Manipulation-Robust Prediction*

Daniel Björkegren† Columbia University

Joshua E. Blumenstock‡ U.C. Berkeley

Samsun Knight§ University of Toronto

October 24, 2025

Abstract

An increasing number of decisions are guided by machine learning algorithms. But when consequential decisions are encoded in algorithms, individuals may strategically alter their behavior to achieve desired outcomes. This paper develops an empirical approach that adjusts decision algorithms to anticipate manipulation. By explicitly modeling incentives to manipulate, our approach produces decision rules that are stable under manipulation, even when the rules are fully transparent. We stress test this approach through a large field experiment in Kenya. When implemented, linear strategy-robust decision rules outperform standard linear models such as LASSO.

Keywords: machine learning, manipulation, decision making, digital credit, targeting

*We are grateful for helpful conversations with Susan Athey, Jon Bittner, John Friedman, Greg Lewis, Ted Miguel, Paul Niehaus, Ben Roth, and Jesse Shapiro; and for feedback from seminar audiences at BREAD, Brown, U Chicago, Columbia, ETH Zürich, Harvard, Imperial College London, Microsoft Research, MIT, NBER AI, NBER DEV, NBER IO, NYU, Oxford, Stanford, Tufts, World Bank, UC Berkeley, UCSB, and the Simons Institute. We thank Jolie Wei for research assistance, and Chaning Jang, Simon Muthusi, Nicholas Owsley, and the Busara team for their collaboration. We are grateful for funding from the Brown University Seed Fund, the Bill and Melinda Gates Foundation, and the Center for Effective Global Action. Björkegren thanks the W. Glenn Campbell and Rita Ricardo-Campbell National Fellowship at Stanford University, and Microsoft Research for support. Blumenstock thanks the National Science Foundation for support under CAREER Grant IIS-1942702. This study was pre-registered with the AEA RCT Registry (AEARCTR-0004649) and approved by the IRBs of UC Berkeley, Brown, Columbia, and the Kenya Medical Research Institute.

†dan@bjorkegren.com

‡jblumenstock@berkeley.edu

§samsun.knight@utoronto.ca

1 Introduction

An increasing number of important decisions are now made by machine learning algorithms. Algorithms determine what information we see online; who is hired, fired, and promoted; who gets a loan; and whether to grant bail and parole. In many of these applications, an individual's observed behavior is used as input to a decision rule.

However, when algorithms are used to make consequential decisions, they create incentives for people to 'game' the decision rule. When agents understand how their behavior affects decisions, they may alter their behavior to achieve the outcome they desire. Gamed decision rules can yield decisions that are arbitrarily poor or even unsafe. This problem arises because the standard machine learning approach to training decision rules assumes that the relationship between outcomes and behavior will remain stable. But this assumption often fails once a decision rule is implemented and agents have incentives to change their behavior (Lucas, 1976; Goodhart, 1975).

Economists traditionally address the problem of manipulation by modeling behavioral responses when designing policies, an approach that is central in canonical settings like taxation and mechanism design (Mirrlees, 1971; Akerlof, 1978; Ramsey, 1927; Agarwal and Budish, 2021). However, this insight is not commonly applied when training the modern decision rules that are now common in society, which typically rely on atheoretic estimators to uncover high-dimensional correlations in data.

Instead, real-world applications of machine learning commonly use one of two alternate approaches to deal with manipulation. The first is to restrict the decision rule to only include features that are thought to be more stable — effectively adopting a dogmatic prior that behaviors are either impossible to manipulate (for included features) or arbitrarily manipulable (for excluded features). Yet in reality, people can manipulate most behaviors at some cost, and those costs may be heterogeneous and difficult to assess in modern contexts that can have thousands of features.

The second approach relies on secrecy and retraining. First, decision rules are not revealed in order to make them more difficult for people to game (i.e., 'security through obscurity' (NIST 2008)). However, such secrecy is at odds with the societal demand for a 'right to explanation' about how algorithmic decisions are made (Goodman and Flaxman, 2016). And even when rules are not disclosed, people may still figure out

how to game them, which can cause great harm at unanticipated times in high-stakes settings like finance or governance. Thus, secrecy is often coupled with some degree of re-training to respond to the changing relationship between features and outcomes (Bruckner and Scheffer, 2011). However, each training iteration typically computes a myopic best response; as a result, the process may not converge or may lead to a suboptimal equilibrium. The limits of this approach have become central to policy debates about the regulation of machine learning and artificial intelligence.[1]

This paper considers an alternate approach, which integrates an economic model of behavior into machine learning algorithms. This approach has been explored in computer science (Hardt et al., 2016; Perdomo et al., 2020), but the focus of prior work is on theory. We construct an empirical model that can be estimated from data, and use it to derive an estimator. We then — to our knowledge for the first time — implement, deploy, and evaluate this approach in a real-world setting. Through a field experiment in Kenya, we illustrate how machine learning models trained to anticipate manipulation can lead to better decisions.

The paper is organized into two main parts. The first part explores how this estimator can produce 'strategy-robust' decision rules that are stable under manipulation. We consider a policymaker who seeks a decision rule $\pi(\cdot)$ that produces a decision for individual i based on features $\mathbf{x}_i$. The policymaker obtains loss equal to the square of the difference between the decision $\pi(\mathbf{x}_i)$ and a label y_i, which represents the optimal decision from the perspective of the policymaker. Labels may be observed for a subset of individuals in *training* instances but not in *implementation* instances when the rule is deployed. Whereas the standard approach selects a decision rule that is optimal for the distribution $(\mathbf{x}_i, y_i)$ observed in training, a strategy-robust approach anticipates how individuals will adjust behavior in response to the incentives generated by a decision rule; that is, it models $\mathbf{x}_i(\pi(\cdot))$. Characterizing how behavior will respond to the rule requires a structural model, which we embed within the machine learning estimator. This leads to a 'Stackelberg' solution which obtains better performance by allowing the policymaker to commit to a decision rule. While this general approach

[1]For example, the European Union's General Data Protection Regulation mandates that 'meaningful information about the logic' of automated systems be available to data subjects (European Union, 2016). The White House's Blueprint for an AI Bill of Rights (OSTP, 2022) calls for such explanations as well as for risks to be anticipated and mitigated *before* deployment.

can be applied to arbitrary estimators, we focus primarily on linear decision rules of the form $\pi(\mathbf{x}) = \alpha + \boldsymbol{\beta}\mathbf{x}$ and assume manipulation costs are quadratic in $\mathbf{x}$.

The second major part of the paper shows how the framework can be applied in a real-world setting, through a field experiment in Kenya that we designed specifically to stress-test strategy-robust decision rules. This experiment enables us to train strategy-robust and standard decision rules, and evaluate their performance when implemented. Specifically, we built a smartphone app that passively collects data on how people use their phones, and disburses rewards based on predictions formed from the data collected. The app was designed to mimic 'digital credit' products that have transformed consumer credit in the developing world (Bharadwaj and Suri, 2020; Björkegren et al., 2022). Digital credit products similarly collect user data and use machine learning algorithms to convert that data into a credit score. However, as these systems have scaled, they increasingly face fraud resulting from manipulation, as borrowers learn which behaviors will increase their credit limits (Bloomberg, 2015; Crosman, 2017).

The field experiment produces several results. First, consistent with prior work showing that mobile phone data can predict credit repayment (Björkegren, 2010; Björkegren and Grissen, 2020) and socioeconomic status (Blumenstock et al., 2015; Blumenstock, 2018), we find that the data collected through our smartphone app can predict phone owners' characteristics such as income and intelligence. Second, in a training sample, we structurally estimate the cost parameters that determine each $\mathbf{x}_i(\pi(\cdot))$ in our model; that is, how behaviors would shift if a decision rule $\pi(\cdot)$ were implemented. We construct these estimates using a series of experiments that randomly assign decision rules, offering financial rewards based on behaviors observed through the app. For example, participants faced decision rules that reward them based on frequency of outgoing calls in a given week, or the number of text messages they receive. Average weekly payouts were similar in size to typical digital credit loans in Kenya at the time ($4.80 in Bharadwaj and Suri (2020)). The shifts in behavior that we estimate are intuitive: for instance, outgoing communications are less costly to manipulate than incoming communications, and text messages, which are relatively cheap to send, are more easily manipulated than calls. Complex behaviors (such as the standard deviation of talk time) are harder to manipulate than simpler behaviors

(such as the average duration of talk time). We also find substantial heterogeneity in manipulation ability; much of this heterogeneity arises from unobservables, but people who self-identified as tech-savvy found it easier to manipulate behavior.

Third, we evaluate the trained decision rules $\pi(\cdot)$ in an implementation phase of the experiment where we observe only participants' incentivized behavior. When implemented on real decisions that affect people, strategy-robust decision rules performed substantially better than LASSO decision rules, which do not account for manipulation. We make this comparison by exposing participants to decision rules that offered financial rewards if they used their phones like a particular type of person. For instance, some people received a message stating, 'Earn up to 1000 Ksh if the app guesses that you are a high income earner, based on how you use your phone,' while others received messages that offered rewards for acting like an 'intelligent' person, and so forth. Using a variety of such decision rules, we find that classifications made by the algorithm trained with the strategy-robust approach were more accurate on average than classifications made with the standard approach.

Finally, we estimate the performance cost of algorithmic transparency: the loss from disclosing the details of the decision rule. In the experiment, we experimentally varied how much information subjects had about the decision rule $\pi(\cdot)$. Transparency reduced the performance of standard decision rules by 23% (s.e. 5.9 p.p.). However, when the strategy-robust rule was transparently disclosed, the performance decline was only 9.3% (s.e. 4.2 p.p.). Thus, switching to strategy-robust decision rules reduced the performance cost of transparency by 59% (s.e. 18.7 p.p.).

Taken as a whole, our paper provides a framework for implementing empirical decision rules that are robust to manipulation. We expect similar approaches will be valuable across many domains as human and machine intelligence increasingly interact. While it may seem obvious that predictions that account for incentives will perform better than those that do not, if the assumptions behind strategic response are wrong, real-world performance could be much worse. Our approach combines experiments that measure how behavior responds to perturbations in a decision rule with a structural model to anticipate the response to any rule, which is embedded in an estimator suitable for high dimensional data. Similar approaches are likely to be relevant in a range of applied settings – especially when stakes are high or decision

rules cannot be kept secret, in new implementations where there is limited evidence of historical manipulation, and when updating decision rules is costly or slow.

1.1 Connection to Literature

The conceptual problem of manipulation is not new. Goodhart (1975), in what has since been referred to as 'Goodhart's Law', noted that once a measure becomes a target, it ceases to be a good measure. Lucas (1976) observed that historical patterns can deviate when economic policy changes. Empirically, agents attempt to game decision rules in a wide range of settings, including New York high school exams (Dee et al., 2019), pollution monitoring in China (Greenstone et al., 2019), fish vendors in Chile (Gonzalez-Lira and Mobarak, 2019), and census questions in Indonesia (Banerjee et al., 2018). Economics has a long tradition of developing canonical models for specific settings to anticipate behavioral responses, for example, when setting taxes (Ramsey, 1927; Mirrlees, 1971; Akerlof, 1978), or in market design algorithms (e.g. Agarwal and Budish, 2021).

However, an increasingly important setting falls outside these canonical models. Across society, many consequential decisions are now automated using behavioral 'big data.' These empirical decision rules are often trained atheoretically, based on correlations in high-dimensional data (Breiman, 2001). Many practical implementations assume that the behavior observed in training will remain fixed. Yet manipulation is pervasive; for instance, companies spend many millions of dollars each year manipulating their websites in order to be ranked higher by search engine algorithms (Borrell Associates, 2016). To address manipulation, most systems periodically retrain decision rules, treating it as a generic covariate shift (cf. Sayed-Mouchaweh and Lughofer, 2012). These approaches are typically agnostic about the forces that lead to shifts, and learn by making mistakes. By contrast, we focus on shifts arising from strategic behavior, which can be anticipated.

Theoretical literature in economics and computer science has started to bridge these approaches, suggesting that behavioral responses can, in principle, be incorporated in decision rules trained from data. In economics, work in mechanism design (Frankel and Kartik, 2019, 2020; Ball, 2019; Hennessy and Goodhart, 2023) develops conceptual foundations, and shows that in settings like ours the revelation principle can fail.

In computer science, a theoretical literature on 'strategic classification' considers how behavioral responses can be modeled in classification algorithms. In early work, Hardt et al. (2016) explores how the performance of classifiers could deteriorate in the presence of strategic gaming; they show that, in general, designing a near-optimal classifier is NP-hard (i.e., no known algorithm can solve the problem in polynomial time), and provide a computationally efficient algorithm for learning classifiers with certain types of cost functions. This built on prior work by Bruckner and Scheffer (2011), who show how to compute Stackelberg equilibria of one-shot classification settings by embedding agents' best responses within the loss function. A series of more recent papers analyzes optimal decision rules and extends strategic classification theory to related settings, including iterative environments (Dong et al., 2018), incomplete information (Jagadeesan et al., 2021), heterogeneous agent cost functions (Hu et al., 2019), varying degrees of opacity (Ghalme et al., 2021), and strategic rankings (Liu et al., 2022).[2]

While this literature suggests that behavioral responses can be incorporated in machine decision algorithms in theory, we are unaware of any work that estimates, implements, and evaluates such algorithms under real manipulation. Our paper thus makes two main contributions. First, we develop a tractable empirical model that adjusts how decision rules are trained to anticipate manipulation and from this derive an estimator that produces rules that are effective even when fully transparent and manipulated. Second, to our knowledge for the first time, we design and implement a field experiment that deploys strategy-robust empirical decision rules in a real-world setting.

2 Theoretical Model

This section introduces the model underlying our strategy-robust adjustment. We focus on the stylized case where the decision rule is linear and costs are quadratic,

[2]Also related, Perdomo et al. (2020) introduces a broader notion of 'performative' prediction, formally characterizing settings where repeated risk minimization (retraining) can address more general distribution shifts. Kleinberg et al. (2019) and Miller et al. (2020) consider incentive design and distinguish between strategic behavior that does and does not provide utility. Hu et al. (2019) and Miller et al. (2020) extend these costs to concerns of inequality and fairness.

which we use to derive solutions that we will test with a field experiment.[3]

2.1 Setting

A policymaker seeks a decision rule $\pi(\mathbf{x}_i) = \alpha + \boldsymbol{\beta}'\mathbf{x}_i$ for entities i based on behavior captured in a vector of features $\mathbf{x}_i$. The decision rule could represent, for example, the amount of aid or credit to grant based on a person's observed assets or digital behavior; how much a social network will prioritize a piece of content based on its characteristics and initial engagement; whether to interview an individual based on the text in their resume; and so forth. For convenience, we refer to entities as individuals.

When making a decision on individual i, the policymaker incurs loss equal to the square of the difference between the decision $\pi(\mathbf{x}_i)$ and the label y_i, which represents the optimal decision from the perspective of the policymaker.[4]

Individuals are defined by three-dimensional types $(y_i, \underline{\mathbf{x}}_i, C_i) \sim F$, which have a joint distribution that we take as given. Individuals possess some *natural behavior* $\underline{\mathbf{x}}_i$, which represents their ideal behavior in the absence of incentives. However, i chooses their *observed behavior* $\mathbf{x}_i$, which can differ from this natural level, by incurring a *manipulation cost* $c_i(\mathbf{x}_i, \underline{\mathbf{x}}_i)$, which is parameterized by C_i.

Individuals obtain utility from the policy's decision, minus any cost from manipulation[5]

$$u_i(\pi(\cdot), \mathbf{x}_i) = \pi(\mathbf{x}_i) - c_i(\mathbf{x}_i, \underline{\mathbf{x}}_i).$$

When manipulation costs are quadratic,

$$c_i(\mathbf{x}_i, \underline{\mathbf{x}}_i) = \frac{1}{2}(\mathbf{x}_i - \underline{\mathbf{x}}_i)' C_i (\mathbf{x}_i - \underline{\mathbf{x}}_i)$$

[3] The setting with linear rules and quadratic costs falls within the class of 'separable' cost functions explored in Hardt et al. (2016).

[4] That is, the total realized loss is $\sum_i (y_i - \pi(\mathbf{x}_i))^2$. In the field experiment, we report the square root of the mean loss as it has more natural units and can be compared across samples of different sizes.

[5] We focus on the benchmark case where the utility from the decision coincides with the implemented decision, which is a first approximation of many settings. See Section 5.2 for a more general formulation.

for a matrix of costs

$$C_i = \begin{bmatrix} c_{11i} & \cdots & c_{1Ki} \\ \vdots & \ddots & \vdots \\ c_{K1i} & \cdots & c_{KKi} \end{bmatrix}$$

that is symmetric and positive definite. This parameterization allows for heterogeneity by behavior (indices jk) and person (index i). Some behaviors may be harder to manipulate than others, either by themselves (the diagonal) or in conjunction with other behaviors (the off-diagonals). Different types of individuals may also find it more costly to manipulate behaviors; for example, clever people or those with low opportunity costs may face lower costs.

Behavioral response i chooses optimal behavior $\mathbf{x}_i^*$ to maximize utility:

$$\mathbf{x}_i^*(\alpha, \boldsymbol{\beta}) = \arg\max_{\mathbf{x}_i} \left[\alpha + \boldsymbol{\beta}'\mathbf{x}_i - c_i(\mathbf{x}_i, \underline{\mathbf{x}}_i)\right] = \underline{\mathbf{x}}_i + C_i^{-1}\boldsymbol{\beta}. \tag{1}$$

Since the optimal behavior does not depend on α, we write $\mathbf{x}_i^*(\boldsymbol{\beta})$, omitting α for convenience. When the decision does not depend on behavior ($\boldsymbol{\beta} = \mathbf{0}$), i's optimal behavior is their natural level ($\mathbf{x}_i^*(\mathbf{0}) = \underline{\mathbf{x}}_i$). However, as $\boldsymbol{\beta}$ moves away from zero, it creates incentives for behavior to follow: i's behavior moves in the same direction, down-weighted by their cost of manipulation (highlighted in blue).

2.2 Decision rules

The policymaker believes the joint distribution of labels, natural behaviors, and manipulation costs is $(y_i, \underline{\mathbf{x}}_i, C_i) \sim \tilde{F}$.

A **strategy-robust decision rule** is then given by

$$\alpha^{SR}, \boldsymbol{\beta}^{SR} = \arg\min_{\alpha,\boldsymbol{\beta}} \mathbb{E}_{\tilde{F}}\left[\left(y_i - \alpha - \boldsymbol{\beta}'(\underline{\mathbf{x}}_i + C_i^{-1}\boldsymbol{\beta})\right)^2\right] \tag{2}$$

which anticipates how each individual will manipulate behavior.

In contrast, a best linear predictor (BLP, as estimated by OLS) best responds to observed behavior, anticipating that behavior will remain the same. It results in a decision rule

$$\alpha^{BLP}, \boldsymbol{\beta}^{BLP} = \arg\min_{\alpha,\boldsymbol{\beta}} \mathbb{E}_{\tilde{G}} \left[\left(y_i - \alpha - \boldsymbol{\beta}' \mathbf{x}_i \right)^2 \right], \tag{3}$$

given the joint distribution $(y_i, \mathbf{x}_i) \sim \tilde{G}$.

2.3 Intuition and Discussion

We provide intuition for how the method works using Monte Carlo simulations. These simulations involve a policymaker who implements a linear decision rule ($\pi(\mathbf{x}_i) = \beta_1 x_{1i} + \beta_2 x_{2i} + \alpha$) based on two observed behaviors, x_1 and x_2. The policymaker trains the decision rule on unincentivized behavior, $\underline{\mathbf{x}}$. x_1 is initially more predictive of the preferred decision than x_2, but it is also more susceptible to manipulation.

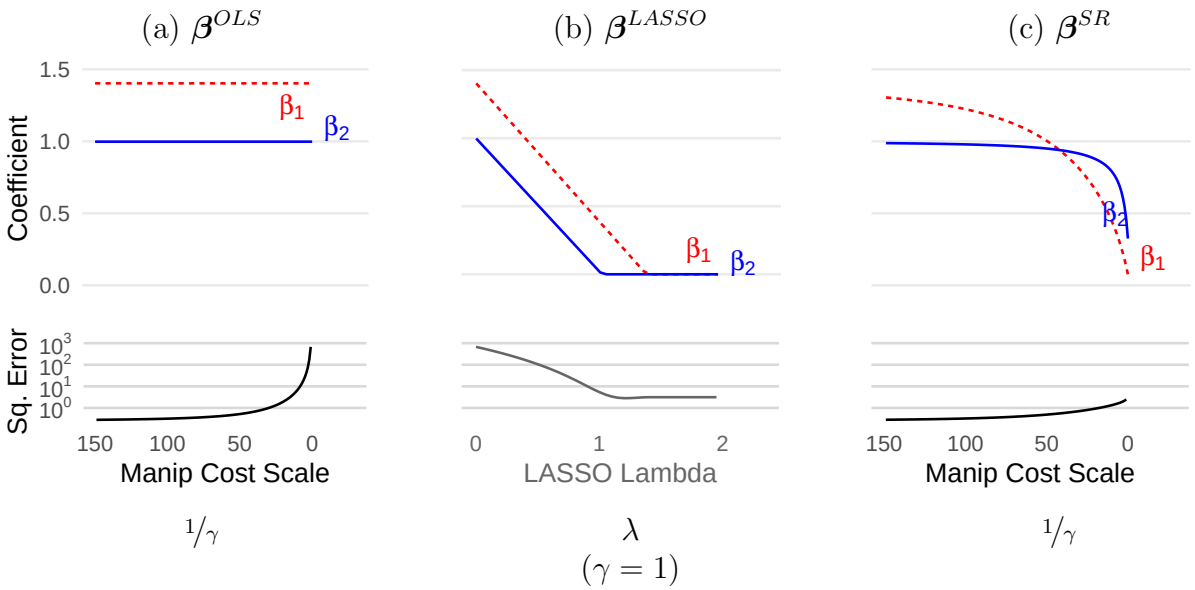
We consider a diagonal cost matrix with elements $c_{kki} = \frac{c_{kk}}{\gamma_i^{het}\gamma}$, with $c_{11} \ll c_{22}$. We allow for heterogeneity between people, defined by γ_i^{het}, which is independent of y_i and $\underline{\mathbf{x}}_i$, and scale the overall ease of manipulation with common factor γ. The policymaker knows the distribution of costs. Figure 1 compares the performance of three different approaches to designing decision rules in this setting.

Panel (a) of Figure 1 illustrates the parameters selected by OLS. These parameters do not depend on manipulation costs—either the relative levels (c_{11i} vs. c_{22i}), or the absolute scale ($1/\gamma$), which decreases from left to right (as indicated in the bottom panel). Instead, OLS maximizes predicted performance in the unincentivized sample where for each i, $(\mathbf{x}_i^*(\mathbf{0}), y_i) = (\underline{\mathbf{x}}_i, y_i)$.

An alternate approach would be to add a LASSO penalty $R_\lambda^{LASSO}(\boldsymbol{\beta}) = \lambda \sum_k |\beta_k|$ to the objective when estimating the standard decision rule in equation (3). Panel (b) shows the parameters selected by LASSO, fixing the scale of manipulation costs ($\gamma = 1$) and varying the LASSO regularization penalty (λ) along the x-axis. Like OLS, LASSO places the most weight on x_1, since it has the strongest relationship to y in the unincentivized sample. As λ increases, both parameters are penalized similarly. Since LASSO assumes that behavior will remain fixed at $\underline{\mathbf{x}}$, it drops x_2 first. Thus, regularization does not address the key issue that x_1 is more easily manipulated.

The strategy-robust approach, shown in panel (c), is related to regularization in that it systematically alters how features are expressed in a decision rule. However, unlike LASSO, the strategy-robust approach adjusts the decision rule to account for

Figure 1: Common vs. Strategy-Robust Decision Rules



Notes: The policymaker's desired allocation is $y_i = \mathbf{b}'\underline{\mathbf{x}}_i + e_i$. The first behavior is more predictive ($b_1 > b_2 > 0$), but is easily manipulable ($c_{11i} \ll c_{22i}$) and has more manipulation noise. **(a)** OLS performance deteriorates when behavior can be manipulated. (**b**) LASSO penalization with hyperparameter λ favors x_1, which will be manipulated as soon as the decision rule is implemented. **(c)** Our method anticipates that x_1 will be manipulated, and shifts weight to x_2 as behavior becomes manipulable. Simulation parameters:
$\underline{\mathbf{x}}_i \overset{iid}{\sim} N\left(\mathbf{0}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$, $\mathbf{b} = \begin{bmatrix} 1.4 \\ 1 \end{bmatrix}$, $C_i = C\frac{1}{\gamma_i^{het}\gamma}$, $C = \begin{bmatrix} 4 & 0 \\ 0 & 32 \end{bmatrix}$, $\gamma_i^{het} \overset{iid}{\sim} Pareto(1, 0.1)$, $e_i \overset{iid}{\sim} N(0, 0.25)$.
The policymaker knows C and γ but only the distribution of γ_i^{het}, and takes B=10 draws of the distribution when estimating the decision rule. N=10,000. Squared error computed on out of sample draw from same population, incentivized by that decision rule.

how each person's features $\mathbf{x}_i^*(\boldsymbol{\beta})$ will be manipulated in the counterfactual state of the world where the decision rule is implemented. We show how the solution varies with the scale of manipulation costs ($1/\gamma$). When manipulation costs are high, the strategy-robust solution approaches OLS; as manipulation becomes easier, the solution adjusts, in this case by penalizing x_1.[6]

Understanding the strategy-robust adjustment

Unlike revelation mechanisms where a person's type can be inferred from their behavior, in our setting, individuals may be heterogeneous both in their ability to shift behavior and other dimensions of type (like the theoretical settings of Frankel and Kartik (2019,

[6]The method can also exploit cost interactions, adjusting behaviors that make it easier to shift other behaviors (akin to Ramsey (1927) taxation). See Online Appendix S3.2.

2020) and Ball (2019)). As a result, the people with more desired behaviors may include those with more desired types and those with higher ability to game, regardless of type. Because the mapping may not be one-to-one, types may not be fully revealed.

The strategy-robust solution for $\boldsymbol{\beta}$ is given by the moment condition derived from equation (2):

$$\mathbb{E}_{\tilde{F}}\left[\mathbf{x}_i^*\left(\boldsymbol{\beta}\right)\cdot\varepsilon_i\left(\alpha,\boldsymbol{\beta},\mathbf{x}_i^*\left(\boldsymbol{\beta}\right)\right)\right]=-\mathbb{E}_{\tilde{F}}\left[C_i^{-1}\boldsymbol{\beta}\cdot\varepsilon_i\left(\alpha,\boldsymbol{\beta},\mathbf{x}_i^*\left(\boldsymbol{\beta}\right)\right)\right] \tag{4}$$

given residual[7]

$$\varepsilon_i(\alpha,\boldsymbol{\beta},\mathbf{x})=y_i-\alpha-\boldsymbol{\beta}'\mathbf{x}.$$

The strategy-robust estimator differs from standard estimators in two ways. First, it anticipates that the levels of behaviors will shift in response to the choice of $\boldsymbol{\beta}$. The left side of equation (4) is analogous to the moment condition of OLS, but replaces observed behaviors $\mathbf{x}_i$ with counterfactual behaviors $\mathbf{x}_i^*(\boldsymbol{\beta})$. These behavioral responses may be heterogeneous between individuals.[8] If the people who find it easier to manipulate (low C_i) have differential values of the outcome (y_i), manipulation itself serves as a signal of type (as in Spence (1973) and Nichols and Zeckhauser (1982)). Our method tends to increase the weight on manipulable behaviors if y_i is negatively correlated with C_i (and decrease if positively correlated), provided this correlation is present in $\tilde{F}$. However, there tends to be additional, idiosyncratic heterogeneity in manipulation ability beyond that observable to the policymaker during training. Accounting for the resulting manipulation noise leads the method to attenuate $\boldsymbol{\beta}$ (Frankel and Kartik, 2019).

Second, the method anticipates the *gradient* of those behaviors: how $\mathbf{x}_i$ would respond if $\boldsymbol{\beta}$ were to deviate off path. The right-hand side of equation (4) departs from orthogonality and produces a Stackelberg (subgame-perfect) equilibrium. In contrast, standard estimators compute a one-step best response, assuming that $\mathbf{x}_i$ would remain fixed if $\boldsymbol{\beta}$ were to deviate.[9]

[7] As well as moment condition $\mathbb{E}_{\tilde{F}}\left[\varepsilon_i\left(\alpha,\boldsymbol{\beta},\mathbf{x}_i^*\left(\boldsymbol{\beta}\right)\right)\right]=0$, which pins down α.

[8] If behavior cannot be manipulated ($C_i\to\boldsymbol{\infty}$), our solution corresponds to OLS. If each behavior j has the same manipulation cost for all people ($c_{jki}\equiv c_{jk}$), the method will expect each person to shift behavior the same amount in response to a given decision rule $\boldsymbol{\beta}$.

[9] Thus, even if one trained a standard decision rule on data from a strategy-robust equilibrium $(y_i,\mathbf{x}_i^*(\boldsymbol{\beta}^{SR}))$, it may learn a different decision rule that escapes the equilibrium.

Performance

Further simulations illustrate how the strategy-robust approach can produce better decisions when people manipulate behavior. Table 1 simulates a scenario with three observed behaviors, with a decision rule of the form $\pi(\mathbf{x}_i) = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \alpha$. Here, x_1 is initially more strongly related to the preferred decision than the other two features, but is more manipulable on average (the cost matrix is diagonal, with $c_{11i} = {}^1/_{\gamma_i}$, $c_{22i} = {}^2/_{\gamma_i}$ and $c_{33i} = {}^4/_{\gamma_i}$). Manipulation costs are 10 times higher for a random half of individuals ($\gamma_i = 10$ or $\gamma_i = 1$, each with probability 0.5); the policymaker knows this distribution but not who finds it easy to manipulate. These parameters illustrate a case where manipulation can have a large impact on performance.

Panel B of Table 1 illustrates the performance of two status-quo approaches to constructing decision rules. In the first row, OLS captures the static relationship between features and the outcome. This approach would perform well if behavior were fixed (as indicated by the low squared loss in training data); however, once people respond to the decision rule, the OLS rule leads to poor decisions (the loss 'when implemented'). The next set of rows in Panel B illustrate a common status quo approach, in which the OLS decision rule is periodically retrained. After observing behavior in the first period (when the rule $\boldsymbol{\beta}^{OLS}$ is active), the rule is then re-trained to obtain $\boldsymbol{\beta}^{OLS(2)}$, which places lower weight on the manipulated x_1. However, once people respond to this new rule, it also performs poorly. As this process is iterated, the rule always appears to predict well on the training sample but makes poor decisions when actually implemented. In this particular case, the process does not converge; it alternates between decision rules that place high and low weight on x_1.[10] Thus, standard approaches can perform poorly even in stable settings with perfect information (Perdomo et al. (2020) provides more formal discussion of this point). In settings with noise or friction in learning, a system might unexpectedly and catastrophically fail when the other side discovers how to exploit it.

In contrast, the strategy-robust decision rule ($\boldsymbol{\beta}^{SR}$ in Panel C) adjusts the coefficients by penalizing the behavior x_1 that has more manipulation noise, shifting

[10] In some cases, this process can converge; or, oscillations can be dampened using cumulative data from prior periods. However, this may take several iterations and the resulting equilibrium may be inferior to the strategy-robust equilibrium (Online Appendix S3.3.1 and Section 5.1).

Table 1: Manipulation Can Harm Prediction (Monte Carlo)

	Decision Rule				**Performance** (squared loss)	
	β_1	β_2	β_3	α	On training data	When implemented
Panel A: **Data Generating Process (Unmanipulated)**						
$\boldsymbol{b}^{DGP}$	3.00	0.10	0.10	0.20	0.25	4147.95
Panel B: **Status Quo Approaches**						
$\boldsymbol{\beta}^{OLS}$	3.01	0.10	0.10	0.20	0.25	4198.12
Iterative retraining						
$\boldsymbol{\beta}^{OLS(2)}$	−0.02	2.84	−2.34	0.16	1.90	1491.03
$\boldsymbol{\beta}^{OLS(3)}$	3.01	0.09	0.10	0.20	0.25	4214.57
$\boldsymbol{\beta}^{OLS(4)}$	−0.02	2.84	−2.34	0.16	1.89	1487.82
⋮						
$\boldsymbol{\beta}^{OLS(1001)}$	3.54	−0.52	0.70	0.18	0.32	8401.06
$\boldsymbol{\beta}^{OLS(1002)}$	0.30	2.60	−2.16	0.18	1.57	1101.25
Panel C: **Strategy-Robust Method**						
$\boldsymbol{\beta}^{SR}$	0.29	0.50	−0.01	−0.94	7.27	6.86

Notes: Monte Carlo simulation results. Panel A shows the coefficients that relate the outcome to natural behaviors under the data generating process (DGP), $y_i = \mathbf{b}'\underline{\mathbf{x}}_i + e_i$. Panel B shows coefficients from OLS. For the retraining approach, the training data for $\boldsymbol{\beta}^{OLS(n)}$ is the manipulated data from when $\boldsymbol{\beta}^{OLS(n-1)}$ is assigned; $\boldsymbol{\beta}^{OLS(1)} = \boldsymbol{\beta}^{OLS}$. Panel C shows coefficients estimated with the strategy-robust method. Performance is assessed on the same sample of individuals under the training data, and when the data is manipulated.

Parameters: N=10000, actual costs are $C_i = \frac{1}{\gamma_i}C$, $\gamma_i = 10^{Bernoulli(0.5)}$. Policymaker knows C and the distribution of γ_i but not the value for each i. Policymaker averages over B=10 draws of the distribution, see equation (6). $\underline{\mathrm{x}} \overset{iid}{\sim} N\left(\mathbf{0}, \begin{bmatrix} 1.0 & 1.0 & 0.1 \\ 1.0 & 2.0 & 1.0 \\ 0.1 & 1.0 & 1.0 \end{bmatrix}\right)$, $C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{bmatrix}$, $e_i \overset{iid}{\sim} N(0, 0.25)$.

weight to behaviors that are harder to manipulate (x_2 and x_3). It anticipates how manipulation would change under other values of $\boldsymbol{\beta}$, sacrificing performance in the environment in which it is trained (in-sample, no manipulation) for performance in the counterfactual implementation environment where there will be manipulation. When individuals manipulate as modeled, the strategy-robust decision rule exceeds the performance of the standard estimator.

3 Estimation

The previous section took as given that the policymaker had formed a belief of the joint distribution $(y_i, \underline{\mathbf{x}}_i, C_i) \sim \tilde{F}$. This section considers how $\tilde{F}$ can be estimated to empirically learn strategy-robust decision rules.

We estimate $\tilde{F}$ in two steps in a *training sample* where label y_i is known for each individual. First, we implement and communicate different decision rules $\boldsymbol{\beta}$ and observe resulting behavior $\mathbf{x}_i^*(\boldsymbol{\beta})$; this makes it possible to estimate $\underline{\mathbf{x}}_i$ and C_i, under restrictions on their dependence. Second, we use these estimates to learn the strategy-robust decision rule $\boldsymbol{\beta}^{SR}$, which best predicts y_i anticipating $\mathbf{x}_i^*(\boldsymbol{\beta}^{SR})$. This decision rule can then be deployed to a full population (an *implementation sample*), including individuals without observed labels.

We extend our model of behavior to allow for panel data. When facing a decision rule $\boldsymbol{\beta}$ in time period t, i's empirical behavior is given by

$$\mathbf{x}_{it}^*(\boldsymbol{\beta}) = \underline{\mathbf{x}}_i + C_i^{-1}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{it},$$

which may deviate from i's natural level ($\underline{\mathbf{x}}_i$) due to manipulation, or an idiosyncratic shock that rationalizes variation across time ($\boldsymbol{\epsilon}_{it}$, with $\mathbb{E}[\boldsymbol{\epsilon}_{it}] \equiv \mathbf{0}$ and $\mathbb{E}[\epsilon_{itk}\underline{\mathbf{x}}_i] \equiv \mathbf{0}$).[11]

3.1 Behavioral responses

The most direct approach to estimating the distribution of behavioral responses ($\underline{\mathbf{x}}_i$ and C_i) relies on assigning each individual i to random decision rules $\boldsymbol{\beta}$ in a training sample – Section 5.3 discusses non-experimental approaches to estimate behavioral

[11] This arises from the utility function $u_{it}(\alpha, \boldsymbol{\beta}, \mathbf{x}_{it}) = \alpha + \boldsymbol{\beta}'\mathbf{x}_{it} - c_i(\mathbf{x}_{it}, \underline{\mathbf{x}}_i) + \boldsymbol{\epsilon}_{it}' C_i(\mathbf{x}_{it} - \underline{\mathbf{x}}_i)$.

responses. Randomization ensures that $\mathbb{E}[\epsilon_{itk}\boldsymbol{\beta}] \equiv \mathbf{0}$, and $\mathbb{E}[\underline{\mathbf{x}}_i\beta_j] = \mathbf{0}$ for each j. After a decision rule is communicated to i, we can estimate parameters from observed responses. In the linear setting, all parameters can be estimated by assigning decision rules that are simple perturbations of a base decision rule $\boldsymbol{\beta}_0$. In particular, for a decision rule $\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \beta_\delta \boldsymbol{\delta}_k$ that has been perturbed by amount β_δ along dimension k (where $\boldsymbol{\delta}_k$ represents the k^{th} unit vector), the expected vector of manipulated behavior is

$$\mathbb{E}\left[\mathbf{x}^*_{it}(\boldsymbol{\beta}_0 + \beta_\delta \boldsymbol{\delta}_k)\right] = \mathbb{E}\left[\mathbf{x}^*_{it}(\boldsymbol{\beta}_0)\right] + \beta_\delta C_i^{-1}\boldsymbol{\delta}_k.$$

with expectations conditional on $\underline{\mathbf{x}}_i$ and C_i. We can thus learn how sensitive behaviors are to incentives by observing how the average vector of behaviors changes as we vary β_δ for different k.

Heterogeneity and noise Our cost specification distinguishes between observed heterogeneity and manipulation noise. We recover observed heterogeneity under the assumption that C_i is independent of $\boldsymbol{\epsilon}_{it}$ conditional on a set of variables observed during training, $\mathbf{z_i}$. We use residual variation to form the policymaker's uncertainty about costs (which results in manipulation noise); our decision rule estimator integrates over this distribution.

To improve small sample performance, we parameterize manipulation costs to allow separable heterogeneity by behavior (jk) and person (i),

$$c_{jki} = c_{jk} \cdot \frac{1}{\gamma_i},$$

where individual gaming ability is modeled as $\gamma_i = f(\boldsymbol{\omega}, \mathbf{z}_i) + v_i$. This has two components. Observed heterogeneity is described by the function f, which is parameterized by $\boldsymbol{\omega}$.[12] Manipulation noise is captured by $v_i \sim V$, which is assumed to be mean zero and independent of $\mathbf{z}_i$. This unobservable portion will be treated as i.i.d. random effects when estimating the decision rule. Altogether, our procedure accounts for the signaling value of manipulation only across people differing in $\mathbf{z}_i$.

Separability implies that the noise in element c_{jki} is proportional to the base costs

[12] Although $\mathbf{z}_i$ is observed for the training sample, the decision rule cannot condition on it: it is not observed when the decision rule is implemented.

of manipulation c_{jk}. As a result, behaviors that are more manipulable on average will be subject to more manipulation noise, and our method will tend to attenuate their coefficients in decision rules.

3.1.1 Estimation procedure

We develop three estimation procedures for different settings. Our main focus is on the setting created in our field experiment, where we perturb decision rules relative to a control base rule $\boldsymbol{\beta}_0 = \mathbf{0}$ (this mimics 'greenfield' settings that arise before a decision rule has been implemented and where behavior is initially unmanipulated), and where each individual i is observed over multiple perturbed decision rules $\boldsymbol{\beta}_{it}$. The appendix presents two alternate procedures. Appendix A1.2 extends this approach to settings where a decision rule is already in use (a 'brownfield' environment); Appendix A1.3 extends to 'one shot' settings where each individual's behavior is observed only once.

Natural behaviors ($\underline{\mathbf{x}}_i$) In a greenfield setting, we first estimate $\underline{\mathbf{x}}_i$ from behaviors observed in unincentivized periods. This estimate, $\hat{\underline{\mathbf{x}}}_i$, can be calculated as the average of i's observed behavior during periods with no incentives. Alternatively, finite-sample performance may improve by modeling shocks as having both common and individual-specific terms, $\boldsymbol{\epsilon}_{it} = \boldsymbol{\mu}_t + \boldsymbol{\eta}_{it}$, and then estimating the OLS regression,

$$\mathbf{x}_{it} = \underline{\mathbf{x}}_i + \boldsymbol{\mu}_t + \boldsymbol{\eta}_{it}, \tag{5}$$

with time fixed effects, including only unincentivized periods where $\boldsymbol{\beta}_{it} = \mathbf{0}$.

Manipulation costs (C and $\boldsymbol{\omega}$) In the second step, we impose each $\hat{\underline{\mathbf{x}}}_i$ (and $\hat{\boldsymbol{\mu}}_t$ if estimated, otherwise $\boldsymbol{\mu}_t = 0$), and jointly estimate C and $\boldsymbol{\omega}$ using Generalized Method of Moments (GMM) with moment conditions detailed in Appendix A1.1. Moment conditions include that implemented decision rules are orthogonal to idiosyncratic behavior shocks and manipulation noise ($\mathbb{E}[\beta_{itk}\epsilon_{itj}] = 0$ or $\mathbb{E}[\beta_{itk}\eta_{itj}] = 0$, and $\mathbb{E}[\beta_{itk}v_i] = 0$, for each incentive β_{itk} that i faced in period t on behavior k). Additionally, manipulation noise v_i is mean zero ($\mathbb{E}[v_i] = 0$), and orthogonal to each heterogeneity characteristic z_l ($\mathbb{E}[z_{li} \cdot v_i] = 0$ for each l). To reduce estimation variance (at the expense of bias), we include a regularization term in the GMM loss function to penalize the ease of ma-

nipulation towards zero (costs towards infinity; which penalizes the resulting decision rules towards standard estimators). We use LASSO-style penalization with separate hyperparameters for diagonal and off-diagonal costs ($\mathbf{\Lambda} = \{\Lambda_{diagonal}, \Lambda_{offdiagonal}\}$).[13] In our application, we will regularize off-diagonal elements to zero since they are imprecisely estimated, and set $\Lambda_{diagonal}$ with cross-validation.

Manipulation noise (V) An individual's behavior is affected by both idiosyncratic shocks (ϵ_{itk}) and manipulation noise (v_i). After estimating the cost parameters, we estimate the noise distribution V in two steps. First, we compute $\hat{v}_i$ based on whether each individual manipulates more or less than predicted during randomized incentivized periods, using a panel average of behavior by individual (Appendix equation (9)). Second, to reduce finite-sample sensitivity to idiosyncratic shocks, we shrink and winsorize these averages, forming the empirical distribution $\tilde{V} = \{\max(\phi \cdot \hat{v}_i, \underline{v})\}_i$, where $\underline{v}$ is the lowest value of $\hat{v}_i$ that would lead to a nonnegative implied gaming ability, and ϕ is a shrinkage parameter.[14] One can impose ϕ or calibrate it based on fit in initial incentivized periods.

3.2 Decision rules

A decision rule can then be estimated to best predict $\mathbf{y}$ given the anticipated behaviors $\mathbf{x}_i^*(\cdot)$, using the estimates ($\underline{\hat{\mathbf{x}}}$, $\hat{C}$, $\hat{\boldsymbol{\omega}}$, $\tilde{V}$) and the empirical equivalent of equation (2),

$$\hat{\alpha}^{SR}, \hat{\boldsymbol{\beta}}^{SR} = \arg\min_{\alpha,\boldsymbol{\beta}} \left[\frac{1}{N} \sum_i \left[\frac{1}{B} \sum_b \left[y_i - \alpha - \boldsymbol{\beta}'(\underline{\hat{\mathbf{x}}}_i + \hat{C}_{ib}^{-1}\boldsymbol{\beta}) \right]^2 + R_\lambda(\boldsymbol{\beta}) \right] \right]. \quad (6)$$

This procedure uses B draws of the cost matrix, given by $\hat{C}_{ib} = \frac{1}{f(\hat{\boldsymbol{\omega}}, \mathbf{z}_i) + \tilde{v}_{ib}} \hat{C}$, where $\tilde{v}_{ib} \sim \tilde{V}$ are drawn randomly from the estimated distribution.[15] Finally, one may include a regularization term, such as $R_\lambda^{LASSO}(\boldsymbol{\beta}) = \lambda \sum_k |\beta_k|$, which can improve small-sample performance in the presence of statistical noise.

[13] We use the term $R_{costs}^{\mathbf{\Lambda}}(\cdot) = \left[\Lambda_{diagonal} \sum_k (\theta_{kk})^2 + \Lambda_{offdiagonal} \sum_{j \neq k} (\theta_{jk})^2 \right] \left[\frac{1}{N} \sum_i f(\boldsymbol{\omega}, \mathbf{z}_i)^2 \right]$, with θ_{jk} representing the elements of inverse costs C^{-1}.

[14] That is, $\underline{v} = \min_i(\hat{v}_i | \phi \cdot \hat{v}_i \geq -\min_{i'}(f(\hat{\boldsymbol{\omega}}, \mathbf{z}_{i'})))$.

[15] Note that this step imposes that v_i is orthogonal to $\underline{\mathbf{x}}_i$, though that is not imposed when estimating the objects. We use different sets of draws of $\boldsymbol{v}$ when training, and when reporting decision rule performance.

4 Field Experiment in Kenya

We designed a field experiment to test the performance of strategy-robust decision rules in a real-world setting. Working with the Busara Center in Nairobi, we developed and deployed a new smartphone application (the 'Smart Sensing' app) to 1,557 research subjects.

The app was designed to mimic key features of digital credit applications, which have become widely popular in recent years and are transforming how consumers in developing countries access credit (Bharadwaj and Suri, 2020; Francis et al., 2017). In a typical digital credit application, lending decisions are based on an 'alternative credit score' that is constructed by applying machine learning algorithms to data on how the loan applicant uses their phone (Björkegren, 2010; Björkegren and Grissen, 2020). At the time of our field experiment, CGAP (2018) estimated that 27% of Kenyan adults had an outstanding 'digital credit' loan. Yet, there is mounting evidence that digital credit is a domain where manipulation is problematic. In one example, Bloomberg covered a story where 'a scam artist studied the loan-approval patterns for several months, using 30 different sim cards to generate data sets and deciphering the lender's algorithms. He fleeced the firm of $30,000 in one day and then vanished' (Bloomberg Technology, Sep. 22, 2015). The potential for manipulation is also salient to everyday customers: in a survey conducted in Kenya and Tanzania, respondents listed the desire to obtain larger digital loans as one of the top five reasons for saving money in their mobile money accounts (FSD Kenya, 2018).

This section describes the app and experimental design; estimates costs of manipulation and derives strategy-robust decision rules using our method; and compares the performance of these new algorithms to traditional learning algorithms. Our design was pre-registered and pre-specified in a pre-analysis plan (AEARCTR-0004649).

4.1 Experimental design and smartphone app

We designed our experiment to create incentives similar to those of a digital credit lending app. These apps run in the background on a smartphone, and collect data on phone use (including data on communications, mobility, social media behavior, and much more). Digital credit apps use this information to allocate loans to people who

appear creditworthy (i.e., for whom $\pi(\mathbf{x}_i)$ exceeds some threshold). Since financial regulations prevented us from actually underwriting loans to research subjects, we instead focused on analogous problems where a decisionmaker wishes to allocate resources to individuals with specific characteristics — for instance, by paying individuals based on a linear prediction of their income level or other characteristic (e.g., level of intelligence or education).[16]

Smartphone app The Smart Sensing app had two key features. First, it ran in the background to capture anonymized metadata on how individuals used their phones, such as when calls or texts were placed, which apps were installed and used, battery usage, wifi connections, and so on. In total, we captured over 1,000 behavioral indicators ('features'). Second, the app delivered weekly 'challenges' to participants, which appeared on the user's phone, and which provided financial rewards based on the user's behavior (see Figure 2). We describe these challenges in greater detail below. Participants were paid a base amount of 100 Ksh. for uploading data, plus any challenge earnings, directly via mobile money at the end of each week.

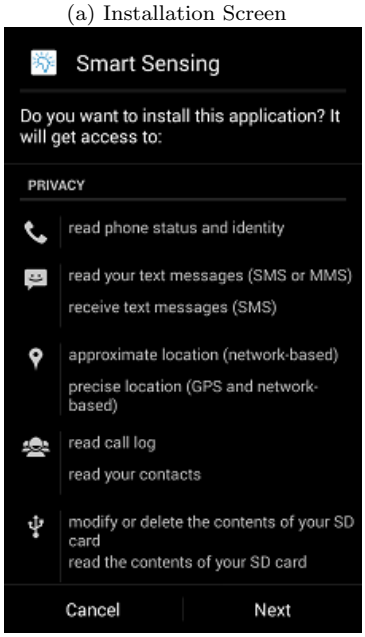
Study population The subject population consisted of Kenyans aged 18 years or older who owned a smartphone and could travel to the Busara center in Nairobi. Participants were recruited in public spaces in Nairobi, and were invited to enroll at the Busara center. During enrollment, participants completed a baseline survey and installed the Smart Sensing app on their phones. During the consenting process, participants were told about the data the app would gather and were given the opportunity to ask questions. Participants generally understood the privacy tradeoffs involved in participation.

Weekly rhythm The study followed a weekly rhythm. Each Wednesday at noon, each participant received a generic notification on their phone that said, 'Opt in to see this week's challenge!' If the participant opted in, they were shown information about the decision rule they faced that week (see Figure 2b). Challenges were valid

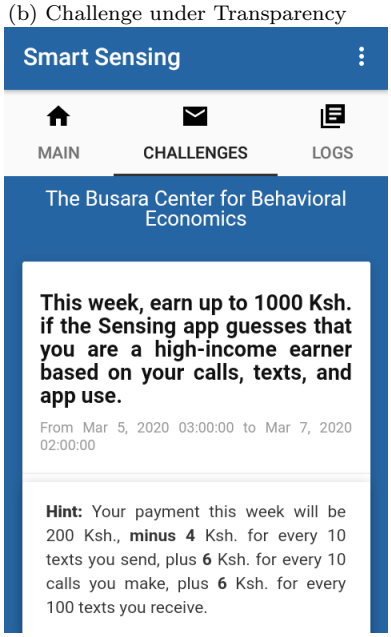
[16]While these prediction targets differ from creditworthiness, there are many settings where similar characteristics are inferred by digital traces (for example, social assistance programs that target the poor (Aiken et al., 2021), or digital advertisers who target college students).

Figure 2: Smart Sensing App

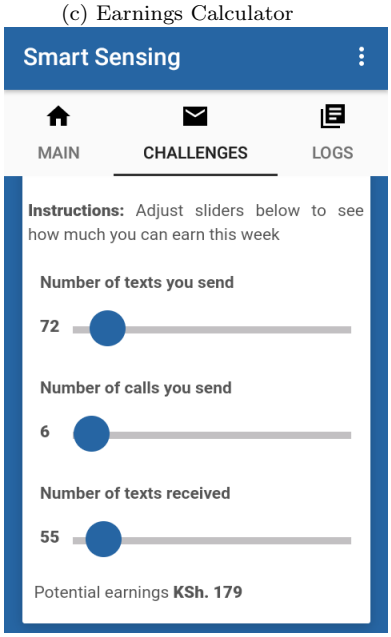
(a) Installation Screen



(b) Challenge under Transparency



(c) Earnings Calculator



until 1pm the following Tuesday. At the conclusion of the challenge, participants had 21 hours to ensure that their data was uploaded (i.e., until 10am Wednesday). Busara then determined how much each participant should be paid, and payments were sent via mobile money by noon Wednesday, at which point the next week's cycle would begin.[17]

Randomization of decision rules Each week, each participant was randomly assigned to one of three types of decision rules: control, simple, or complex. The control decision rules ($\boldsymbol{\beta}_0 \equiv \mathbf{0}$), which were deployed during the first few weeks of the experiment, did not require any action from participants; each individual who successfully uploaded their data received the same reward irrespective of how they used their phone in that week.[18] Each simple rule made decisions based on one specific behavior (e.g., $\boldsymbol{\beta}_{it} = \beta_{it}\boldsymbol{\delta}_{k_{it}}$), and were of the form, 'We'll pay you β_{it} for each behavior

[17]Participants could attrite by not opting in to the weekly challenge or by not uploading their data. The Busara center attempted to contact attriters via text message and phone call, following an attrition protocol detailed in Online Appendix S1.4. Our analysis includes only participant-weeks where the participant opted in and successfully uploaded data.

[18]Specifically, the subject received a challenge of the form, 'Dear user, you do not have to do anything for this week's challenge. You will receive an extra Ksh 100 for accepting this challenge.'

k_{it} you do', where behavior k_{it} and amount β_{it} were assigned randomly. Most incentive amounts were positive but some were negative (participants were incentivized to reduce behavior).[19] For example, one simple challenge was, 'You will receive 12 Ksh. for every incoming call you receive this week, up to Ksh. 250.' The control and simple decision rules were used to collect training data.

Finally, in the last part of the study, we assigned complex decision rules. These were designed to mimic real-world implementations of machine learning, in which people can receive a desirable benefit based on how they are classified. The complex challenges were of the form, 'We'll pay you m if the Sensing App guesses you are...'. For example, Figure 2 illustrates a highlighted challenge, 'Earn up to 1000 Ksh. if the Sensing app guesses you are a high-income earner.' Because the underlying decision rules were linear, payment was a smooth function even if the outcome was binary (in which case our decision rules represent a scaled linear probability model). Our empirical analysis presents results for this highlighted challenge as an intuitive example, and also presents results that pool over all complex challenges to gauge representative performance.

Predicting user characteristics from app data Using data from the 'control' weeks, where the app collected data on user behavior but did not provide incentives for people to change their behavior, we assess the extent to which a user's characteristics can be predicted based on how they use their phone. In Table 2, we observe that phone data can weakly predict monthly income and intelligence (above-median performance on Raven's matrices).[20]

[19] Each individual's payment was drawn from $\{-2r_k, -r_k, r_k, 2r_k, 4r_k, 8r_k\}$, for scalar r_k. We scaled the payout for each behavior so that the maximum payout ($8r_k$) was triggered by the 90th percentile of baseline behavior. For a small number of challenges, the maximum payout was predicted to be quite high so we slightly reduced the highest payout. Due to budget constraints, we could not assign simple challenges for all measurable behaviors; instead, we selected behaviors k that were predictive of main outcomes in control weeks, or which were similar to a predictive behavior. For example, if outgoing calls were predictive, we also include a corresponding measure based on incoming calls. See Online Appendix S1.6.

[20] The relatively low predictive power ($R^2 \approx 0.03$) is likely due to the fact that we have a small sample of relatively homogeneous users who are observed for a short time.

Table 2: Behavior Predicts Individual Characteristics

		Monthly Income			**Intelligence** (Above Median Ravens)	
Mean Duration of Evening Calls		-0.559	(3.702)		0.0001	(0.0002)
Mean Duration of Outgoing Calls		-1.770	(8.965)		-0.0007	(0.0004)
Calls with Non-Contacts		-42.023	(14.033)	▲●	-0.002	(0.0006)
Outgoing Text Count	▲●	10.211	(12.396)		0.0004	(0.0006)
Incoming Text Count	▲	3.888	(7.974)	▲●	-0.0002	(0.0004)
Evening Text Count	●	-9.029	(7.815)		-0.0002	(0.0003)
Outgoing Call Count	▲●	76.752	(18.133)		0.002	(0.0008)
Missed Outgoing Call Count		-84.533	(31.636)	▲	-0.003	(0.0014)
Outgoing Texts on Weekdays		-15.023	(15.210)		-0.001	(0.0007)
Max Daily Incoming Text Count		2.901	(21.212)	●	0.003	(0.0009)
Intercept		5651.04	(430.141)		0.480	(0.019)
N (individuals)		1539			1557	
R^2		0.026			0.027	

Notes: Each column represents a regression of the outcome characteristics (column header) on behaviors measured through the Sensing app (rows). Observations include data collected during the first week the participant used the sensing app. We estimate the regression model over the subset of features which were selected as one of the top-5 predictive variables by LASSO or were assigned in an SR model (including SR based on estimated costs or expert costs) for the focal two outcomes, and for which we estimate costs in the experiment (Section 4.2.3); for further details on SR models, see Section 4.3. Standard errors in parentheses. ▲ : included in incentivized naive LASSO decision rule, ● : included in incentivized strategy-robust (SR) decision rule.

4.2 Behavioral responses to simple decision rules

We next characterize how behavior responded to the simple algorithms and estimate the shape of behavioral responses, i.e., the parameters underlying each $\mathbf{x}_i^*(\boldsymbol{\beta})$.

4.2.1 Reduced form evidence

Participants changed behavior when facing simple algorithms. We demonstrate this using data from the 'simple' rules that base decisions on one specific aspect of phone use (such as increasing the number of incoming calls). Table 3 presents regressions of each participant's weekly level of different behaviors (columns) on randomly assigned incentives to change specific behaviors (rows), for a subset of the behaviors incentivized. There are three main takeaways. First, individuals manipulate the behaviors that are incentivized, as shown by the diagonal. A joint F-test that the diagonals all equal zero is rejected with $p < 0.001$.[21] Second, some behaviors are more manipulable than others. For example, the number of texts sent was 49 times more responsive to incentives than the number of people called during the workday. Finally, incentivizing one behavior can affect others, as shown in the off-diagonal elements. For example, incentivizing missed incoming calls also increased the number of texts sent (it may be that people sent messages to ask their contacts to call them back). In theory, our method can exploit these cross-elasticities, though in practice many are imprecisely estimated in our data (we find 94.5% of off diagonals are not statistically significant ($p < 0.05$), 3.6% are significantly positive and 1.8% are significantly negative).

4.2.2 Testing model assumptions

We use experimental variation in incentives to test two assumptions of our model. Here we summarize these results; details are in Online Appendix S3.1.

Quadratic costs. We find that behavior is approximately linear in incentive amounts, as would be implied by quadratic costs. However, people appear less responsive to negative incentives.

[21] Evaluated using a seemingly unrelated regression (SUR) to allow for correlation across regressions.

Table 3: Behavior Changes when Incentivized

Behavior incentivized	**Behavior observed** (change per ¢ of incentive)				
	# Texts sent	# Missed calls (incoming)	# Missed calls (outgoing)	# People called (Workdays, i.e. M-F, 9am-5pm)	# Calls w non-contacts (weekends)
# Texts sent	**24.51**	-0.052	-0.836	-0.305	-0.022
	(16.114)	(0.384)	(0.515)	(0.14)	(0.208)
# Missed incoming calls	4.15	**0.708**	0.825	0.128	-0.002
	(1.345)	(0.402)	(0.582)	(0.126)	(0.152)
# Missed outgoing calls	-0.213	0.324	**1.187**	0.22	0.502
	(1.237)	(0.24)	(0.83)	(0.194)	(0.247)
# People called (workday)	2.308	0.156	0.679	**0.497**	0.108
	(2.49)	(0.265)	(0.624)	(0.282)	(0.215)
# Calls w non-contacts (weekends)	-2.019	-0.056	1.234	0.015	**1.233**
	(2.651)	(0.165)	(0.718)	(0.164)	(0.908)
Individual Fixed Effects	X	X	X	X	X
Week Fixed Effects	X	X	X	X	X
N (person-weeks)	7966	7966	7966	7966	7966
R^2	0.704	0.552	0.637	0.604	0.491

Notes: Standard errors in parentheses, clustered at the individual level. Bold indicates diagonal: effect on behavior k when behavior k is incentivized. Each column represents a separate regression over the full set of behaviors assigned; only the first five coefficients reported here. N represents person-weeks during which 'simple' (single behavior) challenges were issued. ¢ defined as one U.S. cent, which was 1 Ksh. based on contemporaneous exchange rates.

Separable heterogeneity. We find that behaviors that are easier to manipulate on average also have higher variance across individuals when incentivized. That is, behaviors that are hard to manipulate tend to be hard for everyone, but behaviors that are easy to manipulate on average tend to be differentially manipulable for different people. This relationship is nearly proportionate, which supports the parameterization allowing separable heterogeneity by behavior and by individual.

4.2.3 Model estimates

Having confirmed that participants manipulate behavior when facing a decision rule, we use the simple and control challenges that were deployed during the first few

weeks of the experiment to estimate the parameters underlying each $\mathbf{x}_i^*(\boldsymbol{\beta})$, following Section 3. We regularize off-diagonal elements of the manipulation cost matrix to zero because they are otherwise imprecisely estimated in our sample; this results in a diagonal cost matrix C. We regularize diagonal elements using $\Lambda_{diagonal} = 1.0$, set via cross-validation.[22] We parameterize $\tilde{\gamma}_i = f(\boldsymbol{\omega}, z_i) + \tilde{v}_i = e^{-\boldsymbol{\omega}' z_i} + \tilde{v}_i$, allowing observable heterogeneity to vary with self-reported tech skills $z_i \in \{0, 1\}$, as this characteristic explained the most heterogeneity in preliminary analysis.[23]

The estimated costs of manipulation are displayed in Table 4 for the main behaviors selected by our decision rules; the costs associated with additional behaviors are shown in Appendix Table A1. We observe several intuitive patterns in these costs (top panel of Table 4). For instance, outgoing communications are less costly to manipulate than incoming communications. Text messages, which are relatively cheap to send, are more manipulable than calls, which are relatively expensive. And simpler behaviors (such as the number of texts sent) are more manipulable than complex behaviors (such as the standard deviation of texts sent by day; see Appendix Table A1).

Costs are also heterogeneous across people, as shown in the bottom panel of Table 4. On average it is 9% easier for individuals who report advanced or higher tech skills to manipulate behaviors. Including unobserved heterogeneity, the 90th percentile of gaming ability finds it twice as easy to manipulate behavior as the 10th percentile.

4.3 Complex decision rules

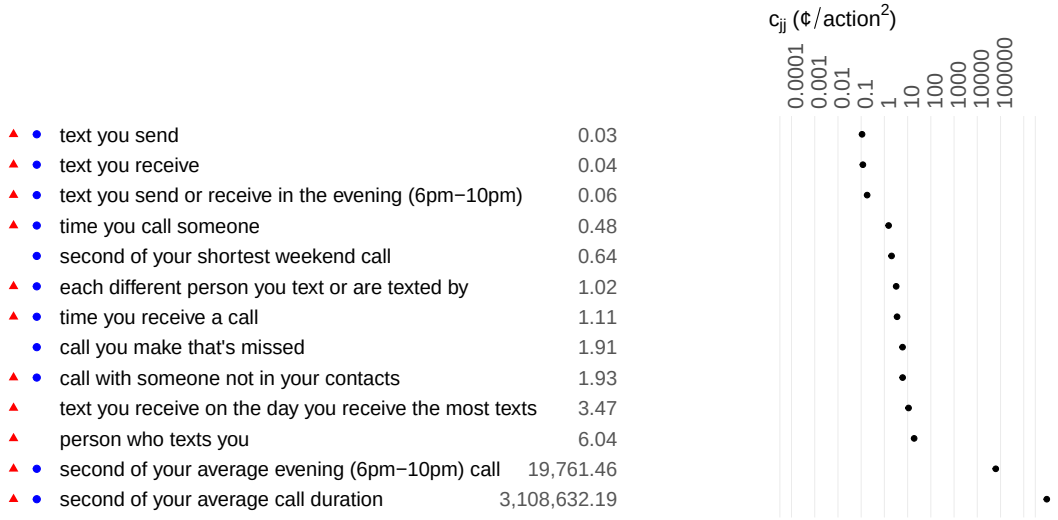
The final stage of the experiment allows us to compare the implemented performance of decision rules trained with the standard approach against those trained with the strategy-robust approach, which anticipates manipulation. In this stage, participants were randomly assigned into different target outcomes (y), decision rules (standard $\boldsymbol{\beta}^{LASSO}$ or strategy-robust $\boldsymbol{\beta}^{SR}$), and whether the decision rule was opaque or transparent to the user. Under the opaque treatment, users were told only the target outcome and the reward (i.e., Figure 2b without the Hint). Under the transparent

[22] We split the training sample randomly into thirds, estimate costs in two-thirds of the sample, penalized by a $\Lambda_{diagonal}$, and predict incentivized behavior in the held-out third. We select the $\Lambda_{diagonal}$ that yields the lowest average hold-out loss from among $\{2000, 1000, 800, 600, 400, 200, 100, 50, 25, 10, 7.5, 5, 2.5, 1, 0.5, 0.1, 0.05, 0.005, 0.0005, 0.00001, 0.0\}$.

[23] Prior to implementing phase 2, we estimated interaction effects between incentives and candidate characteristics; tech skills explained the most heterogeneity in behavioral responses.

Table 4: Estimated Manipulation Costs

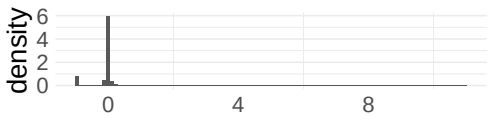
Heterogeneity by Behavior (**C** diagonal; subset of behaviors selected by decision rules)



Heterogeneity by Person ($\tilde{\gamma}_i$)

$$\tilde{\gamma}_i = e^{-\omega z_i} + \tilde{v}_i$$

Low tech skills	1.000
High tech skills	1.087



Parameters estimated using GMM. Top panel shows only behaviors used in decision rules (▲ : naive LASSO, ● : strategy-robust); for all behaviors see Appendix Table A1. In cost matrix, off-diagonal elements are regularized to zero ($\Lambda_{offdiagonal} \to \infty$), diagonal elements are regularized with $\Lambda_{diagonal} = 1.0$, set via cross validation. $\tilde{v}_i$ plot omits top 5 percent of observations.

treatment, users saw the coefficients of the decision rule, which revealed how much they would be rewarded for each behavior (the hint in Figure 2b), and were given an interactive earnings calculator (Figure 2c). Because the transparent treatment revealed information about potential decision rules, after a person had seen a transparent challenge for a given outcome, they did not later receive an opaque challenge for the same outcome.

Importantly, there was a clear separation between training and implementation: aside from the decision rule itself, no information collected during the first stage of the experiment (such as responses to the baseline survey, or any other individual-specific information) was used to determine the implemented decision.[24]

4.3.1 Estimating decision rules

Once the primitives are estimated using training data, decision rules for each outcome can be constructed. The naïve decision rule $\boldsymbol{\beta}^{LASSO}$ best predicts y_i given $\underline{\mathbf{x}}_i$ — assuming agents continue with their natural behavior. In contrast, $\boldsymbol{\beta}^{SR}$ best predicts y_i given beliefs about anticipated behavior $\mathbf{x}_i^*(\boldsymbol{\beta}^{SR})$, following equation (6). We include a LASSO regularization term tuned to ensure all decision rules have at most three features. This can improve small sample performance, and also helps ensure that participants—who tend to have low-cost mobile phones with small screens—can more easily view and understand the decision rule.[25]

For $\boldsymbol{\beta}^{SR}$, we used simple heuristics to select the shrinkage parameter ϕ for unobserved gaming ability $\tilde{V}$. In the initial weeks of implementation, we had no data on performance of full decision rules and so set ϕ to be large enough that in simulations the strategy-robust decision rule differed meaningfully from the naïve one. After several weeks had passed, we assessed the fit of the models to actual behavior as a function of ϕ; based on this we opted to keep the initial selection of ϕ (with the

[24] One concern is that performance assessed in our final stage may be artificially high since average costs are estimated based on the responses of a sample that includes some of the same individuals. In a robustness test, we exclude individual-weeks with complex challenges if that individual received a simple challenge for any of the complex challenge's constituent behaviors during the training period. Results are similar; see Online Appendix Table S1, columns 3-4.

[25] Specifically, for each outcome, we selected the smallest penalty $\underline{\lambda}^{3var}$ that yielded a naïve decision rule with at most three features. When estimating the strategy-robust decision rule, we consider decision rules representing all combinations of at most three features, each estimated using the same $\underline{\lambda}^{3var}$ penalty selected for the naïve rule.

exception of one week in 2020 where we tried a different setting).

Differences between implemented and final models. The experimental timeline was such that we could not pause the experiment in order to compute parameter estimates from incoming data. Thus, the decision rules implemented in the experiment used a few simplifications. First, when estimating costs, we estimated $\underline{\mathbf{x}}_i$ from regressions including week fixed effects, but when estimating decision rules from equation (6), we plugged in the simple average of $\mathbf{x}_i$ during control weeks for $\underline{\mathbf{x}}_i$ (without week fixed effects). The former may yield better small sample performance but both yield valid estimates of $\underline{\mathbf{x}}_i$. Second, we used preliminary estimates of the cost parameters.[26] Third, the decision rules assigned prior to January 21, 2020 included an error in computing the distribution of unobserved heterogeneity where one component was inverted; this led to a minor change in the noise distribution (see Online Appendix S1.7.2 for details). After the experiment's conclusion, we reestimated decision rules using final costs (which resolve the latter two issues); we report these resulting parameters and predicted losses in the paper. These final costs imply a different scale for V, so we use a value of ϕ that produces a similar set of optimal decision rules as those implemented in our study. Specifically, we calibrated $\phi = 10^{-6}$ to minimize the L2 distance between the coefficients estimated with the final procedure and those that were implemented, for outcomes intelligence and income.[27] Small changes resulted in a slightly different LASSO hyperparameter $\underline{\lambda}^{3var_{final}}$.[28]

The deviations between the implemented and final decision rules are minor. The cost estimates are very similar, as are the resulting decision rules; see Online Appendix S1.7.2. In particular, we can compute the expected performance improvement of decision rule $\boldsymbol{\beta}$ as $\Delta L(\boldsymbol{\beta}) = \frac{L_{\boldsymbol{C}}(\hat{\boldsymbol{\beta}}^{naïve}) - L_{\boldsymbol{C}}(\boldsymbol{\beta})}{L_{\boldsymbol{C}}(\hat{\boldsymbol{\beta}}^{naïve}) - L_{\boldsymbol{C}}(\hat{\boldsymbol{\beta}}^{SR_{final}})}$, where $L_{\boldsymbol{C}}(\boldsymbol{\beta})$ represents the objective used in equation (6). If $\boldsymbol{\beta}$ is expected to achieve the same loss as the optimal decision rule, $\Delta L(\boldsymbol{\beta}) = 100\%$; if it is expected to achieve loss as the naïve rule (and thus is

[26] At the time, we thought these were final. However, after the experiment concluded, we re-checked the optimality conditions and found that the costs estimated during the experiment were a local optimum.

[27] Final costs use the same cost matrix penalization $\Lambda_{diagonal} = 1$ as the implemented costs.

[28] The updated penalization protocol yields decision rules closer to the boundary of 3 coefficients, and we also changed the LASSO sample to coincide with that used for the SR rule (including only individuals with nonmissing tech skills, which drops 1.5 percent of the sample). Minor changes were also made to simplify the method.

expected to offer no strategy-robust improvement), then $\Delta L(\boldsymbol{\beta}) = 0$. By this metric, the median implemented rule achieves 90% of the loss improvement of the optimal final rule.

4.3.2 Indicative results: decision rules shift behavior

Table 5 provides suggestive evidence of how decision rules affect behavior. The first panel simply indicates the estimated naïve decision rule: high-income people make more outgoing calls, send fewer texts, and receive more texts. In the second panel, each person was rewarded 'if the app guesses you are a high-income earner', we see that if people are not told the decision rule (corresponding to the 'Opaque challenge'), the response is not statistically significant and often in the wrong direction on average (i.e., participants place fewer calls and send more texts). However, participants assigned the transparent treatment change their behavior broadly in the direction rewarded by the algorithm, though the response is imprecise. This pattern holds when pooling over all complex decision rules we assigned: we find that for the opaque treatment, 38.5% of estimated effects are in the same direction as the assigned behavior incentive and 14.0% are in the same direction and statistically significant, while 61.4% are in the opposite direction and 21.1% are in the opposite direction and statistically significant; but for the transparent treatment, 75.4% of point estimated effects are in the same direction as the assigned behavior incentive and 16.7% are in the same direction and statistically significant, and 25.6% are in the opposite direction and 3.5% are in the opposite direction and statistically significant.

4.4 Results: Naïve vs. strategy-robust decision rules

Our main empirical results, shown in Table 6, compare the performance of the naïve and strategy-robust decision rules implemented during our experiment.[29] The first two columns (under 'Income') show results for the challenge that rewarded participants for using their phones like a high-income earner; the last two columns show performance

[29]We note that these performance results differ in nature from typical evaluations of models in computer science and structural modeling in that we assess performance on prospective rather than retrospective data: our experiment implemented the decision rules, and the evaluation is thus equivalent to assessing each decision rule on a hold-out set drawn from a different (manipulated) distribution.

Table 5: Agents Game Algorithms

	Calls (outgoing)	Texts (outgoing)	Texts (incoming)	Calls w con-contacts (incoming + outgoing)	Avg call length (evening, seconds)
Panel A: **Incentives generated by algorithm (¢/action)**					
$\boldsymbol{\beta}^{LASSO}$	0.625	-0.395	0.065	0	0
Panel B: **Regression of $\mathbf{x}_{it}$ (column label) on treatment assignment (row label)**					
Opaque challenge	-4.7	12.5	17.4	0.8	-4.3
	(8.3)	(13.4)	(22.1)	(3.5)	(4.7)
Transparent challenge	13.7	-17.5	-6.5	0.3	-2.1
	(12.9)	(11.0)	(18.8)	(3.3)	(4.4)
N (Person-weeks)	1651	1651	1651	1651	1651

Notes: Panel A reports the decision rule for the challenge, 'Earn up to 1000 Ksh. if the app guesses you are a high-income earner!' Panel B reports how behaviors (columns) depend on whether participants are randomly assigned to the opaque challenge (which provides no information about the decision rule) or the transparent challenge (which reveals the details of the decision rule). The sample includes all people who were assigned this challenge, in the week they were assigned that challenge as well as control weeks. Standard errors in parentheses, clustered at the individual level.

averaged across all outcomes. The decision rules and associated manipulation costs are shown in Panel A; the relative performance of the different decision rules is shown in Panel B. We note several results.

First, Panel A highlights differences in the decision rules. LASSO places weight on the behaviors that were most correlated with the outcome at baseline: outgoing calls, outgoing texts, and incoming texts. However, some of these behaviors, particularly text messaging, are quite manipulable (as shown in the 'Costs' column) and thus subject to more manipulation noise. Although adjustments made by the strategy-robust approach can be subtle depending on how gaming ability correlates with the outcome, here the decision rule attenuates or drops more manipulable behaviors (for example, dropping incoming texts in favor of evening texts, which we estimate are harder to manipulate).

We evaluate predictive performance using root mean squared error (RMSE), in units of US dollars, in Panel B. This measures how far off the payments we gave to people (based on the decision rule and their behavior that week) were from what we desired to give to them (based on their fixed characteristic that was targeted). The first

pair of rows report the prediction error that would be expected ex ante. The first row shows that if there were no manipulation and behavior were the same as control weeks, LASSO would be expected to perform marginally better than our strategy-robust estimator (by \$0.01 for income; \$0.08 for all decision rules pooled). The second row shows the error predicted by our model if the rule were made transparent and people were manipulating behavior: here, the strategy-robust method is expected to perform better (by \$0.11 for income; \$3.76 pooled).

Table 6: Strategy-Robust vs. Standard Decision Rules

	Income		*Costs*	**Pooled: All Weeks**	
Panel A: **Decision Rule**	$\boldsymbol{\beta}^{LASSO}$	$\boldsymbol{\beta}^{SR}$	c_{kk}	$\boldsymbol{\beta}^{LASSO}$	$\boldsymbol{\beta}^{SR}$
	¢/action		¢/action^2	.	
# Texts (outgoing)	-0.395	-0.107	0.035	.	.
# Texts (incoming)	0.065	0	0.037	.	.
# Texts (6pm-10pm)	0	-0.121	0.057	.	.
# Calls (outgoing)	0.625	0.542	0.480	.	.
Intercept (α)	301.071	304.622		.	.
Panel B: **Prediction Error**	RMSE ($)			RMSE ($)	
Training Data: Control	3.574 (0.058)	3.583 (0.052)‡		3.660 (0.018)	3.737 (0.018)‡
Training Data: Predicted Transparent	3.702 (0.058)	3.591 (0.056)‡		7.777 (0.212)	4.018 (0.024)‡
Implemented: Opaque	3.549 (0.249)	3.525 (0.218)		3.780 (0.078)	3.710 (0.070)
Implemented: Transparent	3.675 (0.179)	3.484 (0.200)		4.641 (0.167)	4.130 (0.127)‡
Average Payout ($)	3.34	3.25		3.96	3.54
N (Control Individuals)	1376	1376		1391	1391
N (Treatment Person-Weeks, Opaque)	75	75		1344	1344
N (Treatment Person-Weeks, Trans.)	90	74		1246	1298

Notes: The first three columns focus on an example challenge (income); the remaining columns pool all challenges (adding marital status, whether self-reported tech skills are advanced, PCA of number of friends, PCA of phone activity, baseline number of texts received (self-reported), number of texts sent in first control week, and intelligence (above-median Ravens)). Panel A reports decision rules and the manipulation cost estimates for included behaviors. Panel B reports performance using root mean squared error (RMSE), during control weeks ('control'), as predicted by the model of behavior ('predicted transparent'), or when assigned in the experiment with/without transparency hints ('implemented transparent/opaque', respectively). Predicted behavior is based on final cost model estimates. Pooled columns evaluate performance in training data by averaging over individuals and then over outcomes; if an outcome was assigned in multiple weeks, we randomly assign each individual to one decision rule for that outcome. Opaque performance is evaluated across all individuals assigned to opaque for that outcome in that week, regardless of which decision rule was used, since individuals were not shown the rule. In parentheses we report bootstrapped standard errors from 50 draws of individuals with replacement. ‡ indicates the performance difference between $\boldsymbol{\beta}^{LASSO}$ and $\boldsymbol{\beta}^{SR}$ is significant at $p < 0.05$. (Because the first three performance rows are evaluated on the same set of individuals between $\boldsymbol{\beta}^{LASSO}$ and $\boldsymbol{\beta}^{SR}$, the differences are estimated relatively precisely.) Average payout indicates the average payout in the transparent treatment. SR decision rule is estimated using preliminary costs estimates.

The next pair of rows report the prediction error we obtained when the decision rules were implemented experimentally. These may differ from the expected prediction error either if people respond differently than anticipated by our model, or because of noise from week to week. Here, we find that the strategy-robust (SR) method performs better than LASSO when participants are given full information about the decision rule, by \$0.19 (5%) for income and \$0.51 (11%) across pooled outcomes. The observed difference in RMSE between transparent naive and transparent strategy-robust decision rules for the single income outcome is not statistically significant ($p = 0.495$), but across all decision rules pooled, the observed difference in RMSE is significant with $p = 0.021$. The strategy-robust method also performs slightly better when the decision rule is opaque (by \$0.02 / 0.6% for income, $p = 0.096$; \$0.07 / 2% pooled, $p = 0.089$), although this difference is imprecise.[30]

Performance cost of transparency This framework also makes it possible to assess the cost of making decision rules transparent. Making the naïve decision rules transparent reduces their performance on average by 23% (\$4.641 vs. \$3.780; s.e. 5.9 p.p.), as shown in the last two columns of Table 6, Panel B. However, if alongside making the decision rule transparent one also switched to strategy-robust rules, performance declines by only 9.3% (\$4.130 vs \$3.780; s.e. 4.2 p.p.). Thus, in this context, strategy-robust rules cut the cost of transparency by 59% (s.e. 18.7 p.p.).

Our model also allows us to estimate the cost of transparency for strategy-robust rules without implementing them, since the calculation can be done with the primitives estimated from the first part of our experiment. Our model predicts that the cost of making strategy-robust decision rules transparent will be \$4.018 - \$3.660 = 0.358 (9.8%), which is close to the implemented cost of \$4.130 - \$3.780 = 0.350 (9.3%).[31]

Anticipating performance Online Appendix S3.5 explores how well our model (and others) predict behavior under complex challenges. We compute the rank

[30] P-values computed using bootstrapped standard errors for the difference in RMSE across treatments, measured by resampling with replacement across individuals. Opaque comparisons face less sampling variation because individuals assigned to opaque can be used to assess either decision rule.

[31] This approach does well at ranking relative performance, but is less accurate at predicting the exact level of performance of the more manipulable naïve decision rules; see Online Appendix S3.5.

correlation between the loss a decision rule attains when implemented and the loss it is predicted to attain by various models. Under the opaque treatment, loss is best predicted by baseline behavior (where the policymaker assumes that behavior will follow $\mathbf{x}^*(\boldsymbol{\beta}) \equiv \underline{\mathbf{x}}$); for the transparent treatment, loss is better anticipated by the behavior predicted with the strategy-robust adjustment.

We also assess the importance of manipulation noise in explaining behavior under the transparent treatment, and find an inverse U-shape relationship. Namely, when $\phi = 0$, the model omits manipulation noise ($\tilde{v}_{ib} \equiv 0$), and is only as good as baseline behavior at predicting loss. The strategy-robust model that best predicts loss uses a value of ϕ slightly below our final choice of $\phi = 10^{-6}$. When ϕ is much higher, the model anticipates too much noise and predicts worse than assuming baseline behavior remains fixed. In other words, much of the strategy-robust correction comes from properly incorporating manipulation noise.

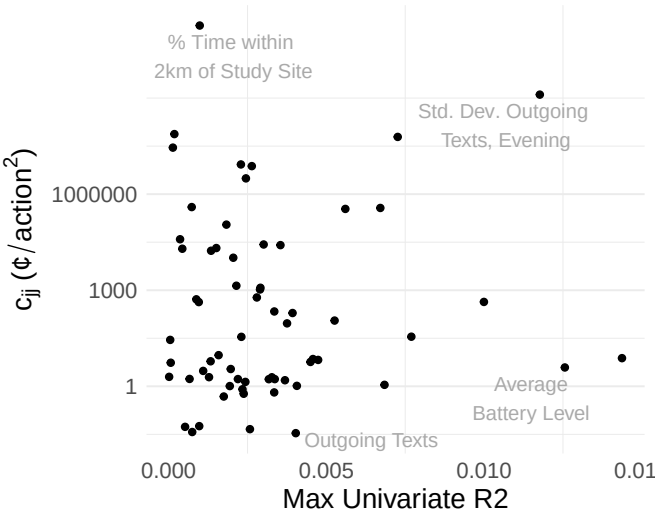
5 Discussion

5.1 Contrast to standard approaches

Standard loss functions evaluate each feature based on its correlation with the outcome within a training dataset, as in equation (3). However, as can be seen in Figure 3, features that appear equally predictive given a current vector of behaviors $\mathbf{x}_i$ can have wildly different manipulability. The figure compares, for a set of features, the estimated manipulability (y-axis) of each feature to the highest univariate baseline predictive power (x-axis) that feature attains for two focal outcomes: income and intelligence (Raven's score). Some of the most predictive features (like the average battery level on the person's phone) are easy to manipulate.

Contrast with the 'intuitive' approach An alternate, 'intuitive' approach would simply exclude the most manipulable features from the decision rule, for instance by only considering features above some y-axis threshold on Figure 3. We assess this approach in Online Appendix S3.4 for income and intelligence. This approach reduces the predicted manipulability of decision rules, but also removes useful features, which in some cases decreases predicted performance. In extreme cases, decision

Figure 3: Manipulation Costs vs. Baseline Predictive Power



Dots indicate behaviors recorded by the Sensing app. Figure compares estimated manipulation cost (y-axis) to the highest R^2 across income and intelligence (x-axis), with illustrative features labeled.

rules resulting from regularized models such as LASSO can be left with no behaviors that are predictive enough to include in the regression. In contrast, our approach can extract signal even from manipulable behaviors, and performs better in these simulations.

Contrast with the iterative retraining approach A second approach, common in industry, retrains a naïve machine learning estimator iteratively. This is equivalent to iterated best response without commitment. In the n^{th} iteration of applying equation (3), the best response decision rule is $\boldsymbol{\beta}^{OLS(n)} = (\mathbf{x}^{(n-1)\prime}\mathbf{x}^{(n-1)})^{-1}(\mathbf{x}^{(n-1)\prime}\boldsymbol{y})$, where $\mathbf{x}^{(n-1)}$ represents the covariate matrix in the $n-1^{th}$ iteration and $\boldsymbol{y}$ the vector of outcomes. In a greenfield application, one initially observes $\mathbf{x}^{(0)} = \underline{\mathbf{x}}$. But our model suggests that when $\boldsymbol{\beta}^{OLS(n)}$ is implemented, behavior changes according to equation (1): $\mathbf{x}^{(n)} = \underline{\mathbf{x}} + C_i^{-1}\boldsymbol{\beta}^{OLS(n)}$. Define the map

$$\Phi(\boldsymbol{\beta}) \coloneqq \left[(\underline{\mathbf{x}} + C_i^{-1}\boldsymbol{\beta})'(\underline{\mathbf{x}} + C_i^{-1}\boldsymbol{\beta})\right]^{-1}\left[(\underline{\mathbf{x}} + C_i^{-1}\boldsymbol{\beta})'\boldsymbol{y}\right]$$

Then the iteration of this process can be written, $\boldsymbol{\beta}^{OLS(n)} = \Phi(\boldsymbol{\beta}^{OLS(n-1)})$. In some cases there may exist a fixed point $\boldsymbol{\beta}^{FP} = \Phi(\boldsymbol{\beta}^{FP})$ — a Nash solution, as explored in related work by Perdomo et al. (2020) and Frankel and Kartik (2020). However, in

others, iteration cycles. In contrast, $\boldsymbol{\beta}^{SR}$ jumps to the solution with commitment (a Stackelberg solution).

In simulations, the performance of an iterated best-response method approaches the strategy-robust method after approximately 4 iterations (see Online Appendix S3.4, which uses a LASSO estimator and the income and intelligence outcomes). However, beyond that point, performance begins to deteriorate. When predicting income this deterioration is small; for intelligence, performance eventually falls below the performance obtained before any retraining.

Note that in our model, behavior responds according to true costs, but the one-shot solution with commitment responds to a belief of the costs (in $\tilde{F}$). $\boldsymbol{\beta}^{SR}$ will theoretically tend to dominate $\boldsymbol{\beta}^{FP}$ so long as the belief is sufficiently correct. However, as we discuss below, in many real-world settings, behavior updates with noise.

5.2 Learning

Individuals may have noisy beliefs about how decisions are made. For example, when the parameters of the actual decision rule are α and $\boldsymbol{\beta}$, individual i might believe that the decision rule is $\tilde{\alpha}_i, \tilde{\boldsymbol{\beta}}_i \sim \tilde{D}_i(\alpha, \boldsymbol{\beta})$. Behavior would then follow the generalization of equation (1),

$$\mathbf{x}_i^*(\alpha, \boldsymbol{\beta}) = \arg\max_{\mathbf{x}_i} \left[\mathbb{E}_{\tilde{D}_i(\alpha,\boldsymbol{\beta})} \left[g_i \left(\tilde{\alpha}_i + \tilde{\boldsymbol{\beta}}_i \cdot \mathbf{x}_i \right) \right] - c(\mathbf{x}_i, \underline{\mathbf{x}}_i) \right]. \tag{7}$$

where $g_i(y)$ represents a utility function. Individuals thus balance the cost of manipulation against their expected utility gain.

Our main utility model is linear ($g_i(y) = y$); because it has no risk aversion, uncertainty would not affect expected behavior. However, if individuals were risk averse ($\frac{\partial^2 g_i}{\partial y^2} < 0$), a mean-preserving spread in $\tilde{\boldsymbol{\beta}}_i$ would reduce the incentive to manipulate and a policymaker could reduce manipulation by obfuscating the decision rule. However, this approach undermines a major goal of transparency: that people know how they are evaluated. In some settings with uncertainty, the linear model may reasonably approximate the distribution of beliefs and risk aversion.[32]

[32]Individuals often have difficulty understanding the complex functional forms that arise from machine learning (Du et al., 2019; Poursabzi-Sangdeh et al., 2021), and commonly use heuristics when facing nonlinear functions (Liebman and Zeckhauser, 2004). To make a decision rule robust to

5.3 Alternate methods to estimate manipulation costs

In some settings, it may be cheaper or easier to use alternative, non-experimental approaches to estimate manipulation costs. We briefly explore an approach that elicits costs through hypothetical questions. We conducted a survey asking 171 individuals to predict how Kenyans would manipulate different phone behaviors when incentivized, in the spirit of DellaVigna and Pope (2016).[33] Although respondents generally predicted costs to be lower than what we found in the experiment, the correlation between the two estimates is 0.30, as shown in Figure A1. If we use hypothetical predictions of manipulation costs to train decision rules, and then predict performance based on the experimentally estimated model of behavior, even these heuristic estimates improve simulated performance substantially for one focal outcome, and have an inconsequential negative effect on the other, as shown in Table A2. See Online Appendix S2.

In some cases, it may also be possible to estimate the cost of underlying manipulations from market prices and first principles. A structural model of costs would allow an implementer to model changes in these underlying parameters, suggesting how manipulation will change if, for example, the price of calls changed.

6 Conclusion

This paper considers the possibility that machine decisions change the world in which they are deployed. We develop an estimable approach that builds on recent theoretical work, which anticipates manipulation by embedding a behavioral model of how individuals will respond within a predictive loss function. We stress test this approach in a field experiment in Kenya, by deploying a custom smartphone app intended to mimic the digital loan products that are now commonplace in sub-Saharan Africa. We find that even some of the world's poorest users of technology—who are relatively recent adopters of smartphones and for whom the concept of an 'algorithm' is quite foreign (Musya and Kamau, 2018)—are savvy enough to change their behavior to game algorithmic decision rules.

manipulation, it may be sufficient to make it robust to these heuristic responses.

[33] Respondents included PhDs in related fields, research assistants, Busara staff who had not worked on the experiment, and Mechanical Turk workers in the U.S.

We document three advantages of the strategy-robust approach. First, strategy-robust decision rules perform better when implemented: when individuals are given information about the rule, strategy-robust rules outperform standard estimators by 12% on average. Second, strategy-robust models better anticipate the relative performance that decision rules will achieve when implemented with transparency. Third, it is possible to estimate the 'cost of transparency': the loss in predictive performance associated with moving from 'security through obscurity' (with a naïve decision rule) to a regime of transparency (with a strategy-robust rule). We estimate this loss to be 9.3% in equilibrium — less than the 23% loss associated with making the naïve rule transparent.

While we focus on the simple case of linear decision rules about which subjects have either no or full information, we envision extensions to more complex rules and more nuanced beliefs. Combining machine learning estimators with models of human behavior is likely to be relevant to a wide range of contexts where systems face changing environments.

This structural approach is different from the approach to machine learning most commonly used in practice, which relies on large amounts of data and flexible functions that impose few assumptions about how the data are generated. A central problem with the status quo approach is that it often performs better in the lab than when implemented (cf. Lazer et al., 2014; Andrews et al., 2023). We study one particular implementation issue — strategic manipulation — and show that the counterfactual world that emerges after implementing $\boldsymbol{\beta}$ has a predictable structure: including a variable in a decision rule tends to induce manipulation and spread in that variable, in proportion to its costs and benefits. While the costs must be estimated, benefits can be inferred directly, because they are a function of the estimand $\boldsymbol{\beta}$.

This structure makes it possible to predict counterfactual fit, and more efficiently identify the decision rules that will perform well when implemented. Our structural approach decomposes decision rules into constituent components, and gathers data on how those components can be manipulated. From these components, the model allows us to understand how *any* proposed decision rule of a given form would be manipulated, and to compute decision rules that are optimal in equilibrium.

In this sense, our paper offers a machine learning interpretation of Lucas (1976),

where algorithmic decisions change the context of the systems they model. In settings like ours, $\boldsymbol{\beta}$ determines not just predictive performance within a given world, but also which counterfactual world comes to exist.

References

Agarwal, Nikhil and Eric Budish, "Market Design," *Handbook of Industrial Organization*, 2021.

Aiken, Emily, Suzanne Bellue, Dean Karlan, Christopher Udry, and Joshua E Blumenstock, "Machine Learning and Mobile Phone Data Can Improve the Targeting of Humanitarian Assistance," *Working Paper*, July 2021.

Akerlof, George A., "The economics of "tagging" as applied to the optimal income tax, welfare programs, and manpower planning," *The American economic review*, 1978, *68* (1), 8–19.

Andrews, Isaiah, Drew Fudenberg, Lihua Lei, Annie Liang, and Chaofeng Wu, "The Transfer Performance of Economic Models," 2023.

Ball, Ian, "Scoring Strategic Agents," *arXiv:1909.01888 [econ]*, November 2019. arXiv: 1909.01888.

Banerjee, Abhijit, Rema Hanna, Benjamin A Olken, and Sudarno Sumarto, "The (lack of) Distortionary Effects of Proxy-Means Tests: Results from a Nationwide Experiment in Indonesia," Working Paper 25362, NBER December 2018.

Bharadwaj, Prashant and Tavneet Suri, "Improving Financial Inclusion through Digital Savings and Credit," *AEA Papers and Proceedings*, May 2020, *110*, 584–588.

Björkegren, Daniel, "'Big data' for development," 2010.

_ **and Darrell Grissen**, "Behavior Revealed in Mobile Phone Usage Predicts Credit Repayment," *The World Bank Economic Review*, October 2020, *34* (3), 618–634.

Björkegren, Daniel, Joshua Blumenstock, Omowunmi Folajimi-Senjobi, Jacqueline Mauro, and Suraj R. Nair, "Instant Loans Can Lift Subjective Well-Being: A Randomized Evaluation of Digital Credit in Nigeria," *arXiv:2202.13540 [econ, q-fin]*, February 2022. arXiv: 2202.13540.

Bloomberg, "Phone Stats Unlock a Million Loans a Month for Africa Lender," *Bloomberg.com*, September 2015.

Blumenstock, Joshua E., "Estimating Economic Characteristics with Phone Data," *AEA Papers and Proceedings*, 2018, *108*, 72–76.

Blumenstock, Joshua Evan, Gabriel Cadamuro, and Robert On, "Predicting poverty and wealth from mobile phone metadata," *Science*, November 2015, *350* (6264), 1073–1076.

Borrell Associates, "Trends in Digital Marketing Services," 2016.

Breiman, Leo, "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)," *Statistical Science*, August 2001, *16* (3), 199–231. Publisher: Institute of Mathematical Statistics.

Bruckner, Michael and Tobias Scheffer, "Stackelberg Games for Adversarial Prediction Problems," in "KDD" ACM New York, NY, USA 2011, pp. 547–555.

CGAP, "Kenya's Digital Credit Revolution Five Years On," *CGAP*, March 2018.

Crosman, Penny, "How fraudsters are gaming online lenders," *American Banker*, March 2017.

Dee, Thomas S., Will Dobbie, Brian A. Jacob, and Jonah Rockoff, "The Causes and Consequences of Test Score Manipulation: Evidence from the New York Regents Examinations," *American Economic Journal: Applied Economics*, July 2019, *11* (3), 382–423.

DellaVigna, Stefano and Devin Pope, "Predicting Experimental Results: Who Knows What?," Working Paper 22566, National Bureau of Economic Research August 2016.

Dong, Jinshuo, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu, "Strategic Classification from Revealed Preferences," in "EC" EC '18 ACM Press New York, NY, USA 2018, pp. 55–70.

Du, Mengnan, Ninghao Liu, and Xia Hu, "Techniques for interpretable machine learning," *Communications of the ACM*, December 2019, *63* (1), 68–77.

European Union, "EU General Data Protection Regulation (GDPR)," 2016.

Francis, Eilin, Joshua Blumenstock, and Jonathan Robinson, "Digital Credit: A Snapshot of the Current Landscape and Open Research Questions," *CEGA White Paper*, 2017.

Frankel, Alex and Navin Kartik, "Muddled Information," *Journal of Political Economy*, August 2019, *127* (4), 1739–1776.

_ and _ , "Improving Information from Manipulable Data," *arXiv:1908.10330 [econ]*, April 2020. arXiv: 1908.10330.

FSD Kenya, "Tech-enabled lending in Africa," 2018.

Ghalme, Ganesh, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen, and Nir Rosenfeld, "Strategic classification in the dark," in "International Conference on Machine Learning" PMLR 2021, pp. 3672–3681.

Gonzalez-Lira, Andres and Ahmed Mobarak, "Slippery Fish: Enforcing Regulation under Subversive Adaptation," IZA Discussion Paper 12179, Institute of Labor Economics (IZA) February 2019.

Goodhart, Charles, *Monetary Relationships: A View from Threadneedle Street*, University of Warwick, 1975. Google-Books-ID: GKwJMwEACAAJ.

Goodman, Bryce and Seth Flaxman, "European Union regulations on algorithmic decision-making and a "right to explanation"," *arXiv:1606.08813 [cs, stat]*, June 2016. arXiv: 1606.08813.

Greenstone, Michael, Guojun He, Ruixue Jia, and Tong Liu, "Can Technology Solve the Principal-Agent Problem? Evidence from Pollution Monitoring in China," 2019.

Hardt, Moritz, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters, "Strategic Classification," in "Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science" ITCS '16 ACM New York, NY, USA 2016, pp. 111–122.

Hennessy, Christopher A. and Charles A. E. Goodhart, "Goodhart's Law and Machine Learning: A Structural Perspective," *International Economic Review*, 2023, *n/a* (n/a). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/iere.12633.

Hu, Lily, Nicole Immorlica, and Jennifer Wortman Vaughan, "The Disparate Effects of Strategic Manipulation," *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, 2019, pp. 259–268. arXiv: 1808.08646.

Jagadeesan, Meena, Celestine Mendler-Dünner, and Moritz Hardt, "Alternative microfoundations for strategic classification," in "International Conference on Machine Learning" PMLR 2021, pp. 4687–4697.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein, "Discrimination In The Age Of Algorithms," Working Paper 25548, National Bureau of Economic Research February 2019.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani, "The Parable of Google Flu: Traps in Big Data Analysis," *Science*, March 2014, *343* (6176), 1203–1205.

Liebman, Jeffrey and Richard J. Zeckhauser, "Schmeduling," 2004.

Liu, Lydia T., Nikhil Garg, and Christian Borgs, "Strategic ranking," in "Proceedings of The 25th International Conference on Artificial Intelligence and Statistics" PMLR May 2022. ISSN: 2640-3498.

Lucas, Robert E., "Econometric policy evaluation: A critique," *Carnegie-Rochester Conference Series on Public Policy*, January 1976, *1* (Supplement C), 19–46.

Miller, John, Smitha Milli, and Moritz Hardt, "Strategic classification is causal modeling in disguise," in "Proceedings of the 37th International Conference on Machine Learning," Vol. 119 of *ICML'20* JMLR.org July 2020, pp. 6917–6926.

Mirrlees, J. A., "An Exploration in the Theory of Optimum Income Taxation," *The Review of Economic Studies*, 1971, *38* (2), 175–208.

Musya, Mercy and Grace Kamau, "How do you say "algorithm" in Kiswahili?," December 2018. Library Catalog: medium.com.

National Institute of Standards and Technology, "Guide to General Server Security," *NIST Special Publication*, July 2008, (800-123).

Nichols, Albert L. and Richard J. Zeckhauser, "Targeting Transfers through Restrictions on Recipients," *The American Economic Review*, 1982, *72* (2), 372–377.

OSTP, "Blueprint for an AI Bill of Rights," Technical Report 2022.

Perdomo, Juan, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt, "Performative prediction," in "International Conference on Machine Learning" PMLR 2020, pp. 7599–7609.

Poursabzi-Sangdeh, Forough, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach, "Manipulating and Measuring Model Interpretability," in "CHI" 2021 Association for Computing Machinery New York, NY, USA May 2021, pp. 1–52.

Ramsey, F. P., "A Contribution to the Theory of Taxation," *The Economic Journal*, 1927, *37* (145), 47–61.

Sayed-Mouchaweh, Moamar and Edwin Lughofer, *Learning in Non-Stationary Environments: Methods and Applications*, Springer Science & Business Media, April 2012. Google-Books-ID: qFWM2nva7xQC.

Spence, Michael, "Job Market Signaling," *The Quarterly Journal of Economics*, 1973, *87* (3), 355–374.

Appendices

A1 Estimation Extensions

We list moment conditions formulated in the case where shocks have been decomposed, $\boldsymbol{\epsilon}_{it} = \boldsymbol{\mu}_t + \boldsymbol{\eta}_{it}$, and common shocks $\boldsymbol{\mu}_t$ are absorbed with time fixed effects. They also apply without time fixed effects, in which case one may omit $\boldsymbol{\mu}_t$ and swap $\boldsymbol{\epsilon}_{it}$ for $\boldsymbol{\eta}_{it}$. These systems are overidentified, and we weight moments equally.

A1.1 Moment Conditions for Greenfield Case

Implemented decision rules are orthogonal to idiosyncratic behavior shocks and manipulation noise ($\mathbb{E}[\beta_{itk}\eta_{itj}] = 0$ and $\mathbb{E}[\beta_{itk}v_i] = 0$). For each pair of behaviors jk (including $j = k$) this yields sample moment condition

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{t\in\mathbb{T}_i}\beta_{itk}\left[x_{ijt} - \underline{x}_{ij} - \mu_{jt} - f(\boldsymbol{\omega}, \mathbf{z}_i)\cdot[C^{-1}\boldsymbol{\beta}_{it}]_j\right] = 0 \tag{8}$$

where $[\mathbf{a}]_k$ indicates the kth element of $\mathbf{a}$.

One can form an estimate of unobserved heterogeneity $\hat{v}_i$ by

$$\hat{v}_i = \frac{1}{\sum_{t\in\mathbb{T}_i^{treatment}}|K_{it}^{eval}|}\sum_{t\in\mathbb{T}_i^{treatment}}\sum_{k\in K_{it}^{eval}}\left[\frac{x_{ikt} - \underline{x}_{ik} - \mu_{kt}}{[C^{-1}\boldsymbol{\beta}_{it}]_k} - f(\boldsymbol{\omega}, \mathbf{z}_i)\right], \tag{9}$$

where K_{it}^{eval} is the set of behaviors to be evaluated for i in period t.[34] Unobserved heterogeneity is mean zero, yielding moment condition, $\frac{1}{N}\sum_i \hat{v}_i = 0$, and orthogonal to each heterogeneity characteristic, yielding moment condition(s) $\frac{1}{N}\sum_i z_{li}\cdot\hat{v}_i = 0$ for each characteristic l.

[34] We set $K_{it}^{eval} = \{k \text{ s.t. } \beta_{itk} \neq 0\}$, so that $\hat{v}_i$ is evaluated on shifts in the incentivized behavior.

A1.2 Moment Conditions for Brownfield Case

In greenfield settings where the base decision rule $\boldsymbol{\beta}_0 = \mathbf{0}$, it is possible to infer natural behavior from baseline behavior (e.g., equation (5)), prior to estimating costs. Our method can also be applied in *brownfield* settings, where a decision rule has already been implemented and baseline behavior may already be manipulated.

In such a setting, one can jointly estimate costs and the parameters describing baseline behavior ($\underline{\mathbf{x}}$ and $\boldsymbol{\mu}$) by appending two moment conditions based on $\mathbb{E}[\eta_{itk}] = 0$. For each individual i and behavior k, we have

$$\hat{\underline{x}}_{ik} = \frac{1}{|\mathbb{T}_i|} \sum_{t \in \mathbb{T}_i} \left[x_{ikt} - \mu_{kt} - f(\boldsymbol{\omega}, \mathbf{z}_i) \cdot [C^{-1}\boldsymbol{\beta}_{it}]_k \right] \tag{10}$$

For each time period t and behavior k we have

$$\hat{\mu}_{kt} = \frac{1}{|\{i | \mathbb{T}_i \ni t\}|} \sum_{i | \mathbb{T}_i \ni t} \left[x_{ikt} - \underline{x}_{ik} - f(\boldsymbol{\omega}, \mathbf{z}_i) \cdot [C^{-1}\boldsymbol{\beta}_{it}]_k \right] \tag{11}$$

Identification still requires observing random variation along each behavior in the decision rule (and ensuing manipulation).[35]

A1.3 One-Shot Estimation

In our experiment, we observed each individual over multiple time periods, which increased statistical power. However, our approach can also be applied if each individual i is observed in only one period t (for example, if loan applicants each apply for a loan once).

In a one-shot setting, C and $\boldsymbol{\omega}$ can be recovered by adjusting the brownfield moment conditions to remove both individual and time fixed effects. This entails replacing the moment condition in equation (8) with

$$\frac{1}{N} \sum_{i=1}^{N} \beta_{itk} \left[x_{ijt} - \chi_j - f(\boldsymbol{\omega}, \mathbf{z}_i) \cdot [C^{-1}\boldsymbol{\beta}_{it}]_j \right] = 0,$$

[35] This inversion will be more sensitive to the specification of the model than when unincentivized behavior can be observed directly in training.

Equation (9) with

$$\hat{v}_i = \frac{x_{ik_it} - \chi_{k_i}}{[C^{-1}\boldsymbol{\beta}_{it}]_{k_i}} - f(\boldsymbol{\omega}, \mathbf{z}_i),$$

Equation (10) with

$$\hat{\chi}_k = \frac{1}{N}\sum_{i=1}^{N}\left[x_{ikt} - f(\boldsymbol{\omega}, \mathbf{z}_i)\cdot[C^{-1}\boldsymbol{\beta}_{it}]_k\right],$$

and dropping equation (11), where natural behaviors are replaced with a term representing common behavior $\boldsymbol{\chi}$ of dimension K, and where k_i is the behavior incentivized for individual i.

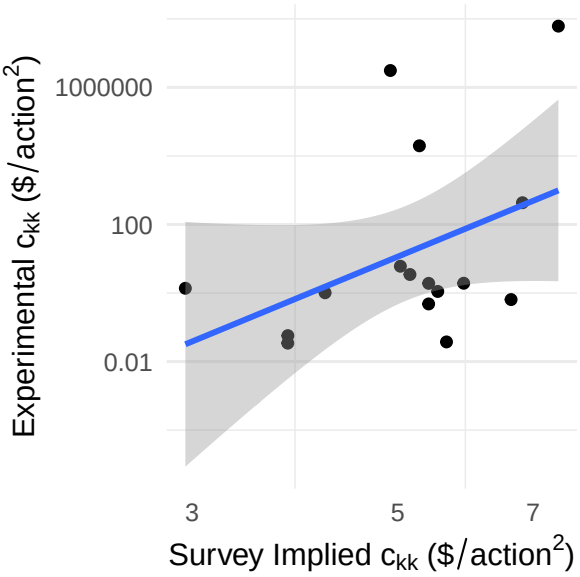
An estimate of each person's natural behavior can then be obtained by undoing any predicted manipulation:

$$\underline{\hat{x}}_{ik} = x_{ikt} - f(\hat{\boldsymbol{\omega}}, \mathbf{z}_i)\cdot[\hat{C}^{-1}\boldsymbol{\beta}_{it}]_k$$

though with just one observation this will be more affected by idiosyncratic noise.

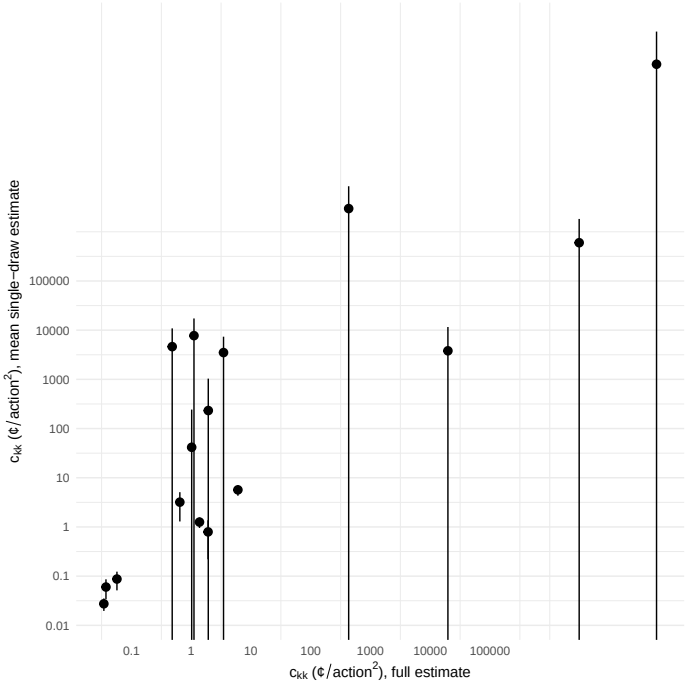
We demonstrate that this approach obtains similar results by mimicking the data that would result if our experiment had observed each individual for only one period. Because this will drastically reduce our sample size, we simulate this over multiple replication draws. For replication draw r, for each individual i we restrict the sample to include only one randomly selected incentivized week $t_{ir} \in \mathbb{T}_i^{treatment}$, and consider the average over replications $r \in \{1...R\}$. Figure A2 shows that the one-shot estimates are similar to the full sample estimates (we report the average $c_{kk} = \frac{1}{R}\sum_r c_{kkr}$); the Pearson correlation coefficient between the two measures is 0.9987. The corresponding estimate for $\omega = \frac{1}{R}\sum_r \omega_r = 0.22$ (standard deviation 0.98), with 30.7% of single-draw estimates less than or equal to the full-sample estimate of -0.083 and 69.2% of estimates greater than the full-sample estimate.

Figure A1: Costs Elicited from Hypothetical Questions and Costs Measured in Experiment



Notes: Each dot represents a behavior captured by the Sensing App. Y-axis indicates the cost of manipulating that behavior, estimated through our experiment (Table 4). X-axis indicates costs elicited from hypothetical questions, inferred as $c_{kk} = \frac{1}{N_{survey}} \sum_i \frac{\beta_k}{\max(0.001, \Delta_{kki})}$ for each i surveyed (see Online Appendix S2.3).

Figure A2: Manipulation Costs Estimated with only One Observation per Person



Notes: Our main estimates (with multiple observations per person) are shown on the x-axis. The average estimate obtained when each individual is observed only once is shown on the y-axis using the one-shot moment conditions in Appendix A1.3. The standard deviation across replications is shown as a whisker in either direction.

Table A1: Estimated Manipulation Costs for All Behaviors

Heterogeneity by Behavior (**C** diagonal; all incentivized behaviors)

▲	●	text you send	0.03
▲	●	text you receive	0.04
		text you send on a weekday	0.05
		each text message you send in the evening hours (after 6pm)	0.05
▲	●	text you send or receive in the evening (6pm−10pm)	0.06
▲	●	time you call someone	0.48
		call with a number not in your contacts during the workday (9am−5pm	0.59
	●	second of your shortest weekend call	0.64
		call with a number not in your contacts on the weekend	0.81
▲	●	each different person you text or are texted by	1.02
		person you text during the early morning hours (12am − 5am)	1.04
▲	●	time you receive a call	1.11
		each person you text in the evening (after 6pm)	1.38
		person you text on the weekday that you send the most texts	1.55
		call between 12am and 5am	1.67
		each different person you text or are texted by on the weekend	1.68
		text you send on the day that you send the most texts	1.68
		person you call during the workday	1.70
	●	call you make that's missed	1.91
▲	●	call with someone not in your contacts	1.93
		person you receive a text from during the evening (5pm − 10pm)	1.98
		text you send on the day you send the least texts	3.03
▲		text you receive on the day you receive the most texts	3.47
		if you keep it at 100%.	3.87
		person you text	5.45
		outgoing call on the day with the least outgoing calls	5.77
▲		person who texts you	6.04
		person who calls you during the workday	6.67
		each person you call for over 30 minutes	7.12
		person you text on the workday that you text the most people	7.57
		time you plug in your phone and it charges	9.39
		if the number of texts you send varies a lot weekday to weekday	28.15
		person you receive a text from during the early morning hours (12am − 5am)	35.06
		day you use a whatsapp−related app	35.44
		if the number of texts you send and receive varies a lot day to day	92.37
		missed call on the day you have the least missed calls	112.94
		tools−type app you use	194.41
		if the number of texts you receive varies a lot weekday to weekday	217.33
		5−minute−or−more call with a non−contact	426.58
		day you use the gmail app	431.59
		day you use the app	519.97
		if you use a whatsapp−related app.	594.14
		if you use a tools−type app.	1,055.35
		if you use any document− or report−related app.	1,194.08
		productivity−type app you use	1,383.08
		if you use the gmail app.	10,349.72
		if you use a productivity−type app.	17,087.82
▲	●	second of your average evening (6pm−10pm) call	19,761.46
		if you use any twitter−brand app.	20,775.27
		text you receive on the day you receive the least texts	25,618.26
		missed call	26,820.04
		text you receive on the evening (6pm−10pm) that you receive the least texts	39,063.42
		if you use the app	111,824.80
		second of your average workday call (monday−friday	346,246.83
		if you use the twitter app at least once	370,683.79
		call that lasts at least an hour	395,022.17
▲	●	second of your average call duration	3,108,632.19
		day you use any instagram−brand app	7,505,847.37
		day that you use a productivity−type app	8,462,705.30
		instagram−brand app used	28,377,875.40
		social−type app you use	61,264,134.39
		document or report−related app you use	75,097,239.06
		times the average day−to−day variation in your texts	1,282,497,178.36
		percentage point of your time that you spend within .5 kilometers of the busara office	184,130,822,787.78

Notes: Parameters estimated using GMM. In cost matrix, off diagonal elements c_{jk}; j$\neq k$ regularized to zero ($\Lambda_{offdiagonal} \to \infty$), diagonal elements regularized with $\Lambda_{diagonal} = 1.0$, set via 3-fold cross validation. ▲ : included in incentivized naïve LASSO decision rule, ● : included in incentivized strategy-robust (SR) decision rule.

Table A2: SR Decision Rules Based on Survey-Estimated Costs

	Costs (Actual)	*Costs* (From Survey)	**Income**			**Intelligence** (above median Ravens)		
			$\boldsymbol{\beta}^{LASSO_{final}}$	$\boldsymbol{\beta}^{SR_{final}}_{SurveyCost}$	$\boldsymbol{\beta}^{SR_{final}}$	$\boldsymbol{\beta}^{LASSO_{final}}$	$\boldsymbol{\beta}^{SR_{final}}_{SurveyCosts}$	$\boldsymbol{\beta}^{SR_{final}}$
Panel A: **Decision Rule**								
text_count_out	0.035	3.804	-0.499	-0.329	-0.093			
text_count_incoming	0.037	5.645	0.141	0.014		0.270	0.223	0.114
text_count_evening	0.057	3.805			-0.115			
call_count_out	0.480	5.4	0.657	0.591	0.501		-0.058	
call_count_outgoing_missed	1.914	5.4				-0.156		
calls_noncontacts	1.929	5.891				-0.547		-0.518
max_daily_texts_incoming	3.471	5.155						0.421
Intercept			296.342	305.309	303.456	489.686	483.529	487.049
$\lambda^{decision}$			759.296	759.296	759.296	1032.37	1032.37	1032.37
Panel B: **Prediction Error**				RMSE ($)			RMSE ($)	
Predicted Opaque			3.572	3.577	3.584	4.972	4.982	4.973
Predicted Transparent			3.876	3.644	3.591	4.988	4.989	4.975

Notes: Panel A reports the decision rules derived from naive LASSO and our strategy-robust model, as well as strategy-robust decision rules that use only control weeks and costs estimated from surveys. It also reports the costs associated with these behaviors. Panel B reports the predicted performance of these decision rules, based on the experimentally estimated model of behavior. $\beta^{LASSO_{final}}$ presented in this table differs slightly from the β^{LASSO} which was implemented. The regularization protocol was updated to select penalization closer to the boundary of 3 coefficients and the sample was changed to coincide with that used for the SR model (it includes only individuals with nonmissing tech skills, dropping approximately 1.5 percent of the sample). For survey costs, we infer heterogeneity in gaming ability using variation in participant responses (see Online Appendix S2).