Unpacking the Black Box: Regulating Algorithmic Decisions

Laura Blattner
Stanford University

Scott Nelson
University of Chicago

Jann Spiess
Stanford University

July 21, 2021

Comments welcome.

Abstract

We characterize optimal oversight of algorithms in a world where an agent designs a complex prediction function but a principal is limited in the amount of information she can learn about the prediction function. We show that limiting agents to prediction functions that are simple enough to be fully transparent is inefficient as long as the bias induced by misalignment between principal's and agent's preferences is small relative to the uncertainty about the true state of the world. Ex-post algorithmic audits can improve welfare, but the gains depend on the design of the audit tools. Tools that focus on minimizing overall information loss, the focus of many post-hoc explainer tools, will generally be inefficient since they focus on explaining the average behavior of the prediction function rather than sources of *mis*-prediction, which matter for welfare-relevant outcomes. Targeted tools that focus on the source of incentive misalignment, e.g., excess false positives or racial disparities, can provide first-best solutions. We investigate the empirical relevance of our theoretical findings using an application in consumer lending.

We thank Mario Curiki and Ruying Gao for outstanding research assistance. We thank Susan Athey, Talia Gillis, Melissa Koide, Amit Seru, Ken Singleton, PR Stark, Chenzi Xu, the FinRegLab team and seminar participants at Stanford University and the OCC for helpful discussions and comments, and 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.

1 Introduction

Decision-makers increasingly rely on complex prediction algorithms to make high-stakes decisions. The adoption of algorithmic decision-making in settings such as lending, medical testing, college admissions, pre-trial detention, and hiring raises new questions about the oversight of such algorithms. In many settings, potential incentive conflicts arise between the agents building the prediction tools and the entities tasked with overseeing their use. An insurance company might worry about a hospital's prediction model over-predicting the risk of heart attack leading to costly over-testing. A financial regulator might worry about lenders' risk models under-predicting credit risk to enable increased leverage. A large employer might worry about a hiring agency using a prediction model that produces low job offer rates for minority job applicants. A key challenge for algorithmic oversight is to determine how to use ex-post audits that can answer the following question: are undesired outcomes due to agents acting on misaligned incentives, or due to other circumstances? For example, high loan default rates could stem from a lender's deliberate risk-taking or could be driven by an unanticipated onset of a recession. Hiring outcomes that appear discriminatory could arise either due to a model that penalizes variables correlated with minority status or an unanticipated shift in the distribution of job applicants.

This paper argues that algorithmic governance faces a new trade-off between complexity and oversight. It is sometimes thought that when algorithms replace humans, decision processes become easier to audit. However, in practice the complexity of state-of-the-art prediction algorithms implies that entities overseeing the use of algorithms have to rely on simplified representations of the prediction functions generated by these algorithms – colloquially also known as "post-hoc explainer tools." Since these tools, by design, cannot preserve all information about the underlying prediction function, algorithmic governance faces a complexity–oversight trade-off. We can restrict algorithms to produce prediction functions that are simple enough to be fully transparent, e.g., a ten-variable logit model, but sacrifice the predictive performance that complex algorithms provide. Alternatively, we can allow complex algorithms but sacrifice some of their ability to understand the model and detect actions that arise from incentive misalignment.

We show how to optimally navigate this complexity–oversight trade-off in a principal-agent setup. In most settings, the optimal policy imposes no ex-ante restriction on the agent's prediction function but conducts an ex-post audit based on an explainer tool. This policy is optimal as long as the bias induced by misalignment between the principal's and the agent's preferences is small relative to the uncertainty about the true state of the world. Intuitively, the agent can induce some distortion ("bias") in the prediction function to achieve their preferred outcome, e.g., they might develop a prediction function that under-predicts default risk in order to take on more credit risk than a regulator would prefer. Ex-ante model restrictions, in contrast, limit an agent's ability to adapt their prediction function to new information about the state of the world, leading to high-variance losses that are (ex-ante) inefficient. Algorithmic audits can improve this bias–variance trade-off because they limit the ability of the agent to distort the prediction function. The optimal algorithmic audit is a targeted explainer that requests information not about what drives the *average* prediction but instead about what drives particular types of *mis*-prediction. Intuitively, the principal uses her knowledge about the sources of misalignment between her and the agent's preferences to inspect parts of the model that are most likely to drive model distortions.

We investigate how our optimal policy prescriptions perform using an empirical application from unsecured consumer lending. Unsecured consumer lending provides an excellent case study because algorithmic credit underwriting is becoming widely used and the realities of regulatory oversight fit our theoretical model well. We design an empirical optimization problem that builds a default prediction model that is maximally distorted in the direction of an agent's misaligned incentives, subject to regulatory constraints. Using a large-scale credit data set, we investigate the complexity–oversight trade-off in the data, and estimate outcomes under different algorithmic audits based on simple explanations.

The key assumption in our work is that algorithmic audits are limited in the amount of information they can reveal about the underlying prediction function. Concretely, we assume that an algorithmic audit reveals only a lower-dimensional representation of the underlying prediction function. This assumption can be motivated by lack of resources or sophistication by the entity overseeing the use of algorithms, by potential legal limitations that prevent regulators from collecting more information, or simply the fact that even the data scientists who build the prediction tool will struggle to fully describe or "explain" highly complex ML/AI models. This assumption also reflects current regulatory exam practices settings such as credit underwriting which rely on simple approximations of the underlying prediction functions.

We approach the question of optimal algorithmic regulation through the lens of a principal-agent framework with asymmetric information, similar to the classic approach of Laffont and Tirole (1993). A principal delegates a prediction task to an agent who builds a statistical model to predict an unknown outcome for a set of individuals using a set of observable characteristics. Both the true data generating process for the outcome of interest and the distribution of observables are unknown ex-ante and are only realized when the prediction model is deployed. This incomplete information allows us to capture the idea that external circumstances may change subsequent to model development. This assumption also reflects that prediction functions are typically not trained on the data on which they are deployed, which makes the potential deterioration of scores due to unforeseen shifts in the deployment data ("covariate shift") or shifts in the data generating process ("model shift") a key concern.

A principal can impose ex-ante restrictions on what types of prediction functions the agent builds and can conduct ex-post audits based on simple descriptions of the prediction function, e.g. she might see a logit model with a handful of variables that approximates an underlying model with hundreds of variables. In designing optimal policy, the principal faces several sources of asymmetric information. First, she faces ex-ante uncertainty about the state of the world when the prediction function is deployed. The state of the world includes both the parameters that govern the data generating process of the outcome of interest, e.g., loan default, worker productivity, a heart attack, or recidivism, and the joint distribution over the observables used in the prediction function, e.g., credit history, job applicants' materials, medical files, or criminal records. Second, she does not know which agents have misaligned preferences. The third source of asymmetric information arises because the principal only observes a noisy public signal generated by the prediction function chosen by the agent, e.g., the amount of disparity across racial groups generated when the prediction function is deployed. Asymmetric information about the state of the world motivates the delegation of building the prediction function to the agent, who receives a signal about the state of the world (a training dataset) prior to designing a prediction function. The second and third type of asymmetric information introduce the hidden-action problem, motivating the need for understanding the prediction function as opposed to simply punishing bad realized outcomes.

The timing of the game between principal and agent is as follows. In the initial rule-setting stage, the principal specifies restrictions on the type of prediction functions and the nature of the ex-post audit to maximize expected welfare. For example, the principal might restrict admissible models to low-dimensional models that can be fully audited and specifies that audits will take the form of a linear projection of the underlying model on a small set of specified variables. In this initial stage, the principal has information about the distribution of potential states of the world. In the subsequent training stage, the agent receives a signal about the state of the world and trains a prediction function to maximize agent utility subject to the restrictions imposed by the principal. This prediction function is then deployed in a third stage, when the state is realized and outcomes are realized. In the fourth stage, the principal audits the prediction function according to her specified audit tool and decides whether the agent passes the audit. Failing the audit imposes infinite negative utility on the agent. In the final stage, welfare (for the principal) and private payoffs (for the agent) are realized. In our main theoretical illustration, welfare and private payoffs take the form of a quadratic loss function around a bliss point which represents a best-fit prediction function – with potential additive terms that reflect additional preferences that drive misalignment, e.g. the principal might have an additional term in her welfare function that reflects the costs of systematic financial risk or a taste for equality across social groups.

We derive several key results from our theoretical model. First, it is optimal not to impose any ex-ante restriction on the prediction function as long as the bias induced by the misaligned agent is small relative to the uncertainty about the state. Intuitively, principals face a type of bias–variance trade-off. If agents are allowed flexibility in their prediction function, misaligned types induce bias to distort the prediction function in their preferred direction. As in our earlier examples, a lender might create a model that systematically under-predicts default for risky subprime loan applicants the lender wants to approve. Or a hiring agency might create a model that under-predicts job performance for minority job applicants relative to the model that the equality-minded employer would prefer. However, the flexibility also allows the agent to adjust their prediction function to the information about the state of the world and, thereby, to reduce variance. Restricting the prediction function can remove the bias but at the cost of a prediction function that is potentially ill-suited for the realized state.

Second, the ability to conduct ex-post audits leads to welfare improvements because it enables at least some partial alignment of preferences. An explainer that focuses on preserving the most information about the average behavior of the prediction function – we term this the "best prediction explainer" – is generally inefficient. Instead, the optimal explainer targets the areas of preference misalignment. Intuitively, the targeted explainer uses knowledge about the preference misalignment to inspect parts of the prediction function that are most likely to reflect the preference misalignment and are also relevant for welfare. Third, this targeted explainer can achieve the first best as long as the preference misalignment is low-dimensional. Intuitively, if the agent wants to distort a slope but not an intercept (one-dimensional misalignment) and the explainer can produce one piece of information (one-dimensional explainer), then a targeted explainer can produce a first-best prediction function. If the dimensionality of the preference misalignment exceeds that of the explainer, the targeted explainer provides a second-best solution.

Finally, it is generally not optimal to regulate based purely on the realized outcomes on the deployment data, e.g., have agents fail their audit if and only if disparities in approval rates across social groups are high, or if realized loan defaults are high. Intuitively, this approach will force the aligned agent to build more restrictive models, which have the same realized outcomes in terms of group parity or false positives across deployment states but at the cost of overall fit. For example, the agent might be forced to build a model that rejects qualified job applicants from one social group in order to ensure that approval disparities are small even under a deterioration in the applicant quality of another social group. Similarly, a lender might adopt a very conservative model rejecting all risky subprime loan applicants in order to limit the number of loans given to subprime applicants in the event of a recession.

We investigate the empirical relevance of our theoretical results by building an empirical counterpart to our model inspired by recent advances in the adaption of Generative Adversarial Networks in economics and econometrics (e.g. Athey et al., 2020a). In our empirical setup, a lender builds loan default prediction functions while a regulator tries to discern whether each prediction function (or credit scoring function) is aligned or misaligned based on an explanation of the model. Similar to our theoretical model, we build deployment data sets that can differ from the training data in a way that generates adverse welfare-relevant outcomes even for an aligned model. For example, minority applicants in the deployment data may be higher risk than in the training data, and rejections of these high-risk applicants may then drive up disparate impact statistics even for an aligned model. To test the ability of different types of regulation to limit the ability of a decision-maker to build a misaligned model, we specify an optimization problem that solves for the prediction model that is maximally distorted in the direction of an agent's misaligned preferences subject to regulatory constraints. We solve this optimization problem using gradient descent in TensorFlow, considering complex neural networks as well as simple logistic regression.

Our empirical counterpart focuses on the setting of unsecured consumer lending, and in particular a random sample of credit reports with newly opened credit cards. Unsecured consumer lending is a good setting to study for at least four reasons: algorithmic decision-making is already in use by some lenders; large-scale data is available; two leading types of preference misalignment, excess risk-taking and disparate impact, that we study are key regulatory concerns; and our model assumptions fit the reality of consumer regulation well. Moreover, credit cards in particular are a widely used credit product for which algorithmic underwriting is already in use by some providers. Our credit scoring model pipeline mimics that of a real-world decision-maker in terms of variable cleaning, variable selection, model tuning, and richness of underlying data.

In our empirical exercise, as in our theoretical framework, we study different regulatory restrictions relative to a baseline where there is no regulation and agents can build a maximally distorted model. Specifically, we compare outcomes under simple (linear-regression) and complex (neural-network) models, and consider audits based on a best-prediction explainer and based on a targeted explainer. Our detailed empirical findings are currently under review by the data provider.

Literature. Our work contributes to a nascent literature that studies algorithmic decision-making (e.g. Athey et al., 2020b) and how to regulate it. Most of the existing work in this area has focused on questions of algorithmic fairness, such as work by Gillis and Spiess (2019) and Gillis (2020) on the limits and design of algorithmic audits. Most related to our approach, Rambachan et al. (2020) study the regulation of algorithmic fairness in principal-agent framework. We make three contributions: First, we offer a framework that nests many types of potential incentive misalignment, including many types of distributional objectives as well as diverging risk preferences. Second, many contributions on algorithmic audits 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, such as a simplified representation of the credit scoring model. Given the complexity of ML/AI tools and potential limitation on the technical or legal reach of regulators, we believe it is important to study optimal algorithmic regulation under informational constraints. Third, we provide empirical validation for our theoretical results in a real-world dataset.

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 et al., 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 et al., 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 a lender to sacrifice model flexibility in ways that reduces, rather than increases, welfare.

Our work is related to a large literature on asymmetric information (Myers and Majluf, 1984; Nachman and Noe, 1994; Diamond and Dybvig, 1983; Leuz and Verrecchia, 2000; Greenstone et al., 2006) and disclosure in the financial system (for a recent literature review see Goldstein and Leitner, 2020). Similar to this literature, we consider the design of disclosure (or audit) in an asymmetric information setting and study a regulator who has to decide ex-ante what rules to follow when presented with information ex-post. Our main contribution lies in studying disclosure when decisions are automated and based on complex risk prediction algorithms. We assume there are (technical or political) limitations on the amount of information the regulator can obtain about the algorithms and ask what the optimal audit 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. In addition, since the result of the audit in our setting is not publicly disclosed, we abstract from frequently questions about the design of optimal *public* disclosure (Goldstein and Leitner, 2017; Faria-e Castro et al., 2017; Williams, 2017; Judge, 2020).

We also contribute to a body of work on the role of credit scoring models in US consumer finance. Most of this work explores properties of credit scores and consequences for welfare in settings where credit scores are used for decision-making, e.g. the role of credit scores in overcoming asymmetric information among new borrowers (Einav et al., 2013; Adams et al., 2009), incentivizing loan repayment (Chatterjee et al., 2020), disparities in credit misallocation due to differential informativeness of credit scores (Blattner and Nelson, 2021) and facilitating loan securitization while discouraging lenders' use of soft information (Keys et al., 2012, 2010). Recent work has also warned that more flexible statistical technology such as machine learning can reduce overall loan approval rates for disadvantaged groups (Fuster et al., 2019), and that modern FinTech underwriting continues to generate cross-group disparities in loan terms (Bartlett et al., 2019); credit scores likewise are seen to play a role in geographic misallocation in the US mortgage market (Hurst et al., 2016). Much of this work echoes persistent policy concerns about equity across consumers in credit scoring (Avery et al., 2009, 2012; Traub, 2013). 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.

This article is organized as follows: Section 2 illustrates our model and main results with simple examples. Section 3 details our empirical analysis. Section 4 sets up the full theoretical model. Section 5 concludes.

2 Model: Illustrative Example

This section presents a simplified version our theoretical setup in order to illustrate our main results. We first describe the setup and timing of the model. We then characterize different regulatory policies in a series of examples. Section 4 provides the full model and general results.

2.1 Setup

A principal delegates a prediction task to an agent who builds a statistical model to predict an unknown outcome for a set of individuals. Each individual has a vector of characteristics $X \in \mathcal{X}$ and an outcome Y. In this illustrative example, we assume there are only two binary observables, $\mathcal{X} = \{(X_1, X_2); X_1, X_2 \in \{0, 1\}\}$, with joint distribution μ.

The outcome Y is generated by a data-generating process that depends on $s(X)$, which can be seen as the input to an inverse link function (e.g. $Y = s(X) + \varepsilon$ if the outcome Y is continuous or $Y = \text{logit}^{-1}(s(X)) + \varepsilon$ if the outcome Y is binary). In this two-variable prediction setup, the data-generating process is governed by

$$s(X_1, X_2) = \alpha + \beta\, X_1 + \gamma\, X_2 + \delta\, X_1 \cdot X_2. \tag{1}$$

Since both observables are binary in this example, writing the scoring model as a fully interacted linear regression is not restrictive.

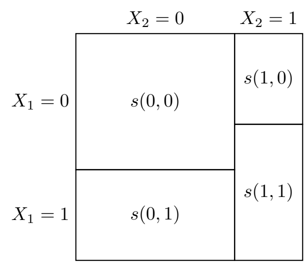

Figure 1: Schematic representation of two-variable prediction setup

The agent builds a prediction function $f(X)$ of $s(X)$, which we can express analogously as a fully interacted linear regression

$$f(X_1, X_2) = a + b\, X_1 + c\, X_2 + d\, X_1 \cdot X_2. \tag{2}$$

For simplicity, we formulate our model in terms of preferences over predictions f of this score s (rather than predictions of Y), and we model the statistics of learning about s from data only implicitly.

Example: Medical Testing An insurance company decides how to reimburse costs for a medical test conducted on the basis of a statistical prediction model. For example, the statistical model may predict the presence of a heart attack in patients visiting the emergency room in order to guide decisions on more invasive testing (Mullainathan and Obermeyer, 2019). The outcome is whether a patient is experiencing a heart attack. The characteristics X summarize information about symptoms currently exhibited by the patient, results from simpler testing (e.g. electrocardiograms or troponin testing) as well as past medical history, age and gender. In this example, we will define X_1 as an indicator variable for whether the patient had a heart attack in the past and X_2 as an indicator variable for whether EKG results returned normal.

Example: Lending A agent builds a credit scoring model to assess default risk of loan applicants subject to oversight by a financial regulator, such as the Federal Reserve or the Consumer Financial Protection Bureau. The outcome is whether an approved borrower defaults on the loan. The characteristics X summarize information in the applicant's credit report, such as past repayment behavior and current credit utilization. In this example, we will define X_1 as an indicator variable for whether the loan applicant had a past default and X_2 as an indicator variable for whether the applicant currently has high credit utilization.

Example: Hiring A large employer uses a hiring agency to hire short-term employees. The hiring agency uses a statistical model to screen job applicants (Hoffman et al., 2018; Li et al., 2020). The outcome is the performance on the job. The characteristics X are comprised of two sets of variables: First, $\tilde{X}$, which summarize information in the applicant's education, socio-demographic characteristics, aptitude tests, and past employment history. Second, an indicator for whether an applicant belongs to a minority group G, which is not used in the prediction function.[1] In this example, we will define X_1 as an indicator variable for whether the job applicant has relevant job experience and X_2 as an indicator variable for whether the applicant has a high-school degree.

Information structure. The agent and principal both face incomplete information. The parameters governing the data-generating process $s(X)$ as well as the distribution μ of observables X are ex-ante unknown to both parties. They jointly make up the deployment state $d = (s, \mu)$. The distribution over the deployment state is ex-ante known by both agent and principal. The deployment state is realized in the final stage of the game when the prediction function is deployed. Outcomes will depend both on the realization of the deployment state as well as on the prediction function f chosen by the agent. Figure 1 provides a visualization of the deployment state. The size of the four cells in the two-by-two matrix represents the distribution over observables, μ, while the score $s(X)$ in each cell describes the conditional distribution driving the outcome of interest.

Agent and principal receive a training signal about the parameters $d = (s, \mu)$ governing the deployment state. In this simple example, we assume that the training signal fully reveals the parameters of the data generating process $s(X)$. Once the agent has observed the training signal and chosen the prediction function f based on this signal, both principal and agent receive a public signal $g(d; f)$ about the realized deployment state.

The principal faces an additional dimension of incomplete information. She does not know the preferences of the agent. In particular, she does not know which agents have preferences misaligned relative to social preferences and which agents do not. In contrast, the agent knows the principal's preferences.

Example: Medical Testing $s(X)$ is the data generating process for the presence of heart attack in ER patients and μ is the joint distribution of patient observables. The medical practitioner who builds the statistical model knows neither the data generating process nor the distribution of patients who will visit the ER when

[1] We take as given that existing anti-discrimination laws prohibit the use of protected class information in the use of predictive models in settings such as hiring.

the prediction function is deployed. The noisy signal about the deployment state is received in the form of data about past patients and outcomes that is used for training the statistical model. Once the prediction function is deployed, the public signal $g(d; f)$ could be the fraction of patients recommended for further testing based on the prediction model who test positive for heart attack when the prediction model is deployed in the ER. This yield depends both on the realized data generating process, the realized distribution of patients who visit the ER when the prediction function is deployed, and the prediction function that determines which patients are recommended for further testing.

Example: Lending $s(X)$ is the data generating process for default and μ is the distribution of loan applicant observables. The lender who builds the credit scoring model does not know the data generating process or the exact distribution of loan applicants they will encounter when they deploy the credit scoring function. The noisy signal about the deployment state is received in the form of training data from either past loans made by the lender or credit bureau data on loans made by other lenders. The public signal $g(d; f)$ is the fraction of loans that default among borrowers with particular values of X, for example those with high credit utilization. The signal depends both on the realized data generating process, the realized distribution of individuals who apply for loans when the credit scoring function is deployed, and the credit scoring function that forms the basis of the lender's underwriting decision.

Example: Hiring $s(X)$ is the data generating process for job performance and μ is the joint distribution of job applicant observables (including minority status G). The hiring agency building the prediction model does not know the data generating process or the exact distribution of job applicants when the prediction tool is deployed. The noisy signal about the deployment state is received in the form of training data from past employees provided by the employer. The public signal $g(d; f)$ is the difference in predicted job performance by minority status. This signal depends both on the realized data generating process, the realized distribution of individuals who apply for jobs when the prediction tool is deployed, and the prediction function used by the hiring agency to score each applicant.

Agent's Problem. The agent builds a prediction function f to maximize expected payoff in the deployment state, given the principal's constraints $(\mathcal{F}, \mathrm{E}, \mathbb{1}_{\text{audit}})$, which we describe below. The agent's payoff is given by

$$\max_{f \in \mathcal{F}} \mathrm{E}_{\mu} U_{\theta}(f; d) - c\mathbb{1}_{\text{fail audit}}(\mathrm{E}f)$$

This payoff depends on the agent's preferences $U_\theta(d)$, the deployment state d, and the cost c of failing the principal's audit. The agent's type θ in U_θ expresses potential sources of preference misalignment. Preference misalignment implies that the agent has different preferences over predictions, either for the whole set of individuals or a sub-group of individuals defined by certain values of X.

We further assume that the agent's utility takes the form of a quadratic loss function

$$U_\theta(f; d) = -\,\mathrm{E}_\mu[(f(X) - s(X) - \theta(X))^2],$$

where $\theta(X)$ depends on the nature of misalignment. Intuitively, this modeling choice offers a tractable way of capturing that the agent cares about accurately assessing the outcome, with some potential deviations due to misalignment either for all individuals or for particular subgroups.[2]

Example: Medical Testing Some medical practitioners (agents) have misaligned preferences relative to the insurance company (principal). These practitioners build a model that predicts a higher risk of heart attack on average relative to a statistical model that simply maximizes predictive fit given the available data. This incentive misalignment could be driven by moral hazard that incentives over-testing either because insurers pay by the test or because malpractice lawsuits push to excess caution (see Greenberg and Green (2014) and O'Sullivan et al. (2018) for evidence in the medical literature, or Acemoglu and Finkelstein (2008) and Baker (2001) for related work in economics, as cited in Mullainathan and Obermeyer (2019)). Alternatively, the moral hazard effect could enter indirectly through properties of the data used for training the statistical model (Mullainathan and Obermeyer, 2017).

The misaligned agent's preferences are then given by

$$U_\theta(f; d) = -\,\mathrm{E}_\mu[(f(X) - s(X) - \Delta_{\text{overall}})^2]$$

where $\theta(X)$ takes the form Δ_{overall} to express the preference for over-predicting heart attack risk on average.

Example: Lending Some lenders (agents) have misaligned preferences relative to the financial regulator (principal) driven by a different taste for risky loans. Due to moral hazard induced by government deposit insurance, the lender would like to build a model that under-predicts default, or equivalently generates higher credit scores, on average in order to be able to approve more risky loans. Alternatively, the lender might prefer a prediction function that under-predicts default risk, or equivalently generates higher credit scores, for loan applicants with

[2] This representation of preferences can also represent a second-order Taylor approximation of more conventional preference specifications. This implies that this formulation imposes only the limitation that only linear and quadratic terms matter for utility.

high credit utilization, if the lender has higher payoffs from these borrowers (e.g., high fees related to credit-card utilization) than the welfare that the regulator ascribes to loans for these borrowers. The misaligned agent's utility is given by

$$U_\theta(f;d) = -\,\mathrm{E}_\mu[(f(X) - s(X) - (\Delta_{\text{overall}} + \Delta_{\text{high utilization}}\, X_2))^2]$$

Here, $\theta(X) = \Delta_{\text{overall}} + \Delta_{\text{high utilization}}\, X_2$, where X_2 is the indicator variable for high credit utilization.

Example: Hiring Some hiring agencies (agents) have misaligned preferences relative to the employer (principal) who has a preference for equality in job offer rates across (racial/ethnic) minority and majority job applicants. The misaligned hiring agency builds a prediction model that simply maximizes fit without regard to the distributional consequences of its prediction model $U_\theta(d) = -\,\mathrm{E}_\mu[(f(\tilde{X}) - s(\tilde{X})]$. Recall, that we assume that minority status G does not directly enter the prediciton function. In contrast, the employer would like to introduce additional terms in the utility function that reflect her preference for greater equality across social groups.

Principal's problem. The principal designs regulation $(\mathcal{F}, \mathrm{E}, \mathbb{1}_{\text{fail audit}})$ to maximize expected welfare $\mathrm{E}W(f;d)$ in the deployment state. We assume that welfare, similar to the agent's problem, takes the form of a quadratic loss function:

$$W(f;d) = -\,\mathrm{E}_\mu[(f(X) - s(X))^2] - \lambda\, \ell(f;d)$$

where $\lambda\, \ell(f;d)$ expresses an additional loss term that arises when the principal's preferences diverge from the best-fit benchmark, e.g., a taste for equality across social groups.

The principal has two regulatory tools available: ex-ante functional restrictions and ex-post audits. First, the principal can restrict the function space over which the agent can choose when building a prediction model. In this illustrative example, the model restriction represents a linear restriction to a lower-dimensional space:

$$\mathcal{F} = \{f \in \mathbb{R}^n; Af = a\}, \qquad\qquad A \in \mathbb{R}^{m\times n}, a \in \mathbb{R}^m.$$

Model restrictions represent ex-ante restrictions that are not sensitive to the training signal, but can constrain the full function. For example, in this setting the principal might restrict scoring functions to simple regressions $f(X) = a + b\, X_1 + c\, X_2$ without an interaction term.

Second, the principal designs an ex-post audit tool in the form of an explanation mapping E. An explanation mapping (short: explainer) is a low-dimensional representation of the prediction function. In our example, we assume that the principal chooses a two-dimensional projection of the scoring function:

$$\text{E} : \mathbb{R}^n \ni f \mapsto \text{E}f \in \mathbb{R}^k, \qquad \text{E} \in \mathbb{R}^{n\times k} \qquad \text{(here: } n = 4, k = 2)$$

This explainer captures the idea that the function f is represented in terms of a few key features that capture the first-order behavior of f, but not all details.

The principal chooses whether or not to fail the agent after conducting the ex-post audit. We assume that an audit will always take place and we leave modeling conditional audits for future work.

Additional simplifications. We make a few additional assumptions for tractability in this example only. For simplicity, we assume that X_1 and X_2 are uncorrelated, $\text{P}_\mu(X_1 = 1) = .5, \text{P}_\mu(X_2 = 1) = p \neq .5$, and that the distribution μ is known. In addition, we assume the jointly Normal form $\mathbb{R}^4 \ni s \sim \mathcal{N}(\bar{s}, \Sigma)$ for the parameter distribution that govern default, where

$$\Sigma = \underbrace{\sigma_0^2 \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}}_{\text{correlation across all cells}} + \underbrace{\sigma_1^2 \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}}_{\text{correlation within } X_1 \text{ cells}} + \underbrace{\sigma_2^2 \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}}_{\text{correlation within } X_2 \text{ cells}} + \underbrace{\sigma_3^2 \mathbb{I}}_{\text{cell-specific variation}}$$

where $\sigma_0^2 > \sigma_1^2 \geq \sigma_2^2 > \sigma_3^2$. For this illustration, we also assume that the agent type θ is fixed, and that the data generating process s is fully revealed to the agent in the training stage.

Information loss and explainers. The key assumption in this setup is that the principal is limited in the amount of information she can learn about the prediction function through the explainer. We express this information loss in the assumption the explainer is a two-dimensional projection of the four-dimensional prediction function.

In Figure 2, we show the two possible explainer mappings the principal can choose in this simple example. Here, the principal can for instance learn about the average difference in predicted values across the X_1 cells, *or* she can learn about the average difference in the X_2 dimension (corresponding to the two different solid lines). For sake of exposition, we assume the variance of X_1 is always greater than the variance of X_2 (individuals are more evenly split between $X_1 = 0$ and $X_1 = 1$ than between $X_2 = 0$ and $X_2 = 1$).[3]

Definition (Best prediction explainer)**.** *The explanation mapping that preserves the most information about the overall behavior of the prediction function.*

[3] In principle, our general model would also allow for explainers that combine the effect of multiple variables into a single parameter. To simplify the exposition, we have chosen this example so that all explainers have a simple representation.

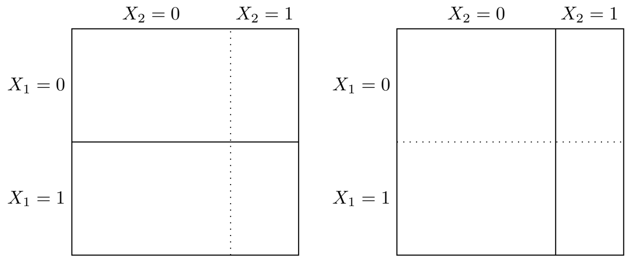

Figure 2: Two of the possible explanation mappings

In this setting, the explainer that summarizes the difference in predicted values across the X_1 cells is the best prediction explainer. Intuitively, since we assume the distribution of individuals across values of X_1 is more even than the distribution across values of X_2, we learn more about the behavior of the model if we split according to the dimension that results in more equally sized groups.

Example: Medical Testing The insurance company decides under what conditions to reimburse costs for invasive tests for heart attack in a setting where testing decisions are based on a statistical prediction model. The company's preferences are given by $W(f; d) = -\,\mathrm{E}_\mu[(f(X) - s(X))^2]$, that is, the company prefers a prediction model that maximizes fit. In order to guard against prediction functions that induce over-testing, the insurance company can decide to limit the admissible types of prediction models and can design an ex-post audit of the prediction model. The insurer's key constraint is that it is limited in the amount of information it can request about the prediction model in the audit. This limitation could arise because the prediction model is proprietary information or because the insurer has limited resources to conduct an audit. The insurer might specify that the only admissible prediction functions are logit models with up to five variables that can be fully inspected in the ex-post audit. Alternatively, the insurer might not specify any restrictions on the admissible prediction function but conduct an audit based on a simpler proxy model of the prediction model.

Example: Lending A financial regulator designs regulatory policy for the lending industry which relies on credit scoring models for loan underwriting. The regulator's preferences are given by $W(f; d) = -\,\mathrm{E}_\mu[(f(X) - s(X))^2]$, that is she prefers a prediction model that maximizes fit. The regulator can issue guidance on the types of credit scoring models allowed and can design an ex-post audit scheme. The regulator's key constraint is that she is limited in the amount of information she can request about the credit scoring model either because she has limited resources to process information or because her legal mandate limits her ability to request more in-depth information from the lender. Financial regulators frequently issue guidance on what is admissible in credit scoring with the most important legislation being the Fair Credit Reporting Act and Equal Credit Opportunity

Act. Enforcement of these regulations relies on regulatory exams that are typically conducted after a credit scoring model has been built and inspect summary statistics and proxy models of the underlying credit scoring models.

Example: Hiring An employer designs criteria for using a third-party hiring agency. The employer's preferences are given by

$$W(f;d) = -\,\mathrm{E}_\mu[(f(x) - s(x))^2] - \lambda\,(\mathrm{E}_\mu[f(X)|G=0] - \mathrm{E}_\mu[f(X)|G=1])^2.$$

where the final term expresses that she has a distaste for inequality across protected groups, such as ethnic and racial minorities. The employer can decide to specify the type of prediction function she wants the hiring agency to employ and can design an ex-post audit tool. The employer's key constraint is that she is limited in the amount of information she can request about the prediction model either because the hiring agency's model is proprietary or because she has limited resources to analyze the model. She might specify that the only admissible prediction functions are logit models with up to five variables that can be fully inspected in the ex-post audit. Alternatively, she might not specify any restrictions on the admissible prediction function but conduct an audit based on a simpler proxy model of the prediction model.

Timing. The game between principal and agent has the following stages.

1. Rule-setting stage. The principal observes the joint distribution over the deployment state d and types θ, and chooses a model restriction $\mathcal{F}$ and an explanation mapping E to maximize expected welfare.

2. Training stage. Agents observe the type θ and receive a signal about the deployment state, and choose a function $f \in \mathcal{F}$ to maximize their expected utility.

3. Deployment stage. The deployments state $d \in \mathcal{D}$ is realized. Along with the deployment state, a public signal $g(f;d)$ about the outcome is realized.

4. Audit stage. The principal audits f based on the signal $g(f;d)$ and the explanation $\mathrm{E}f \in \mathcal{E}$.

5. Payoff stage. Agents' utilities $U_\theta(f;d) - c\mathbb{1}_{\text{audit fails}}$ and welfare $W(f;d)$ are realized.

   (For now, assume $c = \infty$, so that the agent will always avoid a failed audit.)

2.1.1 Optimal Regulatory Policy

Solving the game backwards, we now consider different possible actions for the principal. We show that (i) not imposing any regulation on the agent leads to maximal bias (or distortions) in the prediction function; (ii) restricting the agent to prediction functions that can be fully captured by the best prediction explainer eliminates the bias and but introduces welfare costs from higher variance; (iii) not restricting the scoring function but conducting ex-post audits can achieve the first best depending on the type of audit tool used.

No restrictions, no audit. First, assume there is no regulation (no restriction on the prediction function and no audit). In this case, the misaligned agent will distort the prediction function to reflect his preferences and lead to maximal bias. Expected welfare is reduced due to the bias introduced by the misaligned agents.

Example: Medical Testing The misaligned practitioners distort the prediction model by shifting risk predictions up on average,

$$f(X) = (\alpha + \Delta_{\text{overall}}) + \beta\, X_1 + \gamma\, X_2 + \delta\, X_1 \cdot X_2.$$

Expected welfare is reduced relative to the first-best choice $f \equiv s$ (for which welfare is zero) due to the bias introduced by the misaligned agents,

$$\mathrm{E}[W(f;d)] = -\Delta_{\text{overall}}^2.$$

Example: Lending The misaligned lender distorts the credit score by shifting credit scores up on average and also give applicants with high utilization relatively higher credit scores,

$$f(X_1, X_2) = (\alpha + \Delta_{\text{overall}}) + \beta\, X_1 + (\gamma + \Delta_{\text{high utilization}})\, X_2 + \delta\, X_1 \cdot X_2.$$

Expected welfare is reduced relative to the first-best choice $f \equiv s$ (for which welfare is zero) due to the bias introduced by the misaligned agents,

$$\mathrm{E}[W(f;d)] = -(\Delta_{\text{overall}} + p\Delta_{\text{high utilization}})^2 - p(1-p)\Delta_{\text{high utilization}}^2.$$

Example: Hiring The misaligned hiring agency builds the best fit prediction model. This prediction function leads to an expected welfare loss equal to the employer's weight on distributional equity,

$$\mathrm{E}[W(f;d)] = -\lambda\left(\mathrm{E}_\mu[s(X)|G=0] - \mathrm{E}_\mu[s(X)|G=1]\right)^2.$$

This policy leads to maximal distortion, representing an inefficient welfare outcome.

Ex-ante model restrictions. Second, assume that the principal restricts the admissible prediction functions. This solution is successful at eliminating bias but comes at the cost of introducing variance because the ex-ante model restriction no longer allows the agent to exploit the information in the training signal about the deployment state. The principal faces a form of bias–variance trade-off. If the principal imposes no restriction on the prediction function, she risks that the agent will distort the prediction function in line with his preferences. In contrast, if she restricts the prediction function in order to eliminate this bias, she has to accept higher variance since the prediction function is more likely to mis-predict given the ex-ante restrictions. In the full model below, we show generally that if the size of the preference misalignment is small relative to the variance of the underlying parameters governing the outcome of interest, imposing ex-ante restrictions on the prediction function is never optimal.

Example: Medical Testing The agent and the principal disagree about the intercept term in the prediction function f. If the misalignment is large enough, the principal can fix the overall expectation ex-ante. The misaligned medical practitioner now builds a prediction function of the form

$$f(X) = \underbrace{\mathrm{E}[\alpha + .5\beta + p\gamma + .5p\delta]}_{\text{overall expectation fixed ex-ante}} + b\,(X_1 - .5) + c\,(X_2 - p) + d\,(X_1 - .5)\cdot(X_2 - p)$$

to comply with the restrictions imposed by the principal. While this policy perfectly aligns choices from the restricted set (similar to the aligned delegation solution in Frankel, 2014), it comes at the cost of not fully tracking the realized value of γ, yielding welfare of $\mathrm{E}[W(f;d)] = -\operatorname{Var}(\alpha + .5\beta + p\gamma + .5p\delta)$. The welfare loss is driven by the variance induced by the ex-ante restriction.

Example: Lending When the misalignment between regulator and lender is large and the regulator considers ex-ante restrictions as the only policy options, then the regulator would restrict the agent to build credit scoring functions of the form

$$f(X_1, X_2) = \underbrace{\mathrm{E}[\alpha + .5\beta] + (\mathrm{E}[\gamma + .5\delta])\, X_2}_{\text{overall expectation and effect of } X_2 \text{ fixed ex-ante}} + b\,(X_1 - .5) + d\,(X_1 - .5) \cdot X_2.$$

The agent is still allowed to fit the coefficient on past default X_1 and its interaction with utilization X_2 *after* observing the training signal because the lender and regulator preferences are aligned over this component. The downside of the restriction is that these parameters cannot be adjusted to respond to the realization of the parameters that govern default, and there now is welfare loss due to the variance induced by a scoring function that cannot flexibly respond to the information about the deployment state for a total welfare of

$$\mathrm{E}[W(f;d)] = -\left(\mathrm{Var}(\alpha + .5\beta + p(\gamma + .5\delta)) + p(1-p)\left(\mathrm{Var}(\gamma + .5\delta)\right)\right).$$

Example: Hiring Assume that the first variable X_1, relevant job experience, is uncorrelated with minority status but that the second variable X_2, a high school degree, is negatively correlated with minority status. If the employer's preference for equality is large enough, the employer imposes an ex-ante restriction that fixes the coefficients associated with X_2. The misaligned medical practitioner now builds a prediction function of the form

$$f(X) = a + b\,(X_1 - .5) + \underbrace{\mathrm{E}[\gamma + .5\delta]\, X_2}_{\text{effect of } X_2 \text{ fixed ex-ante}} + d\,(X_1 - .5) \cdot X_2.$$

An extreme version of this restriction would be to force the prediction model to leave out X_2 completely, which would ensure equal distribution of predicted values across groups. While this restriction perfectly aligns choices over the prediction function, it comes at the cost of restricting the agency from adjusting the prediction parameters to the information about the data generating process received in the training signal, leading to welfare loss due to variance induced by the ex-ante restrictions. The expected welfare under this restriction is

$$\mathrm{E}[W(f;d)] = -p(1-p)\,\mathrm{Var}(\gamma + .5\delta).$$

Audit based on ex-post signal. Third, the principal could audit based on the public signal $g(f;d)$. However, the principal faces a challenge. The principal cannot distinguish whether an adverse public signal is likely due to the choice of the prediction function by the agent, or an adverse realization of either the distribution μ over observables X or the parameters of the data generating process $s(X)$. Indeed, we will argue in Section 4 that the principal is unable to rely on the public signal about the realized outcome to take her audit decision: Since the choice f of the agent does not change which signal $g(f;d)$ the principal can observe (even though f makes some outcomes much more likely than others), deciding to fail an audit for any value of $g(f;d)$ would impose expected utility of $-\infty$ on all agents, no matter their choice.

Example: Medical Testing The public signal $g(f;d)$ consists of the fraction of patients who undergo further testing who test positive for heart attack. Penalizing the agent whenever this yield is low is generally inefficient. Since yield could be low either because of a distorted prediction model, or because there were a large number of patients with characteristics that make a heart attack difficult to predict.

Example: Lending The public signal $g(f;d)$ consists of the default rate among a high-risk group of subprime loan applicants. Since only the overall default rate within this group is observed, the regulator cannot distinguish whether an adverse outcome (i.e. a very high default rate) is likely due to the choice of credit score by the lender, or the distribution of subprime borrowers, or an economic shock that increased their propensity to default endogenously.

Example: Hiring The public signal $g(f;d)$ consists of the differences in job performance predictions between minority and majority groups. If there is substantial uncertainty about the distribution of minority job applicants, then the employer will not generally be able to infer whether high disparities $g(f;d)$ is due to the agency's choice of f or the presence of a high number of job applicants with an abnormally low job performance given observables.

Audit based on prediction explainer. Fourth, we consider the case where the principal does not restrict the prediction function, but conducts an ex-post audit based on the best prediction explainer. Recall that the best prediction explainer is equivalent to a linear regression of the predicted values produced by the prediction function on an intercept and X_1. In order to pass the audit, the misaligned agent now builds a prediction function that is indistinguishable from the aligned agent's model based on the best prediction explainer. The prediction explainer partially aligns choices by enforcing $\mathrm{E}_0 f = \mathrm{E}_0 s$. However, the agent can still induce a distortion to reflect his preference which is not captured by the prediction explainer. Expected welfare is reduced due to this remaining bias.

Example: Medical Testing The medical practitioner now builds the following credit scoring function to pass the insurance company's audit: $f(X) = s(X)$. The intercept in the prediction function can no longer be distorted because this distortion would be revealed by the best prediction explainer. Expected welfare is zero and we obtain the first best.

Example: Lending The lender now builds the following credit scoring function to pass the regulator's audit

$$
\begin{aligned}
f(X_1, X_2) &= \alpha - p\Delta_{\text{high utilization}} + \beta\, X_1 + (\gamma + \Delta_{\text{high utilization}})\, X_2 + \delta\, X_1 \cdot X_2 \\
&= \underbrace{s(X_1, X_2)}_{\text{first best}} + \underbrace{\Delta_{\text{high utilization}}\,(X_2 - p)}_{\text{not detectable by } \mathrm{E}_0}.
\end{aligned}
$$

The lender can still induce a distortion to reflect his preference for additional risk for applicants with high credit utilization because the prediction explainer cannot distinguish between models that distort the coefficient on X_2 and models that do not. Expected welfare is reduced due to this remaining bias,

$$
\mathrm{E}[W(f;d)] = -p(1-p)\Delta^2_{\text{high utilization}}.
$$

Example: Hiring The employer audits the hiring agency's prediction function based on the best prediction explainer, which is equivalent to a regression of the predicted values on an intercept and the indicator variable of relevant job experience (X_1). The hiring agency now builds the best-fit prediction function since the best prediction explainer focuses on the model behavior with respect to X_1 but not X_2 (recall that we assume that X_1 and X_2 are uncorrelated in this example). This prediction function leads to an expected welfare loss equal to the employer's weight on distributional equity,

$$
\mathrm{E}[W(f;d)] = -\lambda\left(\mathrm{E}_\mu[s(X)|G=0] - \mathrm{E}_\mu[s(X)|G=1]\right)^2.
$$

This policy leads to maximal distortion, representing an inefficient welfare outcome.

Audit based on targeted explainer. Fifth, we consider the case of no ex-ante restriction and an ex-post audit based on a targeted explainer. Unlike the prediction explainer, the optimal (targeted) explainer is context-specific and depends on the source of the preference misalignment. This audit successfully eliminates the incentive misalignment as long as the explainer is informative about all dimensions of misalignment. This solution can achieve the first-best outcome for the principal.

Example: Medical Testing The optimal (targeted) explainer in this case is $\mathrm{E}^* f = \mathrm{E}_\mu[f(X)]$, corresponding to the intercept in a regression on a constant only. Similar to the case of the best prediction-based audit, the targeted audit successfully eliminates the incentive misalignment and achieves the first best.

Example: Lending The optimal (targeted) explainer corresponds to linear regression of the agent's credit score on a constant and X_2, or equivalently,

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

This targeted audit successfully eliminates the incentive misalignment because the explainer is informative about both dimensions of misalignment. This solution achieves the first-best credit score $f = s$ for the regulator.

Example: Hiring The optimal (targeted) explainer corresponds to the coefficient on X_2 in a linear regression of the agent's prediction function on a constant and X_2, whether the applicant has a high school degree, or equivalently,

$$\mathrm{E}^* f = \Big(\mathrm{E}_\mu[f(X)|X_2 = 1] - \mathrm{E}_\mu[f(X)|X_2 = 0] \Big).$$

This targeted audit successfully eliminates the incentive misalignment because the explainer is informative about the variable that drives the incentive misalignment. Recall that we assume that the agent knows the principal's weight on equality λ. This implies that in order to pass the audit, the agent builds a prediction function that reflects the principal's taste for equal predictions across groups. This solution achieves the first best.

The role of dimensionality. In the above example, the ex-post audit based on the targeted explainer was able to achieve the first best solution. Intuitively, we were able to achieve first best both because the sources of misalignment were known and because the dimensionality of the explainer ($k = 2$) was greater or equal to the dimensionality of the preference misalignment. If the dimensionality of the preference misalignment exceeds that of the explainer, we are no longer able to achieve first best.

Example: Medical Testing The practitioner's misalignment could take on one additional dimension beyond the misalignment over the intercept. In this case, the targeted explainer could still achieve first best. However, if the misalignment were to also apply to specific sub-groups, e.g., those defined by a past heart attack and unfavorable EKG tests, then we are no longer able to achieve the first best with a two-dimensional explainer.

Example: Lending Assume the lender's preference misalignment has an additional dimension, $\Delta(X_1, X_2) = \Delta_{\text{overall}} + \Delta_{\text{high utilization}} X_2 + \Delta_{\text{interaction}} X_2 \cdot X_1$. Now, we are no longer able to achieve the first-best score with a two-dimensional explainer.

Example: Hiring If minority shares also differed by the interaction between college degree and job experience, then a two-dimensional explainer would no longer achieve the first best. A two-dimensional explainer could still achieve a second-best as long as the variation in minority share and the employer's preference for equality are not too high.

3 Documenting the Empirical Relevance of Optimal Regulation

This section demonstrates the empirical relevance of our theoretical results on the optimal regulation of algorithmic predictions in the context of consumer lending. We describe our data, the empirical setup, and our main results. We show that different regulatory tools discussed above produce meaningful differences in the extent to which the lender can generate misaligned credit scoring models. Our demonstration is based on neural networks of varying complexity.

[EMPIRICAL SECTION CURRENTLY UNDER REVIEW BY DATA PROVIDER]

4 A Model of Oversight of Algorithms with Explainers

This section develops our general model of the strategic interaction between a principal and a agent that includes explainers. There is a conflict of interest between the principal and the agent. Only the agent has the technology to calculate complex prediction functions. The principal can impose ex-ante restrictions on the prediction functions the agent can use, and ex-post restrictions based on a simple explanation of the prediction function. We spell out the timing and interaction in this model in Section 4.1 and sketch the general backward-induction solution.

Having set up the general structure of the game, we consider specific quadratic objective functions and linear explanation mapping to provide insight into the roles of ex-ante restrictions and ex-post explanations in regulation. Section 4.2 spells out the specific setup. Section 4.3 argues that ex-ante restrictions and multi-purpose explainers are generally inefficient, while Section 4.4 provides conditions under which a *targeted* explainer permits a first-best solution and develops some general properties of second-best solutions. We discuss extensions of the model in Section 4.5.

4.1 A General Regulation and Explanation Game

We consider a game between a agent and a principal with misaligned preferences. The *agent* of type $\theta \in \Theta$ observes an initial (training) state $s \in \mathcal{S}$ and chooses an allocation or prediction function $f = f(s) \in \mathcal{F}$. The *principal* puts constraints on the prediction function f chosen by the agent. After the prediction function f is chosen, a subsequent (deployment) state $d \in \mathcal{D}$ is realized, leading to agent utility $U_\theta(f; d)$ and principal welfare $W(f; d)$.

We assume that there is a conflict of interest between the agent and the principal; specifically, utility U_θ and welfare W are not generally the same. For example, the agent and the principal may have different risk assessments or preferences, or incorporate different distributional considerations.

To overcome this misalignment, the principal can impose static *ex-ante restrictions* on the prediction functions $\mathcal{F}$ employed by the agent. For example, the principal could restrict the agent to using simple prediction functions. However, these cannot take into account the realization of the state $s \in \mathcal{S}$ or the agent type $\theta \in \Theta$, both of which are ex-ante unknown to the principal.

To this standard game between the principal and the agent, we add the ability to impose additional restrictions through an *ex-post audit.* We assume that once the deployment state $d \in \mathcal{D}$ is realized, the principal receives a context-specific signal $g(f; d)$ about the realized outcome, such as the overall rate at which credit was approved or the overall disparate impact in the deployment of the credit policy chosen by the agent. When observing an adverse signal $g(f; d)$, the principal can however not distinguish between the source of the bad signal; specifically, it could stem from a non-aligned choice f or from a bad deployment state d.

To distinguish between bad action and bad state in deployment, we assume that the principal cannot observe the full, potentially very complex, prediction function $f \in \mathcal{F}$ chosen by the agent, and instead only observes a simple explanation $\textsc{e}f \in \mathcal{E}$. In practice, such a simple explanation may be given by coefficients in a simple model or variable importances of a complex model; below, we will restrict it to a projection onto a lower-dimensional space. Based on this explanation, the principal may then impose sanctions on the agent. In this game, we consider different choices of restrictions to the functions $\mathcal{F}$ as well as explainers $\mathrm{E} : \mathcal{F} \to \mathrm{E}$ based on a distribution $(\theta, s, d) \sim \pi$.

Assumption 1 (Setup and timing of the regulation game)**.** *The timing of this game is as follows:*

1. *Rule-setting stage. The principal observes the (joint) distribution π over states $s \in \mathcal{S}, d \in \mathcal{D}$ and types $\theta \in \Theta$, and chooses an explanation mapping $\mathrm{E} : \mathcal{F} \to \mathcal{E}$ (and possibly restricts the function space $\mathcal{F}$) to maximize expected welfare.*
2. *Training stage. The agent observes the type $\theta \in \Theta$ and training state $s \in \mathcal{S}$, and chooses a function $f \in \mathcal{F}$ to maximize expected utility.*
3. *Deployment stage. The deployment state $d \in \mathcal{D}$ and subsequent signal $g(f; d)$ is realized.*
4. *Audit stage. The principal audits f based on the training state s, the signal $g(f; d)$, and the explanation $\mathrm{E}f \in \mathcal{E}$.*
5. *Payoff stage. agent utility $U_\theta(f; d) - c\mathbb{1}_{\textit{audit fails}}$ and welfare $W(f; d)$ are realized.*[4]

We focus for now on the case of $c = \infty$, so that the agent always avoids the failure of the audit, and the principal can effectively dictate $\mathrm{E}f$. We assume that the principal has commitment power.

The roles of agent type θ and state s are similar in that they are observable to the agent when deciding on a prediction function f but are ex-ante unavailable to the principal. We assume that the principal can observe the state s during the ex-post audit but not the type θ.

The explanation here represents a dimensionality reduction that maps the full prediction function $f \in \mathcal{F}$ to a smaller space $\mathcal{E}$ of interpretable explanations. Below, we will provide a concrete restriction of this mapping to represent the idea of lossy information compression.

4.2 Quadratic Loss Functions and Linear Explainers

Having described a general setup, we now aim to describe the consequences of different ex-ante restrictions and ex-post explainers, and ultimately solve for second-best regulation. In order to obtain tractable examples, we focus on a specific implementation with quadratic loss functions and linear explainers that will, among others, nest the examples presented in Section 2. This specific structure will allow us to characterize first-best and second-best solutions.

Assumption 2 (agent utility and principal welfare)**.** *Prediction functions are n-dimensional real-valued vectors $f \in \mathbb{R}^n$, over which agent and principal have preferences given by quadratic loss functions*

$$U_\theta(f; d) = -(f - u)'\Omega_U(f - u) \qquad (u = u(d, \theta) \in \mathbb{R}^n, \Omega_U = \Omega_U(d, \theta) \in \mathbb{R}^{n\times n}),$$

$$W(f; d) = -(f - w)'\Omega_W(f - w) \qquad (w = w(d) \in \mathbb{R}^n, \Omega_W = \Omega_W(d))$$

[4] We assume that the training state s and the deployment state d are jointly distributed; in particular, these would be correlated in typical applications, so that the training state s conveys information about the realization of d. Hence, while utility and welfare vary with s, we express this relationship only implicitly through the deployment state d.

with bliss points w, u and symmetric positive semidefinite weight matrices Ω_U, Ω_W.

For these utility and welfare functions, the (infeasible) oracle choice for agent and principal are u and w, respectively. Such a structure could be motivated by a model that considers prediction functions across buckets of individuals in the population, possibly with different weights between agent utility and social welfare. Indeed, the structure of these utility functions does nest, in particular, that of the illustration in Section 2. Indeed, the distribution over covariates X can be captured by the weight matrices, while additive terms that are linear in the prediction function f can be incorporated by completing the square.[5] The representation of predicted values in terms of (high-dimensional) vectors is by itself not really restrictive, since each dimension could be a (potential) individual defined by a unique intersection of measured characteristics. The varying bliss points could represent different risk preferences or risk assessments which can vary both with the state and the agent type, or different distributional preferences that can also vary across types.

Utility depends on the realization on the deployment state. Since the agent takes choices and the principal evaluates explanations based on the training state s, it will also be helpful to consider expected utility given s and integrating over the conditional distribution of d. Assumption 2 implies that, up to an additive part that does not depend on the choice of function f and is thus not relevant for optimal choices, we obtain a similar structure for the conditional utility given s only:

Corollary 1 (Expected agent utility and expected principal welfare)**.** *Given state s, expected agent utility and principal welfare can be expressed, up to a constant, as*

$$\overline{U}_\theta(f;s) = -(f-\overline{u})'\overline{\Omega}_U(f-\overline{u}) \qquad (\overline{u}=\overline{u}(s,\theta)\in\mathbb{R}^n, \overline{\Omega}_U=\overline{\Omega}_U(s,\theta)\in\mathbb{R}^{n\times n}),$$

$$\overline{W}(f;s) = -(f-\overline{w})'\overline{\Omega}_W(f-\overline{w}) \qquad (\overline{w}=\overline{w}(s)\in\mathbb{R}^n, \overline{\Omega}_W=\overline{\Omega}_W(s))$$

with bliss points $\overline{w}, \overline{u}$ and symmetric positive semidefinite weight matrices $\overline{\Omega}_U, \overline{\Omega}_W$.

For our results below, it would be enough to impose the weaker restriction on utilities and welfare from this corollary, rather than the stronger ex-post structure in Assumption 2.

We next capture the idea of ex-ante restrictions to smaller function spaces and ex-post explanations in terms of simpler features through linear projections.

Assumption 3 (Ex-ante restrictions and ex-post explanations)**.** *Explanations are formed by linear projections*

$$\mathrm{E}: \mathbb{R}^n \supseteq \mathcal{F} \ni f \mapsto \mathrm{E}f \in \mathcal{E} = \mathbb{R}^k, \qquad \mathrm{E} \in \mathbb{R}^{n\times k}.$$

[5] An additional linear term would be without loss since

$$-(f-x)'\Omega(f-x) + 2\underbrace{\omega'(f-y)}_{=(\Omega^{-1}\omega)'\Omega(f-y)} + z = -(f-x-\Omega^{-1}\omega)'\Omega(f-x-\Omega^{-1}\omega) + \text{const.},$$

where we assume wlog that Ω has full rank (otherwise we can modify Ω appropriately to incorporate ω), and const. does not depend on f and is therefore not directly relevant for choices of the agent.

When we consider linear restrictions to the function space $\mathcal{F}$, we consider restrictions of the type

$$\mathcal{F} = \{f \in \mathbb{R}^n; Af = a\}, \qquad A \in \mathbb{R}^{m \times n}, a \in \mathbb{R}^m.$$

We capture explanations by linear mappings to a smaller space to implement the idea that the explanations yield a lower-dimensional representation of the full predicted value in a tractable way. This definition does not constrain explanations to be expressed in specific ways, such as in terms of primitive features of the data. We could add such constraints to the optimization problem faced by the principal, but for now focus on an unconstrained linear explainer.

While many explanation tools may not be linear, linearity allows us precise expressions about what it means to represent a complex function in terms of a simple projection, while avoiding information-theoretic challenges arising from non-linear maps, such as compressing all dimensions into a single real variable without losing any information. Linear restrictions on functions incorporate, in particular, restrictions that ensure that the function is fully explainable (i.e., in which case A would be of rank $m = n - k$).

4.3 Limits of Ex-Ante Restrictions, Ex-Post Assessments, and General-Purpose Explainers

Before we describe optimal principaly solutions in the game outlined in Assumption 1, we consider three extreme cases that represent standard practical approaches: First, to constrain the predicted value ex-ante to simple functions that can be fully audited, which fully aligns choices at the cost of flexibility; second, to audit only based on the ex-post signal about some target quantity; and third, not to constrain choices, but to regulate based on ex-post explanations that capture as much as possible of the variation in the predicted value via what we will call a "prediction explainer" and make precise below. We argue that neither of those three solutions is generally optimal in our model.

First, the principal could constrain the functions to a k-dimensional space that is thus perfectly explainable. For example, the principal could require that the agent only uses linear regression on a fixed set of covariates to perform their prediction task, and reports the coefficients of this simple linear regression. In this case, the preference misalignment is fully resolved and choices are as if the principal takes them. This solution is not generally optimal:

Remark 1 (No restriction without substantial misalignment)**.** *Restricting the agent to fully explainable function is not generally optimal. Specifically, assume for simplicity that $\overline{\Omega}_U = \overline{\Omega}_W$ and both are constant almost surely. If $\mathrm{Var}_\pi(\overline{w}) \succ \mathrm{E}_\pi[(\overline{u} - \overline{w})(\overline{u} - \overline{w})']$ (where $\succ$ represents the order between matrices implied by positive definiteness) then a restriction to explainable functions is never optimal for any $k < n$, and indeed dominated by not restricting the agent at all.*

The intuition behind this result is straightforward: Restricting the predicted value to explainable functions yields an aligned choice from the restricted functions, but involves excess variance on those components that the principal suppresses ex-ante; if that variance is above the cost $\overline{u} - \overline{w}$ of misalignment in the same components induced by the agent's bliss-point choice $\overline{u}$, then no such restriction can be optimal. When there is some substantial misalignment, some amount of restriction can be optimal, but it will not generally be optimal to restrict predicted values to be fully explainable unless the misalignment is large and universal.

Another extreme solution would be to audit solely based on the realized signal $g(f;d)$, which depends on the choice f of the agent as well as the realized deployment state d. However, regulating based on realized outcomes will generally be inefficient, since the principal cannot distinguish whether an adverse outcome stems from an undesired choice by the agent or an unfortunate realization of the deployment state. If the principal still decided to audit based on realizations of $g(f;d)$, then the agent now has to avoid *any* f that can lead to an impermissible outcome. As long as learning $g(f;d)$ does not rule out any choice f, this constellation is not optimal in our setup:

Remark 2 (Limits of simple ex-post assessments)**.** *Assume that the support of $g(f;d)$ given s does not depend on f or on s. If the agent does not participate if expected utility is $-\infty$ (but participates otherwise) and the principal prefers the agent's unrestricted choice to the agent not participating, then the principal will not use the signal $g(f;d)$ in her audit.*

While we do not otherwise consider participation constraints and assume that agents generally participate unless they face $-\infty$ utility, we provide this result to demonstrate the limits of conditioning on the realized signal, and focus below on audits that do not use this information.

To understand this result, assume that the principal were to approve a agent only for some values in the support of g, holding all information about the training state s fixed. In that case, there is always positive probability of failing the audit, giving the agent $-\infty$ expected probability from participating. This results thus relies heavily on a failed audit having arbitrarily bad consequences to the agent. If we instead allowed for a flexible penalty, then ex-post assessments could improve welfare outcomes. We discuss such an extension in Section 4.5 below.

A final solution would leave functions ex-ante unrestricted, but then impose restrictions on their explainer (instead of the ex-post signal g). A natural explainer to consider would be one that recovers as much information as possible about the function it explains, which we call the "prediction explainer":

Definition 1 (Prediction explainer)**.** *For some symmetric positive definite matrix Ω and function $\hat{f} = \hat{f}(s, \theta)$ (and fixed dimension k), the prediction explainer for $\hat{f}$ is the projection $\text{E} : \mathbb{R}^n \to \mathbb{R}^k, f \mapsto \text{E}F$ that preserves most of the information about $\hat{f}$, i.e.*

$$\text{E}_0 = \arg\min_{\text{E}} \mathrm{E}_\pi[(\hat{f} - \underbrace{\mathrm{E}_\pi[\hat{f}|\text{E}\hat{f}]}_{\textit{best prediction based on explainer}})'\Omega(\hat{f} - \mathrm{E}_\pi[\hat{f}|\text{E}\hat{f}])].$$

For example, when Ω simply represents prediction quality as in the example in Section 2, and $\hat{f}$ a prediction of the true prediction function, then the optimal explainer represents coefficients on those features that are most predictive of the true prediction function. It implements the idea of a general-purpose explainer for the predicted value-prediction exercise. We now apply this explainer to the problem at hand:

Remark 3 (Inefficiency of the prediction explainer)**.** *Assume that the principal audits based on the prediction explainer for the first-best choice $\hat{f} = \overline{w}$, the unconstrained choice $\hat{f} = \overline{u}$, or the equilibrium choice by the agent (and the principal's* *or agent's weight matrix, which we assume to be fixed). Then the average welfare of the resulting prediction function is weakly better than not auditing the agent at all, but generally suboptimal relative to ex-post audits based on an optimal explainer.*

We present an example of suboptimality in Section 2 above. The idea behind suboptimality is the same in the general case: since none of the explainers mentioned here is specific to the nature of preference misalignment between agent and principal, it is not generally optimal.

4.4 Optimal Ex-Post Audits Through Targeted Explainers

Above, we saw that neither rigid restrictions nor general-purpose explanations will be optimal since they fail to target the *difference* in the preferences of agent and principal appropriately. In this section, we explore how a *targeted* explainer can improve outcomes, and even obtain the first-best outcome $f = \overline{w}$. The main idea behind such an optimal explainer is that it represents the misalignment between the agent and the principal, rather than focusing its limited expressiveness inefficiently. Throughout, we will focus on audits that do not explicitly use the ex-post signal $g(f;d)$, motivated by Remark 2 where the principal cannot distinguish definitively between the choice f and the realization d.

Our first main result is that when preference misalignment is limited to at most k dimensions, then we can achieve the first-best outcome in terms of welfare:

Proposition 1 (First-best solution through targeted explanation)**.** *Assume that $\overline{\Omega}_U, \overline{\Omega}_W$ are fixed and that* $\operatorname{rank}(\overline{\Omega}_U^{1/2} - \overline{\Omega}_W^{1/2}) + \operatorname{rank} \mathrm{E}_\pi[(\overline{u} - \overline{w})(\overline{u} - \overline{w})'] \leq k$*. Then there is an explainer that achieves the first-best solution $f = \overline{w}$ of the agent, yielding maximal expected welfare $\overline{W}(f;s) \equiv 0$ across all states and for all agent types. Further, the optimal solution does not include any ex-ante restrictions.*

The rank condition expresses that both the differential weighting and the different prediction targets can be aligned by an explainer that projects the full space of predicted values into a k-dimensional feature space of explanations, provided these explanations are chosen optimally. If the weighting is the same, then the optimal explainer takes a particularly intuitive form:

Proposition 2 (Optimal explainer for different prediction targets)**.** *Assume that $\overline{\Omega}_U \equiv \Omega \equiv \overline{\Omega}_W$ with Ω constant and that* $\operatorname{rank} \mathrm{E}_\pi[(\overline{u} - \overline{w})(\overline{u} - \overline{w})'] \leq k$*. If* $\operatorname{rank} \mathrm{Var}_\pi(\overline{u} - \overline{w}) = \operatorname{rank} \mathrm{E}_\pi[(\overline{u} - \overline{w})(\overline{u} - \overline{w})']$ *then the prediction explainer for $\overline{u} - \overline{w}$ (with weight matrix Ω) achieves the first-best welfare outcome $f = \overline{w}$.*

The optimal explainer in this case focuses solely on those components of the predicted value that agent and principal may disagree on, which are those components that express a difference in targets between the two. If $\operatorname{rank} \mathrm{Var}_\pi(\overline{u} - \overline{w}) < \operatorname{rank} \mathrm{E}_\pi[(\overline{u} - \overline{w})(\overline{u} - \overline{w})']$, we can still achieve a first-best by expanding the prediction explainer to components that represent a constant disagreement. Indeed, we can achieve a first-best by combining ex-ante restrictions and ex-post explanations in the latter case:

Proposition 3 (Optimal combination of restrictions and explanation for different prediction targets)**.** *Assume that* $\overline{\Omega}_U = \Omega = \overline{\Omega}_W$ *with* Ω *fixed and that* $\operatorname{rank}\operatorname{Var}_\pi((\overline{u}-\overline{w})(\overline{u}-\overline{w})') \leq k$*. Then we can achieve a first-best solution by a combination of ex-ante restrictions and ex-post explanation, where the explainer is the prediction explainer for* $\overline{u}-\overline{w}$ *(with weight matrix* Ω*).*

The previous results characterize cases in which we can achieve the first-best welfare outcome. In general the misalignment between agent and principal may imply that even an optimal explainer in combination with an optimal ex-ante restriction cannot achieve first-best welfare. In that case, a second-best solution would only include ex-ante restrictions when the misalignment is sufficiently large in those components that are not captured by the explainer:

Proposition 4 (No ex-ante restriction without substantial uncaptured misalignment)**.** *Assume that* $\Omega_U = \Omega = \Omega_W$ *with* Ω *fixed. If* $\lambda_{(n)}(\operatorname{Var}_\pi(\Omega^{1/2}\overline{w})) \geq \lambda_{(k+1)}(\operatorname{E}_\pi[(\Omega^{1/2}(\overline{u}-\overline{w}))(\Omega^{1/2}(\overline{u}-\overline{w}))'])$*, where* $\lambda_{(i)}$ *represents the i-th highest Eigenvalue, then there is a second-best solution of the principal that does not involve any ex-ante restrictions.*

The intuition for this result is a straight extension of the reasoning behind Remark 1: When an optimal explainer targets the biggest disagreements between agent and principal, then any restriction can only ever be adding welfare if the uncertainty around states – as expressed by $\operatorname{Var}_\pi(\Omega^{1/2}\overline{w})$ – is sufficiently low relative to the *remaining* misalignment after aligning choices through ex-post explanations.

Overall, the conditions in these illustrative results are sufficient, but not necessary. Our model allows for solving for optimal explainers and ex-ante restrictions beyond these edge cases. Note also that none of these results hinge on the distinction between unknown and unverifiable type $\theta \in \Theta$ and unknown but ex-post verifiable state $s \in \mathcal{S}$, since only the latter is directly welfare relevant.

4.5 Extensions

We have focused here on a simple linear model to make precise our discussion of the properties of the role of explainers in optimal regulation. We end on noting that the framework in Section 4.1 extends naturally to considering directly the binary assignment of an outcome (rather than the prediction function), more general notions of explainability, and additional restrictions on what the principal can learn about the state (such as restrictions on learning even about welfare-relevant realizations). Within the linear framework, one natural extension would further limit explainers to linear maps that are forced to use certain interpretable features only, rather than allowing for arbitrary linear combinations of available features.

Finally, we have here argued that the use of the ex-post signal $g(f;d)$ – which may represent the realized disparate impact or the realized default rate – is limited unless it allows for a definite inference about the choice f of the agent. This result is linked to the assumption that the cost to the agent of a failed audit is arbitrarily high. If the principal could instead decide on a cost schedule based on, say, f and $g(f;d)$, then this would add additional options for the principal, such as penalties akin to a Pigouvian tax.

5 Conclusion

As the use of machine learning (ML) and artificial intelligence (AI) becomes more ubiquitous, new questions arise about how the use of these algorithms should be governed. Our paper provides policy-relevant guidance for algorithmic oversight. We argue that new trade-offs emerge as algorithmic decision-making opens up the possibility of inspecting, or auditing, the underlying model. In particular, we argue that entities overseeing the use of algorithms (the ‘principal’) face a trade-off between complexity and oversight. On the one hand, we want those who build prediction tools (the ‘agents’) to be able to exploit the full gains from advances in prediction technology in a wide array of settings, including credit markets, labor markets, the justice system, and medical care. On the other hand, if the entities overseeing the use of algorithms are limited in how much they can learn about the underlying models, oversight becomes more challenging.

We show that algorithmic audits, in particular those based on targeted explainers, can mitigate this trade-off. Our results emerge from a principal-agent setup with an agent who predicts an outcome of interest and a principal who can both specify ex-ante restrictions on the prediction function and conduct ex-post audits. Welfare gains from audits are realized if principals request information not about what drives the *average* prediction of risk but instead about what drives particular types of *mis*-prediction. In most settings, the optimal policy allows agents to fully exploit the complexity of ML/AI technology but inspects the parts of the model that are most likely to drive incentive misalignment.

We investigate the empirical relevance of our theoretical results in the context of unsecured consumer lending. We design an empirical optimization problem that mimics the problem that the misaligned agent solves in our theoretical model. Our empirical work aims to demonstrate that the solutions we provide are both computationally feasible and have the ability to meaningfully improve welfare. The detailed empirical results of this exercise are currently under review by the data provider.

References

ACEMOGLU, D. AND A. FINKELSTEIN (2008): "Input and technology choices in regulated industries: Evidence from the health care sector," *Journal of Political Economy*, 116, 837–880.

ADAMS, W., L. EINAV, AND J. LEVIN (2009): "Liquidity Constraints and Imperfect Information in Subprime Lending," *American Economic Review*, 99, 49–84.

ATHEY, S., G. IMBENS, J. METZGER, AND E. MUNRO (2020a): "Using Wasserstein Generative Adversarial Networks for the Design of Monte Carlo Simulations," *arXiv:1909.02210 [econ, stat]*, arXiv: 1909.02210.

ATHEY, S. C., K. A. BRYAN, AND J. S. GANS (2020b): "The Allocation of Decision Authority to Human and Artificial Intelligence," *AEA Papers and Proceedings*, 110, 80–84.

AVERY, R. B., K. P. BREVOORT, AND G. CANNER (2012): "Does Credit Scoring Produce a Disparate Impact?" *Real Estate Economics*, 40, 65–114, zSCC: 0000038.

AVERY, R. B., K. P. BREVOORT, AND G. B. CANNER (2009): "Credit Scoring and Its Effects on the Availability and Affordability of Credit," *Journal of Consumer Affairs*, 43, 516–537, zSCC: 0000048.

BAKER, L. C. (2001): "Managed care and technology adoption in health care: evidence from magnetic resonance imaging," *Journal of health economics*, 20, 395–421.

BARTLETT, R. P., A. MORSE, N. WALLACE, AND R. STANTON (2019): "Algorithmic Accountability: A Legal and Economic Framework," 58.

BHATT, U., A. XIANG, S. SHARMA, A. WELLER, A. TALY, Y. JIA, J. GHOSH, R. PURI, J. M. F. MOURA, AND P. ECKERSLEY (2020): "Explainable Machine Learning in Deployment," *arXiv:1909.06342 [cs, stat]*.

BLATTNER, L. AND S. NELSON (2021): "How Costly is Noise? Data and Disparities in US Mortgage Market," .

CARVALHO, D. V., E. M. PEREIRA, AND J. S. CARDOSO (2019): "Machine Learning Interpretability: A Survey on Methods and Metrics," *Electronics*, 8, 832.

CHATTERJEE, S., D. CORBAE, K. P. DEMPSEY, AND J.-V. RÍOS-RULL (2020): "A Quantitative Theory of the Credit Score," Tech. Rep. 27671, National Bureau of Economic Research, Inc, publication Title: NBER Working Papers.

CHEN, C., K. LIN, C. RUDIN, Y. SHAPOSHNIK, S. WANG, AND T. WANG (2018): "An Interpretable Model with Globally Consistent Explanations for Credit Risk," *arXiv:1811.12615 [cs, stat]*.

DIAMOND, D. W. AND P. H. DYBVIG (1983): "Bank Runs, Deposit Insurance, and Liquidity," *Journal of Political Economy*, 91, 401–419, publisher: University of Chicago Press.

Doshi-Velez, F. and B. Kim (2017): "Towards A Rigorous Science of Interpretable Machine Learning," *arXiv:1702.08608 [cs, stat]*.

Einav, L., M. Jenkins, and J. Levin (2013): "The impact of credit scoring on consumer lending," *The RAND Journal of Economics*, 44, 249–274, zSCC: 0000146.

Faria-e Castro, M., J. Martinez, and T. Philippon (2017): "Runs versus Lemons: Information Disclosure and Fiscal Capacity," *Review of Economic Studies*, 84, 1683–1707, publisher: Oxford University Press.

Frankel, A. (2014): "Aligned Delegation," *The American economic review*, 104, 66–83.

Fuster, A., M. Plosser, P. Schnabl, and J. Vickery (2019): "The Role of Technology in Mortgage Lending," *The Review of Financial Studies*, 32, 1854–1899, zSCC: 0000178.

Gillis, T. B. (2020): "False Dreams of Algorithmic Fairness: The Case of Credit Pricing," *Available at SSRN 3571266*.

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

Goldstein, I. and Y. Leitner (2017): "Stress Tests and Information Disclosure," SSRN Scholarly Paper ID 3029761, Social Science Research Network, Rochester, NY.

——— (2020): "Stress Tests Disclosure: Theory, Practice, and New Perspectives," SSRN Scholarly Paper ID 3637146, Social Science Research Network, Rochester, NY.

Greenberg, J. and J. B. Green (2014): "Over-testing: why more is not better," *The American journal of medicine*, 127, 362–363.

Greenstone, M., P. Oyer, and A. Vissing-Jorgensen (2006): "Mandated Disclosure, Stock Returns, and the 1964 Securities Acts Amendments," *The Quarterly Journal of Economics*, 121, 399–460, publisher: Oxford University Press.

Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, and F. Giannotti (2018): "A Survey Of Methods For Explaining Black Box Models," *arXiv:1802.01933 [cs]*.

Hashemi, M. and A. Fathi (2020): "PermuteAttack: Counterfactual Explanation of Machine Learning Credit Scorecards," *arXiv:2008.10138 [cs, stat]*.

Hoffman, M., L. B. Kahn, and D. Li (2018): "Discretion in hiring," *The Quarterly Journal of Economics*, 133, 765–800.

Hurst, E., B. J. Keys, A. Seru, and J. Vavra (2016): "Regional Redistribution through the US Mortgage Market," *American Economic Review*, 106, 2982–3028, zSCC: 0000118.

Judge, K. (2020): "Stress Testing During Times of War," *SSRN Electronic Journal*.

Keys, B. J., T. Mukherjee, A. Seru, and V. Vig (2010): "Did Securitization Lead to Lax Screening? Evidence from Subprime Loans," *The Quarterly Journal of Economics*, 125, 307–362, zSCC: 0001831.

Keys, B. J., A. Seru, and V. Vig (2012): "Lender Screening and the Role of Securitization: Evidence from Prime and Subprime Mortgage Markets," *Review of Financial Studies*, 25, 2071–2108, zSCC: 0000256.

Laffont, J.-J. and J. Tirole (1993): *A theory of incentives in procurement and regulation*, MIT press.

Lakkaraju, H. and O. 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*, New York, NY, USA: Association for Computing Machinery, AIES '20, 79–85.

Lakkaraju, H., E. Kamar, R. Caruana, and J. Leskovec (2019): "Faithful and Customizable Explanations of Black Box Models," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA: Association for Computing Machinery, AIES '19, 131–138.

Leuz, C. and R. E. Verrecchia (2000): "The Economic Consequences of Increased Disclosure," *Journal of Accounting Research*, 38, 91–124, publisher: [Accounting Research Center, Booth School of Business, University of Chicago, Wiley].

Li, D., L. R. Raymond, and P. Bergman (2020): "Hiring as exploration," Tech. rep., National Bureau of Economic Research.

Lundberg, S. and S.-I. Lee (2017): "A Unified Approach to Interpreting Model Predictions," *arXiv:1705.07874 [cs, stat]*.

Mullainathan, S. and Z. Obermeyer (2017): "Does Machine Learning Automate Moral Hazard and Error?" *American Economic Review*, 107, 476–80.

——— (2019): *Who is tested for heart attack and who should be: Predicting patient risk and physician error*, National Bureau of Economic Research.

Murdoch, W. J., C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu (2019): "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences*, 116, 22071–22080.

Myers, S. C. and N. S. Majluf (1984): "Corporate financing and investment decisions when firms have information that investors do not have," *Journal of Financial Economics*, 13, 187–221.

Nachman, D. C. and T. H. Noe (1994): "Optimal Design of Securities Under Asymmetric Information," *The Review of Financial Studies*, 7, 1–44, publisher: [Oxford University Press, Society for Financial Studies].

O'Sullivan, J. W., A. Albasri, B. D. Nicholson, R. Perera, J. K. Aronson, N. Roberts, and C. Heneghan (2018): "Overtesting and undertesting in primary care: a systematic review and meta-analysis," *BMJ open*, 8, e018557.

Rambachan, A., J. Kleinberg, S. Mullainathan, and J. Ludwig (2020): “An Economic Approach to Regulating Algorithms,” Tech. Rep. w27111, National Bureau of Economic Research.

Ribeiro, M. T., S. Singh, and C. Guestrin (2016): “”Why Should I Trust You?”: Explaining the Predictions of Any Classifier,” *arXiv:1602.04938 [cs, stat]*.

Slack, D., S. Hilgard, E. Jia, S. Singh, and H. 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*, New York, NY, USA: Association for Computing Machinery, AIES ’20, 180–186.

Traub, A. (2013): “Discredited: How Employment Credit Checks Keep Qualified Workers Out of a Job,” .

Williams, B. (2017): “Stress Tests and Bank Portfolio Choice,” 32.