data that are related to differences across groups.

Equipped with the notion of a targeted explainer, we can now compare the ex-ante policy choice from Section II.3 between restricting models ex-ante to some set $S_1$ or leaving models unrestricted and choosing explanation covariates $S_2$, where we assume that the respective sets are chosen optimally from the perspective of the principal.

**Corollary 2** (Optimal regulation)**.** *If misalignment stems from* covariate shift *with $\mu_\theta^P \ll \mu_\theta$, then leaving the function class ex-ante unconstrained is optimal from both agent and regulator perspectives, and explainer choices are inconsequential. If misalignment stems from* model shift *and restrictions are chosen optimally, then leaving the function class ex-ante unconstrained is preferred by the principal if and only if*

$$\min_{S_2 \in \mathcal{S}} \min_{\beta} \int_{\mathcal{X}} (f_\theta(x) - f_\theta^P(x) - x'_{S_2}\beta)^2 \; \mathrm{d}\mu_\theta(x) < \min_{S_1 \in \mathcal{S}} \min_{\beta} \int_{\mathcal{X}} (f_\theta^P(x) - x'_{S_1}\beta)^2 \; \mathrm{d}\mu_\theta(x),$$

*where the targeted explainer from Corollary 1 is optimal in the unconstrained case.*

We close our discussion of the model results by applying them back to the three motivating applications within our simple linear example, in which we contrast the agnostic and targeted explainers.

**Example** (Fully interacted linear regression, continuing from p. 11)**.** *We continue our simple example with two binary covariates $x_1, x_2$, and we assume that the cost of excluding each of the covariates is ex-ante equal. In this case, the misalignment-agnostic explainer focuses on the projection* $\mathrm{E}_1$ *onto $x_1$ (and an intercept), since $x_1$ varies more than $x_2$ and thus provides more information about the agent's choice. However, if choices are aligned along $x_1$, but not along $x_2$, then the targeted explainer instead regresses the agent's choice onto $x_2$ (and an intercept). The resulting explainer* $\mathrm{E}_2$ *is represented in the right panel of Figure 1.*

Within our applications, we illustrate how targeted and agnostic explainers may differ in practice. In Section III, we provide an analogous illustration from an empirical application to consumer lending.

**Application 1** (Disparate impact in job screening, continuing from p. 8)**.** *Assuming that choices are misaligned over the use of gaps in the resume ($x_2$) only, then regulating based on the explainer* $\mathrm{E}_2$ *that is targeted toward* *the misalignment aligns choices. The targeted explainer* $\mathrm{E}_2$ *also has an intuitive interpretation in this case: it describes the role of those covariates that are most related to differences across groups. On the other hand, the misalignment-agnostic explainer* $\mathrm{E}_1$ *(based on college degree,* $x_1$*) is better at capturing the overall behavior of the scoring function, but worse at capturing the aspect that is relevant for disparate impact with respect to gender.*

**Application 2** (Varying risk preferences in credit provision, continuing from p. 8)**.** *The targeted explainer* $\mathrm{E}_2$ *fully captures the source of misalignment, which here is the overall level of risk as well as the role of HELOCs in credit scoring (*$x_2$*). The impact of past default (and the interaction of both) is left unrestricted, which happens to be optimal in this case. The targeted explainer again permits an intuitive interpretation: it captures those features that are most related to differences across economic states* $s$*.*

**Application 3** (Changing target populations in medical testing, continuing from p. 8)**.** *Since choices are aligned without any restrictions in the case of different target populations, the choice of explainer covariates does not affect the outcome. Hence, there is only a cost to simplicity in this example. However, if the hospital were to force the vendor to exclude one of the covariates, then keeping the variable that is most important in the patient distribution of the hospital would be the better among bad options.*

While these illustrations are extreme in that they allow for perfect alignment, we explore empirically below how the trade-off between simple and complex functions as well as agnostic and targeted explainers play out in data from the context of consumer lending, both for disparate-impact concerns as in Application 1 and for divergent risk preferences as in Application 2.

## II.5 Variants and extensions

In our model, we have so far illustrated trade-offs in algorithmic regulation in a stylized setup that focuses on tractability. Here, we discuss extensions.

**Double-selection explainers for disparate impact.** In Application 1 and Proposition 2, we present results on aligning choices when the source of misalignment is the principal's dislike for disparate impact. The targeted explainer takes an intuitive form: its features represent the set of covariates that best describes group membership. This specific result is driven by the linear form of the penalty in the principal's utility function (see also Footnote 2 for a discussion). Different forms of the penalty, such as a squared penalty, may affect the optimal design of the explainer. Specifically, it could be optimal to choose those features that are important for *both* group membership *and* the outcome model. This has a connection to the literature on double selection in machine learning (and specifically Belloni, Chernozhukov, and Hansen, 2014), which emphasizes the selection of all potential confounders that are related to the assignment of treatment *or* the outcome in order to avoid selection mistakes.

**Audits based on realized outcomes.** We assume in our model that restrictions are set on prediction functions before they are deployed and before their consequences are realized, such as employment, credit, and testing decisions. Once the actual decisions and their consequences are measured, audits may effectively uncover *realized* misaligned choices. For example, managers may lose their bonuses for unequal hiring decisions, banks may be fined for experiencing excess defaults, and vendors may be punished for misallocated tests. While such information is often only available with a delay, it could improve adherence at the training stage. Optimal regulation would then combine restrictions on model explanations that can be verified before deployment, and fines for bad outcomes. However, as long as outcomes are uncertain and agents risk averse, an optimal mix of regulatory instruments would include the use of explainers, since regulating only based on ex-post outcomes in those cases cannot distinguish well between bad choices and bad circumstances. As a result, consequence-based regulation may be slow, have limited effect, or lead to overly conservative agent choices. At the same time, the availability of realized consequences may further limit the need for ex-ante complexity constraints.

**Direct regulation of binary decisions.** In our model above, we assume that preferences are over real-valued predictions, such as a hiring manager's assessment, a lender's credit score, or a health provider's risk assessment. We may instead apply our model directly to the decisions taken by the agent, and focus on binary policies that determine who is hired, who receives credit, or who is tested. In those cases, we would expect similar results as in the more tractable continuous case, with the inefficiency of ex-ante complexity restrictions depending on cases in a neighborhood of the agent's decision threshold.

**Ex-post vs. process-based descriptions.** Our discussion of algorithmic regulation has focused on describing resulting prediction functions, rather than regulating the procedure by which training data and machine-learning pipelines yield prediction functions. Considering the entire process beyond ex-post descriptions of resulting decision functions (as, e.g., considered by Kleinberg et al., 2018) may be a necessity when it is relevant for legal and regulatory considerations and could provide additional tools for making algorithmic decisions more transparent without having to rely on inefficient restrictions to complexity.

# III. Empirical Implementation

We build an empirical counterpart to our model using machine-learning predictors subject to constraints. In this large-scale empirical illustration, we evaluate the different regulatory approaches from our model in the context of algorithmic credit scoring. We focus on two cases of preference misalignment between a financial regulator (principal) and a lender (agent). In the first case, the regulator has a taste for more equal outcomes across racial or ethnic groups – or a taste for less disparate impact. In the second case, the regulator cares about model fit in an economic downturn with elevated default rates, similar to a stress-test scenario. Across both cases, we demonstrate the benefit of allowing for complex prediction functions and the value of using targeted explainers.

## III.1 Data setup

Our base dataset is a random sample of about 300,000 credit bureau files that have a newly opened credit card in 2012. We focus on credit cards since they are a widely used credit product for which algorithmic underwriting is already in use by some providers. This sample is drawn from the larger credit bureau panel constructed in Blattner and Nelson (2022), which used a probability-weighted sampling strategy to ensure sufficient representation from minority groups; this feature is important in our disparate impact empirical application. We draw another random sample of 50,000 credit bureau files from the same credit bureau panel for consumers who were rejected for a recent credit card application; these data are useful for our empirical application focused on misaligned preferences over borrower risk. We split the data into a training dataset (80% of the sample) and a test dataset (20% of the sample).

Our main prediction outcome is credit card default of any severity up to 24 months after the card was opened. We capture this outcome by a repayment indicator $y \in \{1, 0\}$, where $y = 1$ denotes repayment and $y = 0$ denotes default. The dataset contains several hundred credit report attributes, including detailed default histories, debt balances, utilization, credit report inquiries, and current and past debt obligations. For each variable, we create outlier and missing value flags and include these in the set of variables to predict default. We transform all balance variables into logs. We obtain a total of 518 variables. Table 3 shows summary statistics for repayment as well as the ten variables with the highest importance value across a set of baseline prediction models of credit card default trained using XGBoost, random forest, elastic net, and neural net algorithms (see Table 4 for performance summaries). For our prediction exercises, we normalize all of the credit bureau variables so that in the main dataset of 300,000 opened credit cards, the variables are centered at zero with standard deviation equal to one.

In addition to the credit report attributes, we use an indicator for whether an individual belongs to a racial or ethnic minority. We draw on prior work in Blattner and Nelson (2022) which computes minority/non-minority status following the industry-standard BISG methodology using name and geographic information to predict race and ethnicity (see Blattner and Nelson, 2022, for extensive validation of this imputation measure using a merge with HMDA data).[3] We aggregate all ethnic and racial minorities into a single minority category and define its complement as the non-minority (or majority) group. Given the probabilistic sampling strategy in Blattner and Nelson (2022), the share of minority applicants in our sample is about 19.7%.

## III.2 Lender and regulator preferences

Across both applications, the role of the principal is played by a financial *regulator* who cares either about financial stability or fair and non-discriminatory lending practices. The agent in this implementation of our game is a *lender* who uses past data on credit histories and defaults to score future loan applicants.

To be consistent with our theoretical approach, we provide results for the case where loss functions are mean-squared error in predicting repayment, $u(f(x), y) = -(y - f(x))^2$ as in Assumption 1. We also report implications for outcomes when these scores are used to approve a given fraction of applicants. We think of the mean-squared error approach as proxying for regulator and lender utility in a way that captures the quality of credit scoring beyond approval decisions. For example, scores may matter not only for approvals, but also for credit terms or for approvals across a range of thresholds.

In the disparate-impact example, the lender's preferred model is simply the best fit of credit repayment given

[3] BISG uses a two-step procedure to assign race and ethnicity. First, the procedure follows an approach developed by Sood and Laohaprapanon (2018). Their approach is implemented in a Python package available at github.com/appeler/ethnicolr. They model the relationship between the sequence of characters in a name and race and ethnicity using Florida Voter Registration data. After implementing this approach, the procedure updates each individual's baseline racial/ethnic probabilities with the racial and ethnic characteristics of the census block associated with her or his place of residence in 2000 using Bayes' Rule, and then updates the Bayesian posterior again using an individual's 2010 address and the 2010 census data. An individual is assigned to a racial/ethnic category if this category has the highest posterior probability for that individual. This two-step method is similar to methods used by the CFPB to construct race and ethnicity in fair lending analysis. Work by the CFPB (2014) shows that combining geographic and name-based information outperforms methods using either of these sources of information alone. BISG classification errors can be correlated with economic disadvantage, as minorities who live in predominately non-minority geographies are more likely to be misclassified by BISG as non-minority (Blattner and Nelson, 2022) and are more likely to have higher education and income (Greenwald et al., 2023). Such issues are important in some contexts but are unlikely to qualitatively affect our conclusions in this setting.

background characteristics of the applicant. The regulator additionally penalizes differences in average repayment predictions across majority and minority applicants. Hence, as in Application 1,

$$U_\theta^A(f) = -\,\mathrm{E}_\theta[(y - f(x))^2] \tag{DI-A}$$

$$U_\theta^P(f) = -\,\mathrm{E}_\theta[(y - f(x))^2] - \lambda^* \left(\mathrm{E}_\theta[f(x)|g{=}\text{majority}] - \mathrm{E}_\theta[f(x)|g{=}\text{minority}]\right) \tag{DI-P}$$

We calibrate the penalty $\lambda^*$ such that for the most flexible model, the correlation of the output with the minority flag is zero. Throughout, we assume that group identity $g$ does not enter the credit scoring function $f$ directly.

For the risk case we assume that the regulator's and the lender's preferences place different weights on good (or "high") and bad (or "low") states of the economy, with the regulator particularly concerned about the low state, similar to a stress-test scenario. We suppose that a particular group of borrowers – the sample of 50,000 individuals recently rejected for a credit card application – are sensitive to the state of the economy and default with probability 1 in the low state and probability 0 in the high state. We then refer to the lender's credit score as having a low-state MSE and a high-state MSE. For purposes of illustration, we assume that the regulator puts weight only on the low-state MSE, while the lender puts $\ell^* = 20\%$ weight on the low-state MSE. As in the illustration in Application 2 above, the lender (agent) and regulator (principal) utilities are therefore given by

$$U_\theta^A(f) = -\,\mathrm{E}_\theta[(y - f(x))^2] = -(1-\ell^*)\ \mathrm{E}_\theta[(y - f(x))^2|s{=}\text{high}] - \ell^*\ \mathrm{E}_\theta[(y - f(x))^2|s{=}\text{low}], \tag{R-A}$$

$$U_\theta^P(f) = -\,\mathrm{E}_\theta[(y - f(x))^2|s{=}\text{low}]. \tag{R-P}$$

Here, we assume that the (future) state is not known when (current) credit scoring decisions are taken, so that the state $s$ does not enter $f$.

In addition, for expositional purposes, we also consider intermediate preferences that interpolate between the regulator's and lender's preferences and show a Pareto frontier of achievable utility. In the disparate impact case, we vary the weight $\lambda$ put on disparate impact, varying the parameter from zero (lender utility) to the calibrated value $\lambda^*$ (regulator utility), showing the tradeoff between mean-squared error and disparate impact. In the risk application, we vary the weight $\ell$ put on the low state from 0% to 100%, where $\ell^* = 20\%$ represents lender utility, and 100% represents the preference of the regulator.

## III.3 Optimal explainers and constrained solutions

To implement explainability restrictions, we assume that the regular chooses a set $S^*$ of $|S^*| = k$ covariates to use in an explainer. We set $k = 5$ in our main implementation, and we also present robustness results for $k = 10$ and $k = 20$. The regulator then constrains the lender to choose only among functions $f : \mathcal{X} \to \mathbb{R}$ for which the regression coefficients $\mathrm{E}f$ in a linear regression of $f(x)$ on the variables in $S^*$ (and an intercept) agree with a target value $\beta^*$ chosen by the regulator. By Proposition 1, the optimal coefficient $\beta^*$ is the linear regression of the regulator's first-best solution on $S^*$ (and an intercept). Likewise, by Theorem 2, the set of $k$ optimal explainer covariates is chosen so that it minimizes mean-square error in a linear regression relating all data features to the difference between the regulator's and the lender's preferred prediction function.

In our two applications, the optimal covariate sets for targeted explainers take a particularly simple form. For the disparate-impact case, the optimal set of covariates solves

$$\arg\min_{S\in\mathcal{S}}\min_{\beta}\mathrm{E}_\theta[(g-x_S'\beta)^2] \tag{DI-Ex}$$

where $g\in\{1,0\}$ denotes group identity (with 1 corresponding to the majority group). Meanwhile, for the risk application, writing $h\in\{1,0\}$ for those applicants whose default is sensitive to the state of the economy, the optimal set solves

$$\arg\min_{S\in\mathcal{S}}\min_{\beta}\mathrm{E}_\theta[(h-x_S'\beta)^2]. \tag{R-Ex}$$

We contrast these two targeted explainers with a misalignment-agnostic explainer that simply explains as much as possible of the overall variation in repayment, namely

$$\arg\min_{S\in\mathcal{S}}\min_{\beta}\mathrm{E}_\theta[(y-x_S'\beta)^2]. \tag{Ex}$$

Except for differences in the estimation sample we choose for our two applications, the misalignment-agnostic explainer is the same across the two settings. Throughout, we include (but do not penalize) intercepts.

We implement these solutions empirically using boosted trees, where we separate training from evaluation in order to ensure valid inference. Details on optimization and sample splitting are presented in Appendix A.

### III.4 Empirical results

The main empirical results for both applications are summarized in Table 1 and Figure 2. Figure 2 illustrates the different components of utility that the regulator and lender optimize for. Table 1 evaluates the various credit scores by examining implied credit approval decisions, assuming that all applicants with a credit-score-predicted repayment probability of at least 85% are approved for a loan. Throughout, results for the disparate-impact application are listed in panel (a), and results for the risk application in panel (b).

We first consider results for the disparate impact application, for which we compare simple, fully explainable models to more complex ones. Figure 2a represents the trade-off between the fit of the model on the $x$-axis (which the lender wants to maximize) and disparate impact on the $y$-axis (which the regulator trades off with fit). The different preferences are exemplified by the first-best lender and regulator solutions; the lender's first-best solution is towards the right, while the regulator's first-best solution achieves lower disparate impact at the cost of some fit. Ignoring explainer constraints for the moment, we note that the more complex XGBoost models (solid black line) Pareto-dominate the less complex, fully explainable linear models (dashed lines), in the sense that there are more complex models that have better fit and lower disparate impact than either the simple regulator or lender models. Moving towards more complex models thus has the potential to achieve better outcomes for both regulator and lender. At the same time, the *unconstrained* model the lender would choose among all complex models (bottom right lender solution) has high disparate impact, and would only represent a marginal improvement over the regulator's preferred simple, fully explainable model. The main question is, therefore, whether the regulator can impose constraints based on simple explanations so that allowing the lender to choose from complex credit scores leads to a clear improvement in the regulator's utility.

To study this question within the disparate-impact application, we consider policies that allow the lender to choose among complex models subject to a constraint on simple explanations based on five explainer variables only. The blue line in Figure 2a represents the Pareto frontier of complex models subject to a constraint based on the agnostic explainer. The agnostic explainer captures the maximal amount of information about the overall variation of the underlying credit-scoring model, and imposing a constraint based on it leads to higher regulator utility than allowing the lender to make unconstrained choices. At the same time, the regulator can further improve disparate impact by instead constraining explanations based on a targeted explainer. The green line represents the Pareto frontier of complex models subject to a constraint based on the targeted explainer, and this frontier dominates that from the agnostic explainer. This targeted explainer focuses on aspects of the credit score that are particularly informative about disparate impact and, thus, about the source of misalignment in our example.

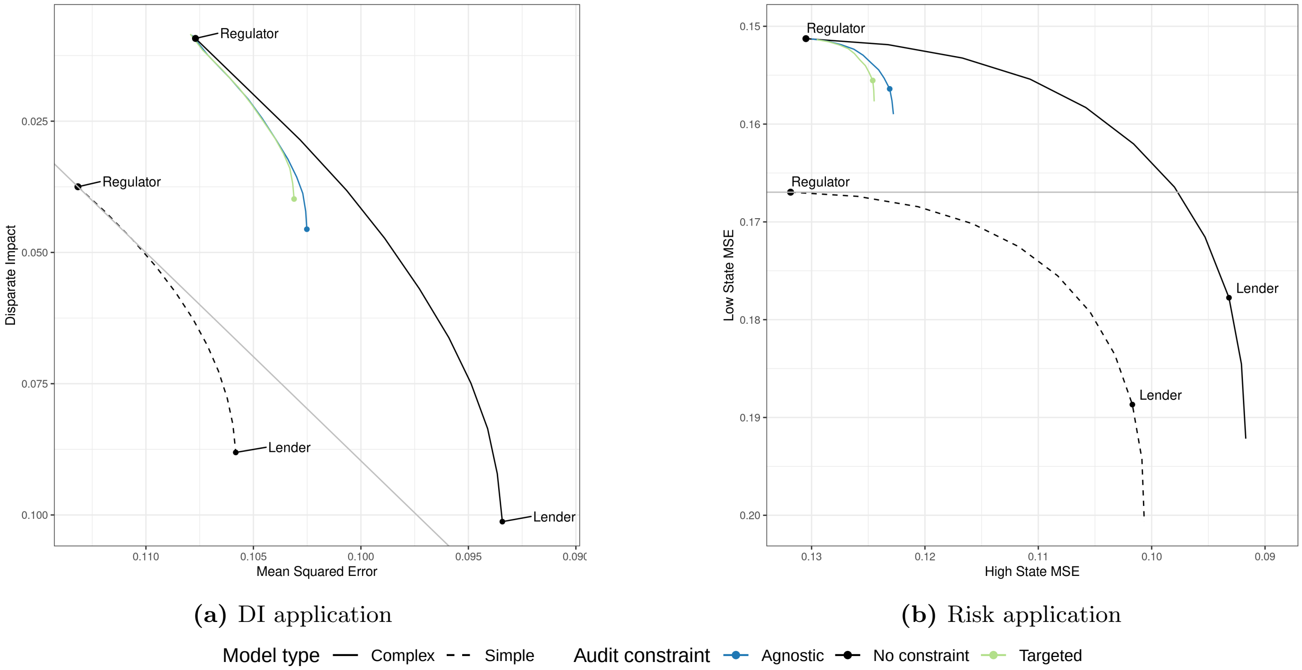
In our analysis so far, we have focused on fit and disparate impact of the credit scoring function itself. In Table 1a we also consider disparate impact of binary loan approval decisions based on the credit score. The "DI on acceptance" column reports the disparate impact of credit approval decisions when all applicants with a predicted repayment probability of 85% or more are approved for credit. These results confirm that imposing constraints on explanations can reduce disparate impact with only a small cost in predictive power. While the unconstrained lender solution leads to a 4.5 percentage point higher approval rate among majority applicants, that gap is reduced to 0.9 percentage points with an agnostic explainer, and approximately zero when the specific targeted explainer is used. The complex solution subject to a targeted explainer constraint thus outperforms the regulator's preferred simple, fully explainable model even in terms of disparate impact. Meanwhile, the DI and MSE columns report the same results as those in Figure 2a for the lender's and regulator's unconstrained and simple models, as well for the lender's constrained solutions.

We next consider results for the risk application, finding overall similar patterns. Figure 2b shows the superior performance of the complex model and the value of explanation constraints: we ask whether all possible convex combinations of the regulator's and lender's preferences would prefer the complex model over a model that is constrained ex-ante to be simple. We trace out a frontier in the space of low-state utility and high-state utility that shows the performance of the complex model for all combinations of regulator and firm preferences. We then trace out analogous curves when the complex model is subjected to an agnostic explainer and a targeted explainer, and finally, analogous curves for the ex-ante simple model (with no explainer). As in the disparate impact case, we see that complex models constrained by either explainer are preferred by both the regulator and the firm to any simple model chosen by any weighted combination of regulator and firm preferences. Constraints play a central role in this result, as the unconstrained complex model is worse for the regulator than the regulator's preferred simple model.

Table 1b provides further details for the risk application. The first four rows show statistics for the lender's complex model when the regulator uses no constraint, an agnostic explainer, an optimal targeted explainer, and an ex-ante constraint to use simple models only; the next two rows show the regulator's preferred complex model and the regulator's preferred simple model. Across columns, the table shows model behavior in terms of false positives (accepted defaults), true positives (accepted non-defaults) and the mean-squared error across both defaults and non-defaults. These statistics are presented for both the high state and the low state; recall that the lender places less preference weight on low-state outcomes than the regulator does.

The table confirms that the different models perform as expected: the constrained models perform better in the low state than the unconstrained models, though not as well as the regulator's preferred model; models that perform better in the low state face a trade-off in that they perform worse in the high state; and complex models generally outperform simple models.

The results across both applications present an empirical illustration of our main theoretical results. First, as in Theorem 1, a complex model with an explainer outperforms a model that is constrained ex-ante to be simple in our data. Here, this holds both for the regulator's preferences and the lender's preferences; both sides are better off using the more complex model with an appropriate explainer. Second, the choice of explainer matters, with the targeted explainer from Theorem 2 providing better results for the regulator across both applications. The empirical application therefore expands our illustration from the case of two binary variables in the example from Section II to a realistic exercise in credit underwriting on over 500 covariates. In Figure 3 in the appendix, we show that these results are robust to a larger number of explainer covariates than the five chosen for illustration in this section.



**(a)** DI application

**(b)** Risk application

**Figure 2:** Out-of-sample performance of unconstrained and constrained prediction models across key objectives for the disparate impact (left) and risk (right) applications. Complex models are XGBoost models on all 518 covariates, while simple models are linear regression on five covariates chosen by the LASSO. The frontiers vary the relative weight put on each objective, where the regulator and lender marks correspond to the solutions maximizing the empirical analog to the objectives (DI-P), (DI-A) and (R-P), (R-A) respectively. The colored lines represent prediction models subject to constraints imposed by the agnostic and targeted explainers, respectively. The gray line represents the points at which the regulator would be indifferent to its preferred simple model.

## III.5 Optimal explainers

Table 2 further illustrates the differences between the agnostic explainer and the optimal, targeted explainer. Recall that the agnostic explainer selects the variables that best summarize the predictions made by the lender's risk model and then projects model predictions onto these variables. The first two blocks of the table show how the regulator places constraints on the coefficients on each of the explainer variables in this projection.

|  |  | DI on acceptance | DI | MSE |
|---|---|---|---|---|
| Lender | Unconstrained | 0.0455 | 0.1013 | 0.0934 |
|  | Agnostic | 0.0087 | 0.0456 | 0.1025 |
|  | Targeted | 0.0004 | 0.0398 | 0.1031 |
|  | Simple | 0.0179 | 0.0620 | 0.1103 |
| Regulator | Unconstrained | −0.0233 | 0.0092 | 0.1077 |
|  | Simple | 0.0100 | 0.0348 | 0.1142 |

**(a)** DI application

|  |  | High state |  |  | Low state |  |  |
|---|---|---|---|---|---|---|---|
|  |  | Accepted defaults | Accepted non-defaults | MSE | Accepted defaults | Accepted non-defaults | MSE |
| Lender | Unconstrained | 0.5337 | 0.8969 | 0.0932 | 0.6480 | 0.9313 | 0.1778 |
|  | Agnostic | 0.5510 | 0.8943 | 0.1231 | 0.6404 | 0.9343 | 0.1564 |
|  | Targeted | 0.5582 | 0.8932 | 0.1246 | 0.6345 | 0.9367 | 0.1555 |
|  | Simple | 0.6316 | 0.8824 | 0.1044 | 0.6775 | 0.9194 | 0.1959 |
| Regulator | Unconstrained | 0.5836 | 0.8895 | 0.1305 | 0.6211 | 0.9421 | 0.1513 |
|  | Simple | 0.6314 | 0.8824 | 0.1333 | 0.6774 | 0.9195 | 0.1752 |

**(b)** Risk application

**Table 1:** Performance of unconstrained and constrained prediction models across key metrics for the disparate impact (top) and risk (bottom) applications, as in Figure 2. For both lender and regulator, unconstrained XGBoost and simple (five-variable) linear models are listed. For the lender, the agnostic model optimizes the lender objective subject to a constraint on the coefficients of the agnostic explainer, while the targeted model optimizes the objective subject to a constraint based on the targeted explainer. In the disparate impact application, "DI" denotes the average difference in credit scores across groups and "DI on acceptance" denotes the average in acceptance probabilities at an approval threshold of 85% probability of repayment. In the risk application, "accepted defaults" and "accepted non-defaults" denotes the fraction of applicants that are accepted among those who ultimately default and among those who ultimately repay, respectively, for an approval threshold of 85% probability of repayment. "MSE" denotes the mean-squared error.

Within each block, the first column shows the case where the regulator constrains these coefficients, and the second column shows what value these coefficients would have taken in the lender's preferred model without any explainer constraint.

The pattern across the first two columns is illustrative. When the constrained coefficient is less positive (or more negative) than the unconstrained coefficient, the constraint has the effect of requiring the model's predicted default rates to be lower for borrowers with higher values of that variable. For example, in the first row of panel (a), the constrained coefficient on percentage of trades ever delinquent is .143, whereas the unconstrained coefficient is .256; the constraint thus lowers the credit score's predicted default rates for individuals with, for example, a 10-percentage-point higher prevalence of delinquency across their past loans by 1.1 percentage points.

The final four columns of the table then shed light on the regulator's preferred changes in these coefficients by showing how these variables are correlated both with default (the outcome the lender's model seeks to predict) and with preference misalignment (an indicator for which borrowers the lender and regulator have divergent preferences over). Univariate correlations are shown in columns (5) and (6), while coefficients from a multivariate regression are shown in columns (7) and (8).[4]

Focusing first on the disparate impact example in panel (a), where the regulator prefers *lower* predicted

[4] Note that the last two columns show coefficients from a multivariate regression that is jointly run for all explainer covariates included in the agnostic explainer or in the targeted explainer. Due to high correlations between covariates that are only included in one but not both, some of the coefficients from the multivariate regression are degenerate. We therefore focus mainly on univariate correlations.

default rates for minority applicants in order to shrink the penalty term in (DI-P), these correlations reveal that the regulator indeed tends to constrain explainer coefficients to be lower when the corresponding variable has a positive multivariate correlation with an indicator for preference misalignment (i.e., an indicator for minority status). Turning next to panel (b), where the regulator prefers *higher* predicted default rates for borrowers more likely to default in a bad state of the economy, the opposite pattern is also intuitive: the regulator tends to constrain explainer coefficients to be higher when the corresponding variable is positively correlated with preference misalignment.

Continuing to the optimal, targeted explainer, Table 2 also illustrates how the targeted explainer achieves dominant outcomes relative to the agnostic explainer. Recall that the targeted explainer selects variables that best predict preference misalignment, rather than those that best summarize model predictions, and the explainer then projects model predictions onto these variables. Similar to the agnostic case, the regulator then places constraints on the coefficients in this projection. While several variables (those in the middle of each panel) are selected by both explainers, the variables selected by only the targeted explainer tend to be more highly correlated with misalignment than the variables selected by the agnostic explainer. In so doing, the targeted explainer leads to model predictions that are more closely aligned with the regulator's preferences, at lower cost in terms of overall model performance, relative to an agnostic explainer.

There are further interesting patterns in which variables are selected by the agnostic and targeted explainers. In panel (a), the disparate impact case, we see that the main drivers of differences between groups and the main drivers of repayment are quite different, such that the agnostic and targeted explainers describe substantially different aspects of the credit scoring function. In particular, the agnostic explainer selects variables that capture a history of credit risk and hence may be particularly predictive of future credit risk; in contrast, the targeted explainer selects variables that reflect broader patterns in credit use and access, which may correlate more closely with the drivers of inequality (e.g., historical exclusion from formal credit markets) that shape the regulator's group-specific preferences. In the risk case (panel (b)), on the other hand, the targeted explainer picks variables that are particularly helpful to describe who potentially defaults in the low state of the economy. These variables are more similar to the variables selected by the agnostic explainer in the sense that both capture a history of credit risk and are selected to be predictive of future risk. These differences help interpret why the choice of explainer makes a relatively minor difference in the risk case, as in Figure 2.

| | Agnostic explainer | | Targeted explainer | | Correlation | | Coefficients | |
|---|---|---|---|---|---|---|---|---|
| | Constrained | Unconstrained | Constrained | Unconstrained | Default | Misalignment | Default | Misalignment |
| Intercept | 0.1511 | 0.1525 | 0.1511 | 0.1525 | | | 0.1501 | 0.1973 |
| Percentage of trades ever delinquent | 0.1433 | 0.2563 | | | 0.3569 | 0.1719 | 0.0432 | 0.0189 |
| Missing: Months since most recent third party collection occurrence | −0.0111 | −0.0334 | | | −0.3097 | −0.1769 | −0.0276 | −0.0066 |
| Number of trades 60 or more days past due ever | −0.1300 | −0.2704 | | | 0.3346 | 0.1218 | 0.0504 | −0.0079 |
| Missing: Balance of most recent, open, premium bankcard trade in 12 months | −0.0026 | 0.0542 | 0.0014 | −0.0082 | 0.2618 | 0.2062 | 9.5044 | −89.1644 |
| Missing: Months since most recent non-medical third party collection occurrence | −0.0196 | −0.0470 | −0.0474 | −0.0965 | −0.3073 | −0.1812 | −0.0247 | −0.0310 |
| Missing: Credit line of most recent, open, premium bankcard trade in 12 Months | | | 0.0014 | −0.0082 | 0.2618 | 0.2062 | −9.4854 | 89.1995 |
| Average bankcard credit limit | | | 0.0079 | −0.0398 | −0.1955 | −0.1803 | −0.0115 | −0.0289 |
| Transactor behavior in past 8 months | | | −0.0104 | −0.0500 | −0.2125 | −0.1674 | −0.0227 | −0.0291 |

**(a)** DI application

| | Agnostic explainer | | Targeted explainer | | Correlation | | Coefficients | |
|---|---|---|---|---|---|---|---|---|
| | Constrained | Unconstrained | Constrained | Unconstrained | Default | Misalignment | Default | Misalignment |
| Intercept | 0.2608 | 0.1153 | 0.2683 | 0.1499 | | | 0.1503 | 0.1458 |
| Missing: Balance of Most Recent, Open, Premium Bankcard Trade in 12 Months | 0.0792 | 0.0499 | | | 0.2848 | 0.1478 | 0.0349 | 0.0091 |
| Missing: Months since most recent third party collection occurrence | −0.0542 | −0.0263 | | | −0.3287 | −0.1941 | −0.0247 | −0.0155 |
| Missing: Months since most recent non-medical third party collection occurrence | −0.0535 | −0.0269 | | | −0.3268 | −0.1887 | −0.0193 | −0.0021 |
| Number of trades 60 or more days past due ever | −0.3562 | −0.1813 | | | 0.3467 | 0.1085 | 0.0535 | −0.0140 |
| Percentage of trades ever delinquent | 0.2824 | 0.1472 | 0.1503 | 0.0964 | 0.3749 | 0.2058 | 0.0387 | 0.0315 |
| Missing: Max aggregate bankcard balance over last 3 months | | | 0.1471 | 0.0688 | 0.1907 | 0.2040 | 0.0074 | 0.0152 |
| Number of unpaid collections | | | 0.0467 | 0.0203 | 0.2767 | 0.2047 | 0.0271 | 0.0265 |
| Missing: Bankcard balance magnitude algorithm over last 24 months | | | 0.0536 | 0.0115 | 0.1722 | 0.1833 | 0.0121 | 0.0123 |
| Missing: Months since most recent bankcard trade opened | | | −0.2209 | −0.0919 | 0.2146 | 0.2101 | 0.0008 | 0.0191 |

**(b)** Risk application

**Table 2:** Explainer coefficients in the disparate impact (top) and risk (bottom) applicants for the solutions presented in Figure 2 and Table 1. The "agnostic explainer" and "targeted explainer" columns report the coefficients on the respective five chosen explainer variables and a constant for different credit scores. The "constrained" coefficients report the coefficients for the regulator's preferred model, which are also those of the lender's model subject to the explanation constraint (dot at the end of the blue and green lines, respectively, in Figure 2). The "unconstrained" coefficients report the explainer coefficients for the unconstrained lender model (lender dot at the end of the solid black lines in Figure 2). The remaining columns provide details of the chosen variables, listing the correlation with default and the variable encoding misalignment (that is, minority identity in the disparate impact application and a rejected application in the risk application), as well as linear regression coefficients (discussed further in Footnote 4 of the text) for a joint linear regression of the variable on all covariates in the table.

# IV. Conclusion

We study the problem of a principal who seeks to regulate an agent's choices over complex algorithms. In an empirical example, we consider a lender (agent) who is choosing a complex credit scoring function to evaluate the credit risk of potential borrowers, whereas a financial regulator (principal) seeks either to reduce disparate impact of the credit scoring function across racial and ethnic groups or to make the credit scoring function more conservative over a group of borrowers who may be particularly likely to default in a bad state of the economy. The example illustrates our theoretical results: restricting the lender ex-ante to simple credit-scoring algorithms makes regulation easier, but creates inefficient results that can come at a cost to both the regulator and the lender. Instead, we show that effective regulation can allow for complex machine-learning algorithms, and should regulate them based on those aspects of the model most related to the source of preference misalignment. We thus add to a broader discussion about regulating algorithms: when should regulators allow the use of complex models, and what aspects of these models should regulators regulate? Our theory and empirical results add further evidence that regulation based on prohibiting the use of certain data or constraining the functional form of models may be ineffective and inefficient. Instead, our work illustrates benefits to allowing for flexible algorithms that are regulated based on descriptions of their key behaviors, and our results provide guidance for what these key behaviors are.

–

Booth School of Business, University of Chicago, Chicago, IL, USA
Graduate School of Business, Stanford University, Stanford, CA, USA

# Appendix

## A Empirical optimization

In empirically solving for optimal credit scores, we face three challenges. First, we only observe a sample, and not the full distribution of the data. Second, solving the regulator's optimization problem of finding good explainer covariates is generally computationally hard. And third, we have to solve for complex prediction functions subject to explainer constraints.

In order to address the first challenge of facing only a noisy sample form the distribution of applicants, we solve all of our optimization problems in a training sample, and later evaluate the solutions on a hold-out set to obtain accurate measurements of performance. We implement the above solutions of the lender and regulator using boosted trees, which has the best out-of-sample performance among a number of parametric and non-parametric standard prediction methods on our data. We provide additional details in the appendix, including a comparison to regularized and non-regularized linear regression, random forests, and neural networks in Table 4 and a comparison of in-sample and out-of-sample performance of our solutions in Figure 4. In the remaining part of this section, all reported results are based on the out-of-sample performance of boosted trees.

For the second challenge, obtaining the optimal set of explainer variables in each case involves a search over all possible sets of $k$ variables, where $k = 5$ in our main specification, out of a total of more than 500. This would generally be computationally infeasible. Instead, we choose the set of variables by a LASSO regression that approximates the solution of (Ex), (R-Ex), and (DI-Ex). Specifically, we fit linear models that minimize the sum of the respective empirical risk and a penalty $\nu \|\beta\|_1$ on the coefficients $\beta$, where $\nu$ is chosen such that the number of non-zero coefficients is $k$. We then use the set of variables with non-zero coefficients as the explainer variables.

To address the third challenge of solving for complex prediction functions subject to explainer constraints, we rely on projection methods. We first solve for the first-best lender and regulator solutions using boosted trees (XGBoost) on the training sample, as well as for preferences that put varying weight $\lambda$ on disparate impact in the DI application and preferences that put varying probability $\ell$ on the low-state MSE in the risk application. We then compute the respective solution subject to explainer constraints by applying Proposition 1. Specifically, from an unconstrained solution $f$ we obtain the constrained solution by setting $\hat{f}(x) = x'_S\hat{\beta}^P + (f(x) - x'_S\hat{\beta}^f)$, where $\hat{\beta}^P$ are the OLS coefficients for the regression of the first-best regulator solution on the explainer covariates in $S$ and $\hat{\beta}^f$ are the analogous coefficients for $f$.

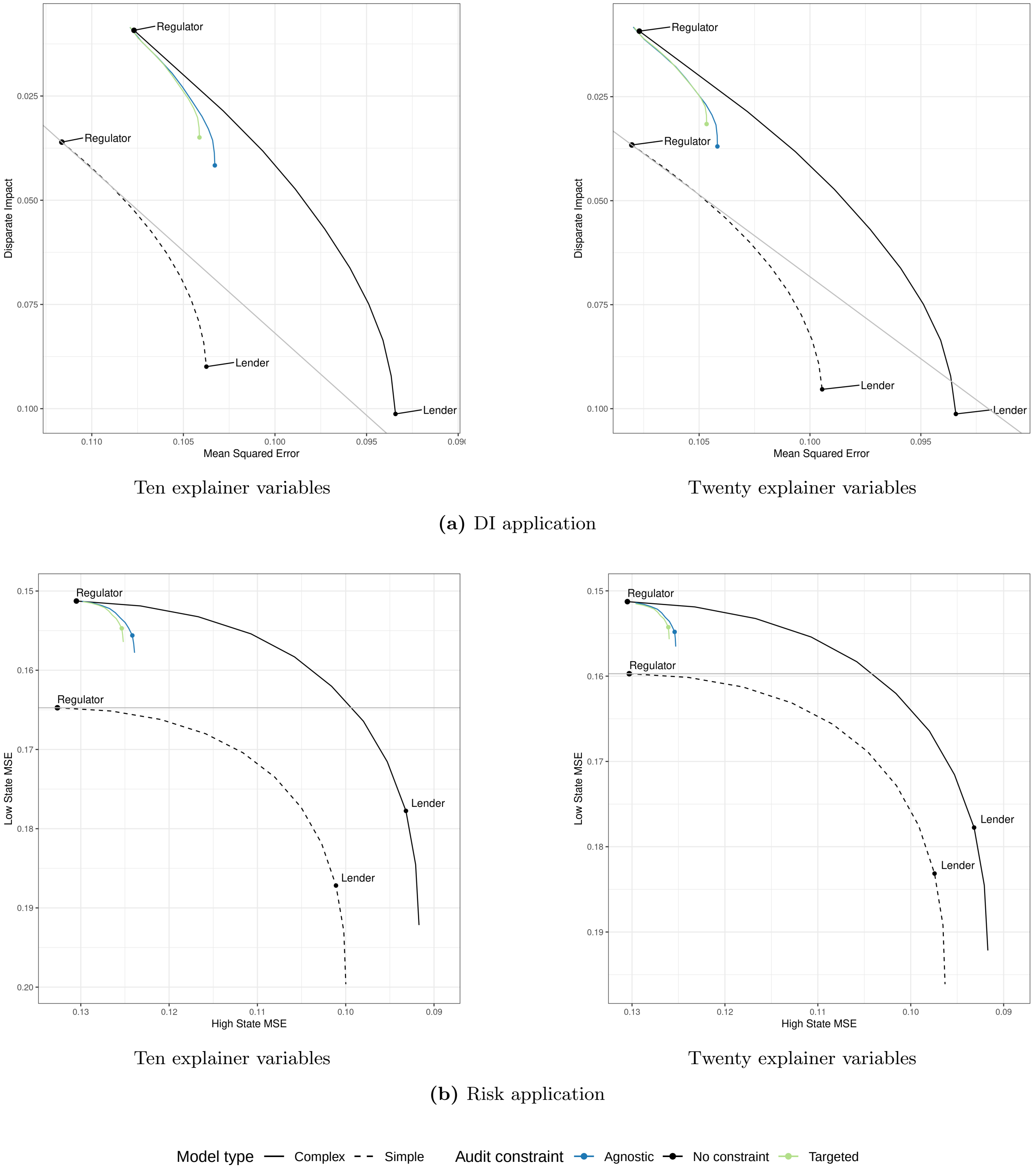
# B Additional Empirical Results

| | Accepted | Rejected | Minority |
|---|---|---|---|
| Repayment | 0.849 | – | 0.841 |
| | (0.358) | | (0.431) |
| Percentage of trades ever delinquent | 12.005 | 24.236 | 18.890 |
| | (19.706) | (28.382) | (23.951) |
| Missing: Months since most recent third party collection occurrence | 0.752 | 0.510 | 0.596 |
| | (0.432) | (0.500) | (0.491) |
| Missing: Months since most recent non-medical third party collection occurrence | 0.823 | 0.612 | 0.683 |
| | (0.381) | (0.487) | (0.465) |
| At least one trade 90 or more days past due ever | 0.269 | 0.478 | 0.424 |
| | (0.443) | (0.500) | (0.494) |
| Missing: Aggregate bankcard amount past due for month 17 | 0.190 | 0.374 | 0.306 |
| | (0.392) | (0.484) | (0.461) |
| Missing: Aggregate bankcard amount past due for month 15 | 0.187 | 0.371 | 0.303 |
| | (0.390) | (0.483) | (0.460) |
| Number of collections excluding medical | 0.529 | 1.446 | 1.032 |
| | (1.575) | (2.727) | (2.168) |
| Missing: Aggregate bankcard amount past due for month 18 | 0.190 | 0.372 | 0.306 |
| | (0.393) | (0.483) | (0.461) |
| Number of unpaid collections | 0.715 | 2.384 | 1.313 |
| | (2.326) | (4.978) | (3.015) |
| Missing: Number of months since overlimit on a bankcard | 0.760 | 0.673 | 0.696 |
| | (0.427) | (0.469) | (0.460) |

**Table 3:** Summary statistics of repayment as well as the ten variables (out of a total of 518) with the highest importance scores for predicting credit-card repayment. Averages are reported separately for subsamples that were accepted or rejected for a loan. An additional column presents average values specifically for the subsample of accepted minority applicants. Standard deviations are in parentheses.

| Model | MSE | ROC AUC | Log loss | $R^2$ |
|---|---|---|---|---|
| XGBoost | 0.0934 | 0.8672 | 0.3007 | 0.2802 |
| Logistic XGBoost | 0.0935 | 0.8680 | 0.3009 | 0.2794 |
| Linear model | 0.0959 | 0.8603 | 0.3230 | 0.2607 |
| Elastic net | 0.0959 | 0.8609 | 0.3210 | 0.2610 |
| Logistic model | 0.0955 | 0.8626 | 0.3103 | 0.2639 |
| Random forest | 0.0946 | 0.8608 | 0.3092 | 0.2711 |
| Logistic random forest | 0.0945 | 0.8632 | 0.3057 | 0.2715 |
| Neural network | 0.0973 | 0.8504 | 0.3569 | 0.2501 |
| Logistic neural network | 0.0950 | 0.8637 | 0.3045 | 0.2683 |
| Finscore logistic | 0.1031 | 0.8429 | 0.3354 | 0.2058 |

**Table 4:** Out-of-sample performance of varying prediction models according to mean-squared error ("MSE," where lower values indicate better fit), the area under the ROC curve ("ROC AUC," where higher values indicate better fit), the negative log-likelihood ("log loss," where lower values indicate better fit), and the coefficient of determination ("$R^2$," where higher values indicate better fit). Our main implementation is based on boosted trees ("XGBoost," first row). As a benchmark, we also report results for using a commercially available credit score as the only covariate in a simple logistic regression ("finscore logistic," last row). This credit score is not used as a covariate in the other models.



Ten explainer variables

Twenty explainer variables

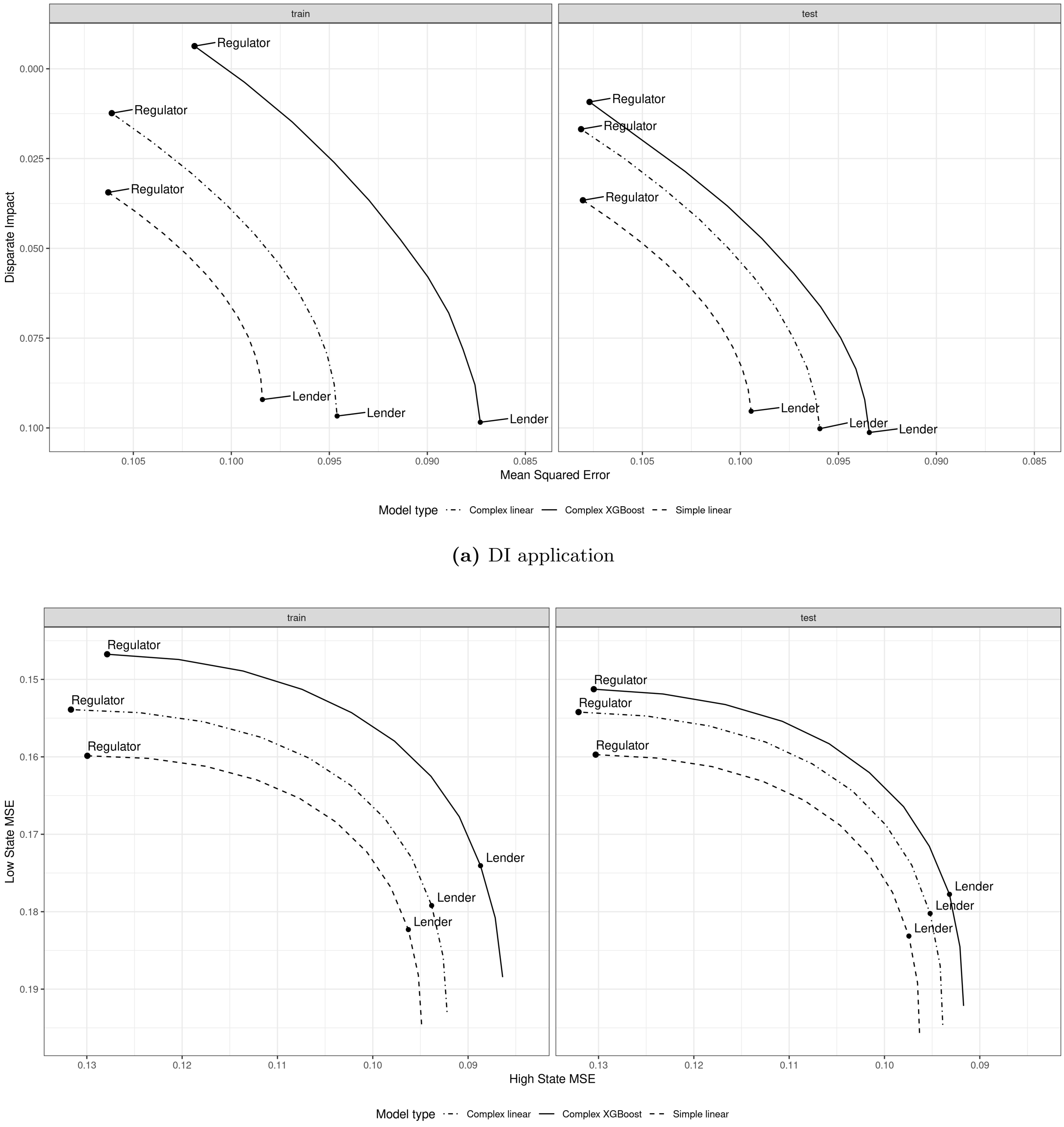
**(a)** DI application

Ten explainer variables

Twenty explainer variables

**(b)** Risk application

**Figure 3:** Out-of-sample performance of unconstrained and constrained prediction models across key objectives for the disparate impact (top) and risk (bottom) applications as in Figure 2, but with varying explainer complexity. The left panels use ten variables for the simple (linear) model and the explainer (in addition to a constant for the intercept), while the right panel uses twenty. This illustrates robustness of the results reported in Table 1 and Figure 2 based on five explainer variables.



**(a)** DI application

**(b)** Risk application

**Figure 4:** Performance of models of varying complexity in the training sample (left panels) and on the hold-out set (right panels) for the disparate impact (top panels) and risk (bottom panels) examples, following Figure 2. The simple linear model is a linear-regression model on five covariates chosen by the LASSO. The complex linear and complex XGBoost models use all 518 covariates to predict repayment. The regulator and lender marks correspond to the solutions maximizing the empirical analog to the objectives (DI-P), (DI-A) and (R-P), (R-A), respectively.

## C Proofs

*Proof of Proposition 1.* Write $s(x) = x'_S \operatorname{E}_\theta^{-1}[x_S x'_S] \operatorname{E}_\theta[x_S y]$ for the linear projection of $y$ on $x_S$. Among the constrained set, the agent chooses among functions $x'_S \beta^P_\theta + r(x)$ subject to $\operatorname{E}_\theta[r(x) x_S] = \mathbf{0}$, minimizing

$$\begin{aligned}\operatorname{E}_\theta[(x'_S \beta^P_\theta + r(x) - y)^2] &= \operatorname{E}_\theta[(x'_S \beta^P_\theta - s(x) + r(x) - (y - s(x)))^2] \\ &= \operatorname{E}_\theta[(x'_S \beta^P_\theta - s(x))^2 + \operatorname{E}_\theta[(r(x) - (y - s(x)))^2].\end{aligned}$$

Hence, the optimality of $r(x)$ does not depend on $\beta^P_\theta$, so $r(x) = r^A_\theta(x)$ is an optimal solution. □

*Proof of Theorem 1.* For *covariate shift*, $\mathcal{F} = \mathbb{R}^{\mathcal{X}}$ achieves the first-best solution $f_\theta$ on the support of $\mu_\theta$, and thus also for the principal. Only if the ex-ante complexity constraint is trivial in the sense that it allows for solutions that are $\mu^P_\theta$-almost surely the same as $f_\theta$ would a constrained solution achieve the same utility. For *model shift*, by Proposition 1, the solution from delegating to the agent subject to an explainability constraint with covariates $S_2$ takes the form $f_\theta(x) - x'_{S_2}\beta$ for some $\beta$ that the principal has control over by requiring $\operatorname{E} f = \mathrm{E} f_\theta - \beta$. Hence, the principal achieves risk $\int_{\mathcal{X}} (f_\theta(x) - x'_{S_2}\beta - f^P_\theta(x)) \, \mathrm{d}\mu_\theta(x)$, where $\beta$ can be set to attain the minimum, yielding the left-hand risk for the principal. At the same time, if the principal restricts the agent to simple linear functions on covariates $S_1$, she can dictate the choice of prediction function and achieve risk $\int_{\mathcal{X}} (x'_{S_1}\beta - f^P_\theta(x)) \, \mathrm{d}\mu_\theta(x)$, where $\beta$ can be set to attain the minimum again, yielding the right-hand risk for the principal. □

*Proof of Proposition 2.* The preference of the principal is equivalent to minimizing

$$\begin{aligned}&\int_X (f(x) - f_\theta(x))^2 \, \mathrm{d}\mu_\theta(x) + \lambda \left( \int_X f(x) \frac{g_\theta(x)}{\bar{g}_\theta} \, \mathrm{d}\mu_\theta(x) - \int_{\mathcal{X}} f(x) \frac{1 - g_\theta(x)}{1 - \bar{g}_\theta} \, \mathrm{d}\mu_\theta(x) \right) \\ &= \int_X \left( (f(x) - f_\theta(x))^2 + \lambda f(x) \left( \frac{g_\theta(x)}{\bar{g}_\theta} - \frac{1 - g_\theta(x)}{1 - \bar{g}_\theta} \right) \right) \mathrm{d}\mu_\theta(x) \\ &= \int_X \left( f^2(x) - 2f(x) \left( f_\theta(x) - \lambda \frac{g_\theta(x) - \bar{g}_\theta}{2\bar{g}_\theta(1 - \bar{g}_\theta)} \right) + f^2_\theta(x) \right) \mathrm{d}\mu_\theta(x) \\ &= \int_X \left( f(x) - f_\theta(x) + \lambda \frac{g_\theta(x) - \bar{g}_\theta}{2\bar{g}_\theta(1 - \bar{g}_\theta)} \right)^2 \mathrm{d}\mu_\theta(x) + \int_X \left( f^2_\theta(x) - \left( f_\theta(x) - \lambda \frac{g_\theta(x) - \bar{g}_\theta}{2\bar{g}_\theta(1 - \bar{g}_\theta)} \right)^2 \right) \mathrm{d}\mu_\theta(x).\end{aligned}$$

Since the right part does not depend on $f_\theta$, this preference is equivalent to Definition 2 with $f^P_\theta(x) = f_\theta(x) - \lambda \frac{g_\theta(x) - \bar{g}_\theta}{2\bar{g}_\theta(1 - \bar{g}_\theta)}$. □

*Proof of Theorem 2.* The result is immediate from the fact that the expression the theorem maximizes over represents the average risk from optimal delegation with explainer, as in the proof of Theorem 1. □

# References

Adams, William, Liran Einav, and Jonathan Levin (2009). Liquidity Constraints and Imperfect Information in Subprime Lending. *American Economic Review*, 99(1):49–84. (Cited on page 5.)

Agrawal, Ajay, Joshua S Gans, and Avi Goldfarb (2019). Exploring the impact of artificial intelligence: Prediction versus judgment. *Information Economics and Policy*, 47:1–6. (Cited on page 5.)

Alonso, Ricardo and Niko Matouschek (2008). Optimal Delegation. *The Review of Economic Studies*, 75(1):259–293. (Cited on pages 2, 6, and 8.)

Angelova, Victoria, Will S Dobbie, and Crystal Yang (2023). Algorithmic recommendations and human discretion. Technical report, National Bureau of Economic Research. (Cited on page 5.)

Arnold, David, Will Dobbie, and Peter Hull (2022). Measuring racial discrimination in bail decisions. *American Economic Review*, 112(9):2992–3038. (Cited on page 5.)

Athey, Susan C., Kevin A. Bryan, and Joshua S. Gans (2020). The Allocation of Decision Authority to Human and Artificial Intelligence. *AEA Papers and Proceedings*, 110:80–84. (Cited on page 4.)

Baker, George P (1992). Incentive contracts and performance measurement. *Journal of political Economy*, 100(3):598–614. (Cited on page 6.)

Barocas, Solon and Andrew D Selbst (2016). Big data's disparate impact. *California law review*, pages 671–732. (Cited on page 5.)

Bartlett, Robert P, Adair Morse, Nancy Wallace, and Richard Stanton (2019). Algorithmic accountability: A legal and economic framework. (Cited on page 6.)

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650. (Cited on page 18.)

Bhatt, Umang, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley (2020). Explainable Machine Learning in Deployment. *arXiv:1909.06342*. (Cited on page 6.)

Blattner, Laura and Scott Nelson (2022). How Costly is Noise? Data and Disparities in US Mortgage Market. (Cited on pages 5, 19, and 20.)

Carvalho, Diogo V., Eduardo M. Pereira, and Jaime S. Cardoso (2019). Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, 8(8):832. (Cited on page 6.)

CFPB (2014). Using publicly available information to proxy for unidentified race and ethnicity. *Consumer Financial Protection Bureau.* (Cited on page 20.)

Chan, David C, Matthew Gentzkow, and Chuan Yu (2022). Selection with variation in diagnostic skill: Evidence from radiologists. *The Quarterly Journal of Economics*, 137(2):729–783. (Cited on page 5.)

Chatterjee, Satyajit, Dean Corbae, Kyle P. Dempsey, and José-Víctor Ríos-Rull (2020). A Quantitative Theory of the Credit Score. Technical Report 27671, National Bureau of Economic Research, Inc. (Cited on page 5.)

Chen, Chaofan, Kangcheng Lin, Cynthia Rudin, Yaron Shaposhnik, Sijia Wang, and Tong Wang (2018). An Interpretable Model with Globally Consistent Explanations for Credit Risk. *arXiv:1811.12615.* (Cited on page 6.)

Chouldechova, Alexandra (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163. (Cited on page 8.)

Corbett-Davies, Sam and Sharad Goel (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023.* (Cited on page 5.)

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806. (Cited on page 5.)

Coston, Amanda, Ashesh Rambachan, and Alexandra Chouldechova (2021). Characterizing fairness over the set of good models under selective labels. In *International Conference on Machine Learning*, pages 2144–2155. PMLR. (Cited on page 4.)

Cowgill, Bo and Catherine Tucker (2017). Algorithmic bias: A counterfactual perspective. *NSF Trustworthy Algorithms*, 3. (Cited on page 5.)

Cowgill, Bo and Catherine E Tucker (2019). Economics, fairness and algorithmic bias. *preparation for: Journal of Economic Perspectives.* (Cited on page 5.)

Doshi-Velez, Finale and Been Kim (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608.* (Cited on pages 5 and 6.)

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. (Cited on page 4.)

Einav, Liran, Mark Jenkins, and Jonathan Levin (2013). The impact of credit scoring on consumer lending. *The RAND Journal of Economics*, 44(2):249–274. (Cited on page 5.)

Faria-e Castro, Miguel, Joseba Martinez, and Thomas Philippon (2017). Runs versus Lemons: Information Disclosure and Fiscal Capacity. *Review of Economic Studies*, 84(4):1683–1707. (Cited on page 6.)

Frankel, Alexander (2014). Aligned Delegation. *American Economic Review*, 104(1):66–83. (Cited on pages 2, 6, and 8.)

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther (2022). Predictably unequal? the effects of machine learning on credit markets. *The Journal of Finance*, 77(1):5–47. (Cited on pages 5 and 6.)

Gillis, Talia B (2021). The input fallacy. *Minn. L. Rev.*, 106:1175. (Cited on page 4.)

Gillis, T B and J L Spiess (2019). Big data and discrimination. *The University of Chicago law review. University of Chicago. Law School.* (Cited on page 4.)

Goldstein, Itay and Yaron Leitner (2017). Stress Tests and Information Disclosure. SSRN Scholarly Paper ID 3029761, Social Science Research Network, Rochester, NY. (Cited on page 6.)

Greenstone, Michael, Paul Oyer, and Annette Vissing-Jorgensen (2006). Mandated Disclosure, Stock Returns, and the 1964 Securities Acts Amendments. *The Quarterly Journal of Economics*, 121(2):399–460. (Cited on page 6.)

Greenwald, Daniel, Sabrina T Howell, Cangyuan Li, and Emmanuel Yimfor (2023). Regulatory arbitrage or random errors? implications of race prediction algorithms in fair lending analysis. Technical report, National Bureau of Economic Research. (Cited on page 20.)

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi, and Fosca Giannotti (2018). A Survey Of Methods For Explaining Black Box Models. *arXiv:1802.01933.* (Cited on page 6.)

Hashemi, Masoud and Ali Fathi (2020). PermuteAttack: Counterfactual Explanation of Machine Learning Credit Scorecards. *arXiv:2008.10138.* (Cited on page 6.)

Holmström, B. (1977/1984). On the theory of delegation. In Boyer, M. and R. Kihlstrom, editors, *Bayesian Models in Economic Theory.* (Cited on pages 2, 6, and 8.)

Holmstrom, Bengt and Paul Milgrom (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *JL Econ. & Org.*, 7:24. (Cited on page 6.)

Holmstrom, Bengt and Paul Milgrom (1994). The firm as an incentive system. *The American economic review*, pages 972–991. (Cited on page 6.)

Judge, Kathryn (2020). Stress Testing During Times of War. *SSRN Electronic Journal.* (Cited on page 6.)

Kamenica, Emir and Matthew Gentzkow (2011). Bayesian Persuasion. *American Economic Review*, 101(6):2590–2615. (Cited on page 6.)

Kasy, Maximilian (2023). Algorithmic bias and racial inequality: A critical review. (Cited on page 5.)

Kasy, Maximilian and Rediet Abebe (2021). Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 576–586. (Cited on page 5.)

Kearns, Michael and Aaron Roth (2019). *The ethical algorithm: The science of socially aware algorithm design.* Oxford University Press. (Cited on page 5.)

Keys, Benjamin J., Tanmoy Mukherjee, Amit Seru, and Vikrant Vig (2010). Did Securitization Lead to Lax Screening? Evidence from Subprime Loans. *The Quarterly Journal of Economics*, 125(1):307–362. (Cited on page 5.)

Keys, Benjamin J., Amit Seru, and Vikrant Vig (2012). Lender Screening and the Role of Securitization: Evidence from Prime and Subprime Mortgage Markets. *Review of Financial Studies*, 25(7):2071–2108. (Cited on page 5.)

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan (2018a). Algorithmic fairness. In *Aea papers and proceedings*, volume 108, pages 22–27. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203. (Cited on page 4.)

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein (2018b). Discrimination in the age of algorithms. *Journal of Legal Analysis*, 10:113–174. (Cited on page 19.)

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein (2019). DISCRIMINATION IN THE AGE OF ALGORITHMS. *Oxford University Press*, page 62. (Cited on pages 15 and 17.)

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv:1609.05807*. (Cited on page 8.)

Korinek, Anton and Avital Balwit (2022). Aligned with whom? direct and social goals for ai systems. Technical report, National Bureau of Economic Research. (Cited on page 5.)

Lakkaraju, Himabindu and Osbert Bastani (2020). "How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pages 79–85, New York, NY, USA. Association for Computing Machinery. (Cited on page 6.)

Lakkaraju, Himabindu, Ece Kamar, Rich Caruana, and Jure Leskovec (2019). Faithful and Customizable Explanations of Black Box Models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pages 131–138, New York, NY, USA. Association for Computing Machinery. (Cited on page 6.)

Lambrecht, Anja and Catherine Tucker (2019). Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management science*, 65(7):2966–2981. (Cited on page 5.)

Leuz, Christian and Robert E. Verrecchia (2000). The Economic Consequences of Increased Disclosure. *Journal of Accounting Research*, 38:91–124. (Cited on page 6.)

Li, Danielle, Lindsey R Raymond, and Peter Bergman (2020). Hiring as exploration. Technical report, National Bureau of Economic Research. (Cited on page 5.)

Liang, Annie, Jay Lu, and Xiaosheng Mu (2023). Algorithm design: A fairness-accuracy frontier. *arXiv preprint arXiv:2112.09975*. (Cited on page 5.)

Lipton, Zachary C (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57. (Cited on page 5.)

Little, Camille Olivia, Michael Weylandt, and Genevera I Allen (2022). To the fairness frontier and beyond: Identifying, quantifying, and optimizing the fairness-accuracy pareto frontier. *arXiv preprint arXiv:2206.00074*. (Cited on page 5.)

Lundberg, Scott M and Su-In Lee (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30. (Cited on pages 3, 6, and 9.)

Melumad, Nahum D and Toshiyuki Shibano (1991). Communication in settings with no transfers. *The RAND Journal of Economics*, pages 173–198. (Cited on pages 6 and 8.)

Menon, Aditya Krishna and Robert C Williamson (2018). The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*, pages 107–118. PMLR. (Cited on page 5.)

Meursault, Vitaly, Daniel Moulton, Larry Santucci, and Nathan Schor (2022). One threshold doesn't fit all: Tailoring machine learning predictions of consumer default for lower-income areas. (Cited on page 5.)

Murdoch, W. James, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080. (Cited on page 6.)

Rambachan, Ashesh, Jon Kleinberg, Sendhil Mullainathan, and Jens Ludwig (2020). An Economic Approach to Regulating Algorithms. Technical Report w27111, National Bureau of Economic Research. (Cited on page 5.)

Rambachan, Ashesh and Jonathan Roth (2019). Bias in, bias out? evaluating the folk wisdom. *arXiv preprint arXiv:1909.08518*. (Cited on page 5.)

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv:1602.04938*. (Cited on page 6.)

Rudin, Cynthia (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215. (Cited on page 5.)

Slack, Dylan, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju (2020). Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pages 180–186, New York, NY, USA. Association for Computing Machinery. (Cited on page 6.)

Sood, Gaurav and Suriyan Laohaprapanon (2018). Predicting Race and Ethnicity From the Sequence of Characters in a Name. *arXiv:1805.02109*. (Cited on page 20.)

Sun, Jian (2021). Algorithmic transparency. (Cited on page 5.)

Williams, Basil (2017). Stress Tests and Bank Portfolio Choice. (Cited on page 6.)

Yang, Crystal S and Will Dobbie (2020). Equal protection under algorithms: A new statistical and legal framework. *Mich. L. Rev.*, 119:291. (Cited on page 5.)