Improving Information from Manipulable Data*

Alex Frankel[†] Navin Kartik[‡]

April 4, 2020

Abstract

Data-based decisionmaking must account for the manipulation of data by agents who are aware of how decisions are being made and want to affect their allocations. We study a framework in which, due to such manipulation, data becomes less informative when decisions depend more strongly on data. We formalize why and how a decisionmaker should commit to underutilizing data. Doing so attenuates information loss and thereby improves allocation accuracy.

JEL Classification: C72; D40; D82

Keywords: Gaming; Goodhart's Law; Strategic Classification

*We thank Ian Ball, Ralph Boleslavsky, Max Farrell, Pepe Montiel Olea, Canice Prendergast, Robert Topel, and conference audiences for helpful comments. Bruno Furtado and Suneil Parimoo provided excellent research assistance.

[†]University of Chicago Booth School of Business; afrankel@chicagobooth.edu.

[‡]Columbia University, Department of Economics; nkartik@columbia.edu.

1. Introduction

In various situations an agent receives an allocation based on some prediction about her characteristics, and the prediction relies on data generated by the agent's own behavior. Firms use a consumer's web browsing history for price discrimination or ad targeting; a prospective borrower's loan decision and interest rate depend on her credit score; and web search rankings take as input a web site's own text and metadata. In all these settings, agents who understand the prediction algorithm can alter their behavior to receive a more desirable allocation. Consumers can adjust browsing behavior to mimic those with low willingness to pay; borrowers can open or close accounts to improve their credit score; and web sites can perform search engine optimization to improve their rankings. How should a designer account for manipulation when setting the allocation rule?

First consider a naive designer who is unaware of the potential for manipulation. Before implementing an allocation rule, the designer gathers data generated by agents and estimates their types (the relevant characteristics). The *naive allocation rule* assigns each agent the allocation that is optimal according to this estimate. But after the rule is implemented, agents' behavior changes: if agents with "higher observables" x receive a "higher allocation" y under the allocation rule $Y(x)$, and if agents prefer higher allocations, then some agents will find ways to game the rule by increasing their x. In line with Goodhart's Law, the original estimation is no longer accurate.

A more sophisticated designer realizes that behavior has changed, gathers new data, and re-estimates the relationship between observables and type. After the designer updates the allocation rule based on the new prediction, agent behavior changes once again. The designer might iterate to a *fixed point*: an allocation rule that is a best response to the data that is generated under this very rule. But the resulting allocation need not match the desired agent characteristics well.

The question of this paper is how a designer with *commitment* power—a Stackelberg leader—should adjust a fixed-point allocation rule in order to improve the accuracy of the allocation. We find that a designer should make the allocation rule less sensitive to manipulable data than under the fixed point. In other words, the designer should "flatten" the allocation rule. Flattening the allocation results in ex-post suboptimality; the designer has committed to "underutilizing" agents' data. Fixed-point allocations, by contrast, are ex-post optimal. However, a flatter allocation rule reduces manipulation, which makes the data more informative about agents' types. Allocation accuracy improves on balance. We develop and explore this logic in what we believe is a compelling model of information loss due to manipulation.

By way of background, note that in some environments, manipulation does not lead to information loss: fixed-point rules deliver the designer's full-information outcome. To see this, think of a fixed-point rule as corresponding to the designer's equilibrium strategy in a signaling game in which the designer and agent best respond to each other. Under a standard single-crossing condition à la Spence (1973)—the designer wants to give more desirable allocations to agents with higher types, and higher types have lower marginal costs of taking higher observable actions—this signaling game has a fully separating equilibrium, i.e., one in which the designer perfectly matches the agent's allocation to her type. Even with commitment power, a designer cannot improve accuracy by departing from the corresponding allocation rule.

To introduce information loss, we build on a framework first presented by Prendergast and Topel (1996). The designer learns about an agent's type by observing data the agent generates, her action $x \in \mathbb{R}$. Agents are heterogeneous on two dimensions of their types, what we call *natural action* and *gaming ability*. The designer is only interested in the natural action $\eta \in \mathbb{R}$, which determines the agent's action x absent any manipulation. Gaming ability $\gamma \in \mathbb{R}$ summarizes how much an agent manipulates x in response to incentives. When drawing inferences from the action x, the designer's information about the agent's natural action η is "muddled" with that about gaming ability γ (Frankel and Kartik, 2019). We assume the designer observes x and chooses an allocation $y = Y(x) \in \mathbb{R}$ with the goal of minimizing the quadratic distance between y and η. We focus on linear allocation rules or policies $Y(x) = \beta x + \beta_0$, and we posit that agents adjust their observable x in proportion to $\gamma\beta$—their gaming ability times the sensitivity of allocations to observables. These linear functional forms arise in the linear-quadratic signaling models of Fischer and Verrecchia (2000) and Bénabou and Tirole (2006), among others.[1]

Our main result establishes that the commitment optimal policy is less sensitive to observables than is the fixed-point policy. (For this introduction, suppose there is a unique fixed point unless indicated otherwise.) Mathematically, for policies $Y(x) = \beta x + \beta_0$, it is optimal for the designer to depart from the fixed point with $\beta > 0$ by attenuating that coefficient towards zero. Information is underutilized in the sense that, given the data generated

[1] Subsection 2.1 microfounds such agent behavior. Subsection 4.2 discusses optimality of linear allocation rules. Note that, following a common abuse of terminology, we say "linear" instead of the mathematically more precise "affine".

by agents in response to this optimal policy, the designer would ex-post benefit from using a higher β. For instance, suppose the sensitivity of the naive policy is $\beta = 1$: when the designer does not condition the allocation on observables, the linear regression coefficient of type η on observable x is 1, and the naive designer responds by matching her allocation rule's sensitivity to this regression coefficient. The fixed-point policy may have $\beta = 0.7$. That is, when the designer sets $\beta = 0.7$ and runs a linear regression of η on x using data generated by the agent in response to $\beta = 0.7$, the regression coefficient is the same 0.7. Our result is that the optimal policy has $\beta \in (0, 0.7)$, say $\beta = 0.6$. After the designer sets $\beta = 0.6$, however, the corresponding linear regression coefficient is larger than 0.6, say 0.75. We emphasize that our argument for shrinking regression coefficients is driven by the informational benefit from reduced manipulation, and in turn, the resulting improvement in allocations. It is orthogonal to concerns about model overfitting.

In comparing our commitment solution with the fixed-point benchmark, it is helpful to keep in mind two distinct interpretations of the fixed point. The first concerns a designer who has market power in the sense that agents adjust their manipulation behavior in response to this designer's policies. Think of web sites engaging in search engine optimization to specifically improve their Google rankings; third party sellers paying for fake reviews on the Amazon platform; or citizens trying to game an eligibility rule for a targeted government policy. In these cases the designer may settle on a fixed point by iterating policies until reaching an ex post optimum. Our paper highlights that this fixed point may yet be suboptimal ex ante, and offers the prescriptive advice of flattening the allocation rule.

A second perspective is that the fixed-point policy represents the outcome of a competitive market. With many banks, any one bank that uses credit information in an ex-post suboptimal manner will simply be putting itself at a disadvantage to its competitors; similarly for colleges using SAT scores for admissions. So the fixed point becomes a descriptive prediction of the market outcome, i.e., the equilibrium of a signaling game. In that case, our optimal policy suggests a government intervention to improve allocations, or a direction that collusion might take.

Before turning to the related literature, we stress two points about our approach. First, our paper aims to formalize a precise but ultimately qualitative point, and make salient its logic. Our model is deliberately stylized and, we believe, broadly relevant for many applications. But it is not intended to capture the details on any specific one. We hope that it will be useful for particular applications either as a building block or even simply as a benchmark for thinking about positive and normative implications. Second, we view our main result—the commitment policy flattens fixed points and underutilizes data—as intuitive once one understands the logic of our environment. Indeed, there is a simple first-order gain vs. second-order loss intuition for a local improvement from flattening a fixed point; see the discussion after Proposition 1. Confirming that the result holds for the global optimum is not straightforward, however; among other things, there can be multiple fixed points.

Related Literature. There are many settings in economics in which a designer commits to making ex-post suboptimal allocations in order to improve ex-ante incentives on some dimension. Our specific interest in this paper is in a canonical problem of matching allocations to unobservables in the presence of strategic manipulation. In this context, we study a simple model in which there is a benefit of committing to distortions in order to improve the ex-ante accuracy of the allocations.

Building on intuitions from Prendergast and Topel (1996), Fischer and Verrecchia (2000), and Bénabou and Tirole (2006), Frankel and Kartik (2019) elucidate general conditions under which an agent's action becomes less informative to an observer when the agent has stronger incentives to manipulate. None of these papers model the allocation-accuracy problem we study here; the latter three papers do not study commitment either. Notwithstanding, our designer faces the following tradeoff suggested by the intuitions in those papers: making allocations more responsive to an agent's data amplifies the agent's manipulation, which makes the data less informative, reducing the optimal responsiveness for allocation accuracy.

At a very broad level, our main result that the designer should flatten allocations relative to the fixed-point rule is reminiscent of the "downward distortion" of allocations in screening problems following Mussa and Rosen (1978). That said, our framework, analysis, and emphasis—on manipulation and information loss, allocation accuracy, contrasting commitment with fixed points—are not readily comparable with that literature. One recent paper on screening we highlight is Bonatti and Cisternas (2019). In a dynamic price discrimination problem, they show that short-lived firms get better information about long-lived consumers' types—resulting in higher steady-state profits—if a designer reveals a statistic that underweights recent consumer behavior. Suitable underweighting dampens consumer incentives to manipulate demand.

A finance literature addresses the difficulty of using market activity to learn fundamentals when participants have manipulation incentives. Again in models very different from ours, some papers highlight benefits of committing to underutilizing information. See, for example, Bond and Goldstein (2015) and Boleslavsky, Kelly, and Taylor (2017). These authors study trading in the shadow of a policymaker who may intervene after observing prices or order flows. The anticipation of intervention makes the financial market less informative about a fundamental to which the intervention should be tailored. Both papers establish that the policymaker may benefit from a commitment that, in some sense, entails underutilization of information. In particular, Bond and Goldstein (2015, Proposition 2) highlight a local first-order information benefit vs. second-order allocation loss akin to our Lemma 1. Unlike us, they do not study global optimality.

A number of papers in economics study the design of testing regimes and other instruments to improve information extraction. Recent examples include Harbaugh and Rasmusen (2018) on pooling test outcomes to improve voluntary participation, Perez-Richet and Skreta (2018) on the benefits of noisy tests when agents can manipulate the test, and Martinez-Gorricho and Oyarzun (2019) on using "conservative" (or "confirmatory") thresholds to mitigate manipulation. Jann and Schottmüller (2018), Ali and Bénabou (2019), and Frankel and Kartik (2019) analyze how hiding information about agents' actions—increasing privacy—can improve information about their characteristics.[2]

Beyond economics, our paper connects to a recent computer science literature studying classification algorithms in the presence of strategic manipulation. See, among others, Hardt, Megiddo, Papadimitriou, and Wootters (2016), Hu, Immorlica, and Vaughan (2019), Milli, Miller, Dragan, and Hardt (2018), and Kleinberg and Raghavan (2019). In a binary strategic classification problem, Braverman and Garg (2019) argue for random allocations to improve allocation accuracy and reduce manipulation costs.

We would like to reiterate that our designer is only interested in allocation accuracy, not directly the costs of manipulation. Moreover, unlike Kleinberg and Raghavan (2019), we model an agent's manipulation effort as pure "gaming": it does not provide desirable output or affect the designer's preferred allocation. By contrast to us, principal-agent problems in economics often focus on how allocation rules interact with incentives for desirable effort. For instance, Prendergast and Topel (1996) study contracts in which incentivizing worker effort provides a firm worse information about the worker's match quality because of an intermediary's favoritism. In a multitasking environment, Ederer, Holden, and Meyer (2018)

[2] Eliaz and Spiegler (2019) explore the distinct issue of an agent's incentives to reveal her own data to a "non-Bayesian statistician" making predictions about her.

study how randomized rewards schemes can reduce gaming and improve effort. Liang and Madsen (2020) show that a principal might strengthen an agent's effort incentives by committing to disregard predictive data acquired from other agents; the benefit can dominate the cost of making less accurate predictions.

Finally, we mention the contemporaneous work of Ball (2020). He extends the linear-quadratic-elliptical specification of Frankel and Kartik (2019) to incorporate multiple "features" or dimensions; on each feature, agents have heterogeneous natural actions and gaming abilities. His main focus is on optimal scoring rules to improve information, specifically in identifying how to weight the different features when aggregating them into a one-dimensional statistic.[3] He also compares his analog of our commitment solution with both his scoring and fixed-point solutions. Similar to us, he finds that under certain conditions, his commitment solution is less responsive to all of an agent's features than the (unique, under his assumptions) fixed-point solution.

2. Model

An agent has a type $(\eta, \gamma) \in \mathbb{R}^2$ drawn from joint distribution F. It may be helpful to remember the mnemonics η for *n*atural action, and γ for *g*aming ability; see Subsection 2.1. Assume the variances $\mathrm{Var}(\eta) = \sigma_\eta^2$ and $\mathrm{Var}(\gamma) = \sigma_\gamma^2$ are positive and finite.[4] Denote the means of η and γ by μ_η, μ_γ, and assume their correlation is $\rho \in (-1, 1)$, with $\rho = \mathrm{Cov}(\eta, \gamma)/(\sigma_\eta \sigma_\gamma)$.

A designer seeks to match an allocation $y \in \mathbb{R}$ to η, with a quadratic loss of $(y - \eta)^2$. The designer chooses $y = Y(x)$ as a function of an observed action $x \in \mathbb{R}$ that is chosen by the agent. Thus, the designer's welfare loss is

$$\text{Welfare Loss} \equiv \mathbb{E}[(Y(x) - \eta)^2]. \tag{1}$$

The agent chooses x as a function of her type (η, γ) after observing the allocation rule Y. In a manner detailed later, the agent will have an incentive to choose a higher x to obtain a higher y. Given a strategy of the agent, the designer can compute the distribution of x and

[3] He interprets the aggregator as produced by an intermediary who shares the decisionmaker's interests, but cannot control the decisionmaker's behavior. That is, the intermediary can commit to the aggregation rule but allocations are made optimally given the aggregation.

[4] Throughout, we use 'positive' to mean 'strictly positive', and similarly for 'negative', 'larger', and 'smaller'.

the value of $\mathbb{E}[\eta|x]$ for any x the agent may choose. A standard decomposition[5] is

$$\text{Welfare Loss} = \underbrace{\mathbb{E}[(\mathbb{E}[\eta|x] - \eta)^2]}_{\text{Info loss from estimating } \eta \text{ using } x} + \underbrace{\mathbb{E}[(Y(x) - \mathbb{E}[\eta|x])^2]}_{\text{Misallocation loss given estimation}}. \tag{2}$$

Holding fixed the agent's strategy, it is "ex-post optimal" for the designer to set $Y(x) = \mathbb{E}[\eta|x]$. However, the agent's strategy responds to Y. So the designer may prefer to use an ex-post suboptimal allocation rule to improve her estimation of η from x, as seen in the first term of (2). That is, the designer may benefit from the power to commit to her allocation rule.

2.1. Linearity Assumptions

Assume the designer chooses among linear allocation rules: the designer chooses policy parameters $(\beta, \beta_0) \in \mathbb{R}^2$ such that

$$Y(x) = \beta x + \beta_0. \tag{3}$$

Also assume that, given the designer's policy (β, β_0), the agent chooses x using a linear strategy $X_\beta(\eta, \gamma)$ that takes the form

$$X_\beta(\eta, \gamma) = \eta + m\beta\gamma \tag{4}$$

for some exogenous parameter $m > 0$. Thus η is the agent's "natural action": the action taken when the designer's policy does not depend on x (i.e., $\beta = 0$). The variable γ represents idiosyncratic responsiveness to the designer's policy: a higher γ increases the agent's action from the natural level by more for any $\beta > 0$. The parameter m captures a common component of responsiveness across all agent types.

The strategy in Equation 4 can be motivated as the best response for an agent who

[5] The right-hand sides of (1) and (2) are equal if

$$\mathbb{E}[(Y(x))^2 - 2\eta Y(x) + \eta^2] = \mathbb{E}\big[\eta^2 - 2\eta\mathbb{E}[\eta|x] + (\mathbb{E}[\eta|x])^2 + (Y(x))^2 - 2Y(x)\mathbb{E}[\eta|x] + (\mathbb{E}[\eta|x])^2\big].$$

Canceling out like terms and rearranging, it suffices to show that

$$2\mathbb{E}\big[(\mathbb{E}[\eta|x] - \eta)Y(x)\big] = 2\mathbb{E}\big[(\mathbb{E}[\eta|x] - \eta)\mathbb{E}[\eta|x]\big].$$

This equality holds by the orthogonality condition $\mathbb{E}[(\mathbb{E}[\eta|x] - \eta)g(x)] = 0$ for all functions $g(x)$.

maximizes a utility of

$$m\gamma y - (x-\eta)^2/2.$$

Here m captures the "stakes" that agents face to obtain higher y, and γ is an idiosyncratic marginal benefit. Alternatively, the strategy is also optimal for an agent with $\gamma > 0$ who maximizes

$$y - (x-\eta)^2/(2m\gamma).$$

Here m parameterizes the "manipulability" of the action x, and γ is an agent's idiosyncratic "gaming ability".

2.2. The Designer's Problem

The designer commits to her policy (β, β_0), which the agent observes and responds to according to (4). Plugging the rule (3) and the strategy (4) into the welfare loss function (1) yields

$$\text{Welfare Loss} = \mathbb{E}[(\beta(\eta + m\beta\gamma) + \beta_0 - \eta)^2].$$

The designer's problem is therefore to choose (β, β_0) to minimize the above loss function, which is quartic in β.[6] We denote the solution as (β^*, β_0^*).

2.3. Discussion

Given the asymmetry between the characteristics η and γ in the agent's strategy (4), it is crucial for our results that the designer seeks to match η rather than γ. The reason is that when the designer's policy puts more weight on the data—when β increases—the agent's action x becomes less informative about η but more informative about γ; Remark 1 below makes this point precise.

It is, on the other hand, straightforward to generalize our analysis to an allocation matching some other characteristic of the agent, τ, that is correlated with η. The assumption we would require is that $\mathbb{E}[\tau|\eta,\gamma]$ is independent of γ and linear in η. The welfare loss $\mathbb{E}[(Y(x)-\tau)^2]$ could then be decomposed as $\mathbb{E}[(Y(x)-\mathbb{E}[\tau|\eta])^2] + \mathbb{E}[(\mathbb{E}[\tau|\eta]-\tau)^2]$. As the

[6] Using standard mean-variance decompositions,

$$\text{Welfare Loss} = (1-\beta)^2\sigma_\eta^2 + m^2\beta^4\sigma_\gamma^2 - 2(1-\beta)m\beta^2\rho\sigma_\eta\sigma_\gamma + (\beta_0 - (1-\beta)\mu_\eta + m\beta^2\mu_\gamma)^2.$$

second term here—the information loss from predicting τ using η—is independent of the allocation rule $Y(x)$, it would not affect the designer's choice. The designer would then effectively be trying to match the allocation to a linear function of η.

2.4. Preliminaries

2.4.1. Linear regression of type η on action x

When the designer uses policy (β, β_0), the agent responds with strategy $X_\beta(\eta, \gamma) = \eta + m\beta\gamma$. Suppose the designer were to gather data under this agent behavior and then estimate the relationship between the dimension of interest η and the action x. Specifically, let $\hat{\eta}_\beta(x)$ denote the best linear estimator of η from x under a quadratic loss objective:

$$\hat{\eta}_\beta(x) \equiv \hat{\beta}(\beta)x + \hat{\beta}_0(\beta),$$

with $\hat{\beta}$ and $\hat{\beta}_0$ the coefficients of an ordinary least squares (OLS) regression of η on x.

Following standard results for OLS,

$$\hat{\beta}(\beta) = \frac{\sigma_\eta^2 + m\rho\sigma_\eta\sigma_\gamma\beta}{\sigma_\eta^2 + m^2\sigma_\gamma^2\beta^2 + 2m\rho\sigma_\eta\sigma_\gamma\beta}, \tag{5}$$

where the right-hand side's numerator is the covariance of x and η given the strategy X_β, and its denominator is the variance of x (which is positive because $\rho \in (-1, 1)$). Correspondingly,

$$\hat{\beta}_0(\beta) = \mu_\eta - \hat{\beta}(\beta)[\mu_\eta + m\beta\mu_\gamma].$$

It is useful to further rewrite the welfare loss (2) as follows, for any policy (β, β_0) defining the linear allocation rule $Y(x) = \beta x + \beta_0$:[7]

$$\text{Welfare Loss} = \underbrace{\mathbb{E}[(\hat{\eta}_\beta(x) - \eta)^2]}_{\text{Info loss from linearly estimating } \eta \text{ using } x} + \underbrace{\mathbb{E}[(Y(x) - \hat{\eta}_\beta(x))^2]}_{\text{Misallocation loss given linear estimation}}. \tag{6}$$

Some readers may find it helpful to note that information loss from estimation (the first term in (6)) is the variance of the residuals in an OLS regression of η on x; put differently, $\mathbb{E}[(\hat{\eta}_\beta(x) - \eta)^2] = \sigma_\eta^2(1 - R_{x\eta}^2)$, with $R_{x\eta}^2$ the coefficient of determination in that regression.

[7] This derivation is identical to that in fn. 5, only replacing $\mathbb{E}[\eta|x]$ by $\hat{\eta}_\beta(x)$ and applying the orthogonality condition $\mathbb{E}[(\hat{\eta}_\beta(x) - \eta)g(x)] = 0$ for all affine functions $g(x)$.

We stress that Equation 6 is simply a convenient decomposition; given our focus on linear allocation rules, using OLS entails no restrictions.

Remark 1. For $\rho \geq 0$, $\hat{\beta}(\beta)$ is decreasing on $\beta \geq 0$. To see why, notice that when β increases the agent's action x depends more on the variable γ. This increases $\mathrm{Var}(x)$ and, when $\rho \geq 0$, also provides the designer with less information about the variable η that she is trying to estimate from x.[8] Both effects lead to a lower $\hat{\beta}$. By contrast, if the designer were trying to estimate γ rather than η (minimizing $\mathbb{E}[(y-\gamma)^2]$ rather than $\mathbb{E}[(y-\eta)^2]$), then for $\rho \geq 0$, the analogous regression coefficient of γ on x need not be deceasing on $\beta \geq 0$.

2.4.2. Benchmark policies

Constant. A rule that does not condition the allocation on the observable corresponds to a constant policy (β, β_0) with $\beta = 0$. A constant policy gives rise to a welfare loss of $\sigma_\eta^2 + (\beta_0 - \mu_\eta)^2$. In the decomposition of Equation 6, the entire welfare loss is due to misallocation; the information loss from estimation is zero because the agent's behavior $x = \eta$ fully reveals the natural action η. Under the constant policy the linear estimator $\hat{\eta}_0$ has coefficients $\hat{\beta}(0) = 1$ and $\hat{\beta}_0(0) = 0$.

Naive. If the designer uses a constant policy (β, β_0) with $\beta = 0$, the agent responds with $X_0(\eta, \gamma) = \eta$. Suppose the designer gathers data produced from such behavior, and—failing to account for manipulation—expects the agent to maintain this strategy regardless of the policy. Then the designer would (incorrectly) perceive her optimal policy to be $(\beta^{\mathrm{n}}, \beta_0^{\mathrm{n}}) \equiv (\hat{\beta}(0), \hat{\beta}_0(0)) = (1, 0)$.

Designer's best response. More generally, suppose the designer expects the agent to use the strategy $X_\beta(\eta, \gamma) = \eta + m\beta\gamma$ regardless of its policy. The designer would find it optimal in response to set an allocation rule $Y(x)$ equal to the best linear estimator of η from x, i.e., a policy $(\hat{\beta}(\beta), \hat{\beta}_0(\beta))$ yielding $Y(x) = \hat{\eta}_\beta(x)$.

Fixed point. We say that a policy $(\beta^{\mathrm{fp}}, \beta_0^{\mathrm{fp}})$ is a *fixed point* if

$$\beta^{\mathrm{fp}} = \hat{\beta}(\beta^{\mathrm{fp}}) \quad \text{and} \quad \beta_0^{\mathrm{fp}} = \hat{\beta}_0(\beta^{\mathrm{fp}}).$$

[8] Less information is not generally in the Blackwell (1951) sense unless the prior on (η, γ) is bivariate normal. Rather, it is in the sense of a higher information loss from linearly estimating η using x: $\mathbb{E}[(\hat{\eta}_\beta(x) - \eta)^2]$ is increasing in β.

A fixed point corresponds to a Nash equilibrium of a game in which the designer's policy is set simultaneously with the agent's strategy. That is, instead of the designer committing to a policy (the Stackelberg solution), the policy is a best response to the agent's strategy that the policy induces. In the decomposition of Equation 6, a fixed-point policy may have a positive information loss from estimation, but it has zero misallocation loss—the designer is choosing the optimal policy given the information generated by the agent.

Figure 1 illustrates some designer best response functions and fixed points.

3. Analysis

3.1. Main Result

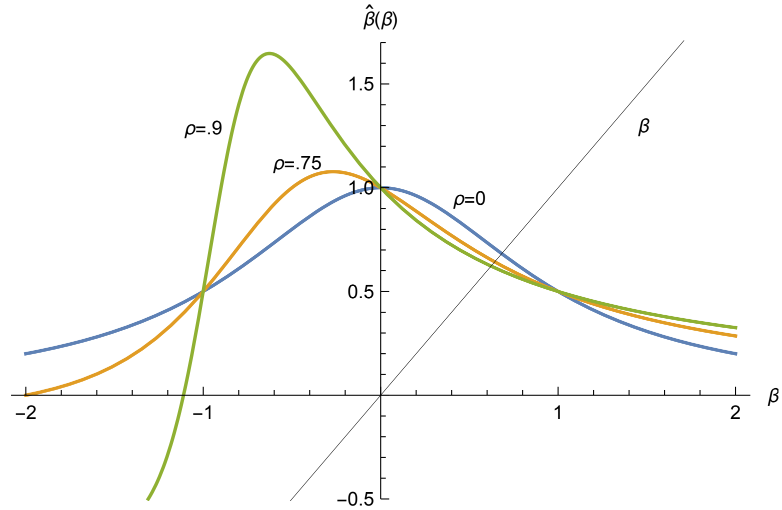
We seek to compare the designer's optimal policy (β^*, β_0^*) with the fixed points $(\beta^{\text{fp}}, \beta_0^{\text{fp}})$. There can, in general, be multiple fixed points, but there is always at least one with a positive sensitivity or weight on the agent's action, i.e., $\beta^{\text{fp}} > 0$. Moreover, when there is nonnegative correlation in the agent's characteristics ($\rho \geq 0$), there is only one nonnegative fixed point, and it satisfies $\beta^{\text{fp}} \in (0, 1)$. See Proposition C.1 in the Supplementary Appendix.

Take any fixed-point sensitivity $\beta^{\text{fp}} > 0$. Our main result is that the optimal policy puts less weight on the agent's action than does the fixed point. Further, the optimal policy underutilizes information by putting less weight on the agent's action than does the OLS coefficient (and hence the best linear policy) given the data generated by the agent in response.

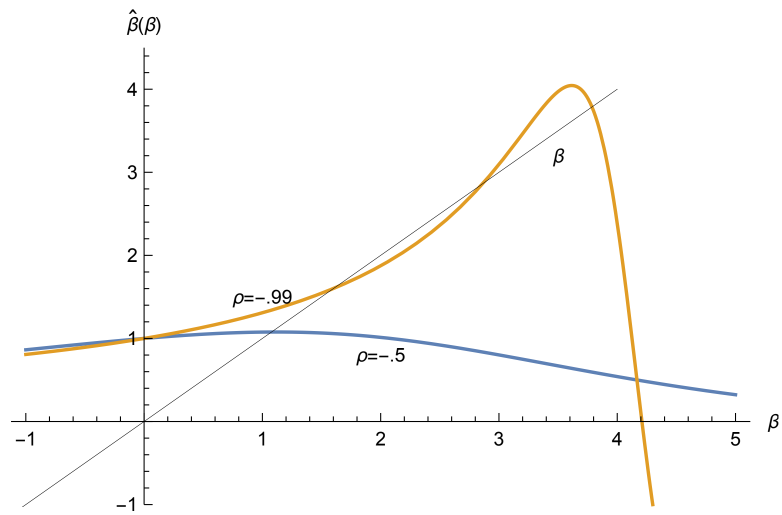
Proposition 1. *There is a unique optimum,* (β^*, β_0^*)*. It has* $\beta^* > 0$ *and* $\beta^* < \beta^{\text{fp}}$ *for any fixed point* $\beta^{\text{fp}} > 0$*. Moreover,* $\hat{\beta}(\beta^*) > \beta^*$*.*

For a concrete example, take $m = \sigma_\eta^2 = \sigma_\gamma^2 = 1$ and $\rho = 0$. Recall that the sensitivity of the naive policy is (normalized to) $\beta = 1$. The unique fixed-point policy has $\beta^{\text{fp}} \approx 0.68$. The optimal policy reduces the sensitivity to $\beta^* \approx 0.59$. Given the agent's behavior under this policy, the designer would ex post prefer the higher value $\hat{\beta}(\beta^*) \approx 0.74$.

Here is the intuition for the proposition, as illustrated graphically in Figure 2. Consider a designer choosing $\beta = \beta^{\text{fp}} > 0$. This designer's policy is ex-post optimal in the sense that misallocation loss (the second term in the welfare decomposition (6)) given the information the designer obtains about η is minimized at zero. Adjusting the sensitivity β in either direction from β^{fp} increases misallocation loss, but this harm is second order because we are



(a) Parameters: $\sigma_\eta = \sigma_\gamma = 1$ and $m = 1$.



(b) Parameters: $\sigma_\eta = \sigma_\gamma = 1$ and $m = 0.24$.

Figure 1 – The best response function $\hat{\beta}$. As shown in Figure 1a, $\hat{\beta}$ is decreasing on $[0, \infty)$ when $\rho \geq 0$. Figure 1b illustrates that this need not be true when $\rho < 0$. In all cases, intersections of $\hat{\beta}$ with β correspond to fixed points β^{fp}.

starting from a minimum. By contrast, at $\beta = \beta^{\text{fp}}$ there is positive information loss from estimation (the first term in (6)) because the agent's action does not reveal η. Lowering β reduces information loss from estimation, which yields a first-order benefit. (The first-order benefit was suggested by Remark 1 for $\rho \geq 0$, and the point is general.) Hence, there is a net first-order welfare benefit of lowering β from β^{fp}. Of course, the designer wouldn't lower β down to 0, since making some use of the information from data is better than not using it at all.[9]

The proof of Proposition 1 in Appendix A establishes uniqueness of the global optimum, rules out that it is negative, and shows that it is less than every fixed point $\beta^{\text{fp}} > 0$. To formalize a key step—the aforementioned first-order benefit of reducing β from any β^{fp}—let $\mathcal{L}(\beta)$ be the welfare loss from policy β (paired with the correspondingly optimal β_0), with derivative $\mathcal{L}'(\beta)$.

Lemma 1. *For any β^{fp}, it holds that $\mathcal{L}'(\beta^{\text{fp}}) > 0$.*

Note that Lemma 1 also applies to negative values of β^{fp} when those exist.

Remark 2. The welfare gains from commitment can be substantial. For suitable parameters, the unique fixed point's welfare is arbitrarily close to that of the best constant policy $Y(x) = \mu_\eta$, while the optimal policy's welfare is arbitrarily close to the first best's.[10]

3.2. Comparative Statics

We provide a few comparative statics below. In taking comparative statics, it is helpful to observe that the designer's best response $\hat{\beta}(\beta)$ defined in Equation 5 depends on parameters m, σ_η, and σ_γ only through the statistic $k \equiv m\sigma_\gamma/\sigma_\eta$, as does the welfare loss $\mathcal{L}(\beta)$ divided by σ_η^2 (see Equation A.1 in Appendix A.1). Therefore, the optimal and fixed-point values β^* and β^{fp} also only depend on these parameters through k. The parameter k summarizes the susceptibility of the allocation problem to manipulation: higher k (arising from higher stakes or manipulability m of the mechanism, greater variance in gaming ability σ_γ^2, or lower variance in natural actions σ_η^2) means that under any fixed policy, agents as a whole adjust their observable action x further from their natural action η, relative to the spread of

[9] Indeed, any fixed-point policy itself does better than the best constant policy $(\beta, \beta_0) = (0, \mu_\eta)$. Note, however, that this constant policy can be better than the naive policy $(\beta^{\text{n}}, \beta_0^{\text{n}}) = (1, 0)$.

[10] The parameters are such that $m\sigma_\gamma/\sigma_\eta \to 1/4^+$ and $\rho \to -1$. Both the first-best welfare and the constant policy's are independent of ρ; the former is 0 (by normalization) while the latter is $-\sigma_\eta^2$, which can be arbitrarily low.

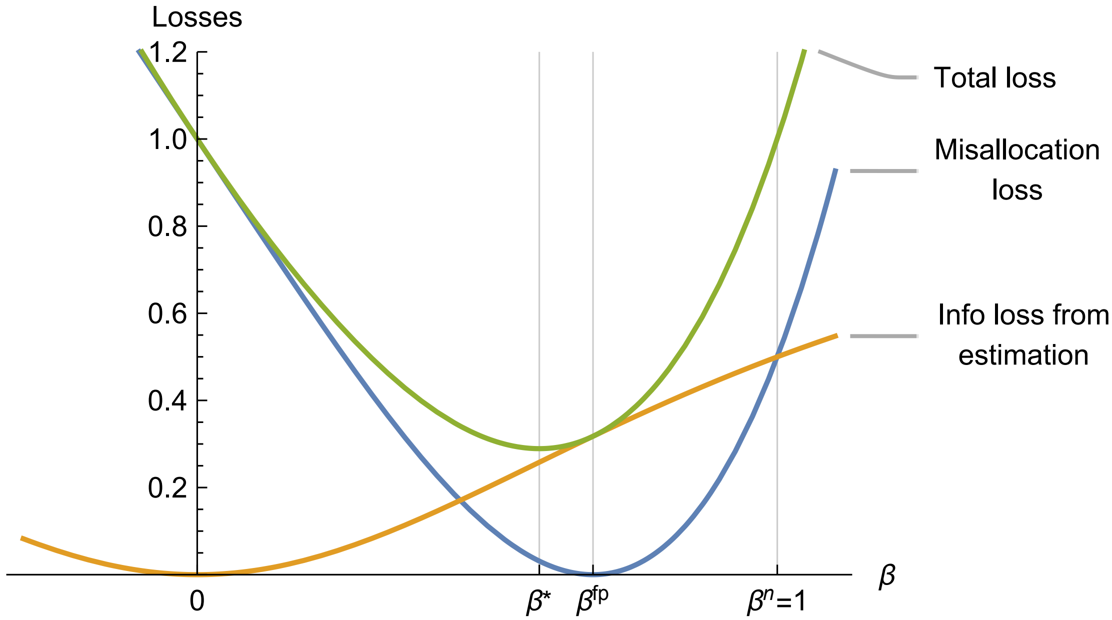


Figure 2 – The welfare loss decomposition from Equation 6 for policy (β_0, β), with the optimal β_0 plugged in for each β on the horizontal axis. Parameters: $\sigma_\eta = \sigma_\gamma = m = 1$ and $\rho = 0$. Numerical solutions: $\beta^* = 0.590$ and $\beta^{\text{fp}} = 0.682$.

observables prior to manipulation. Hence, for comparative statics over model primitives, it is sufficient to consider only the statistic k and the correlation parameter ρ.

Proposition 2. *For $k \equiv m\sigma_\gamma/\sigma_\eta$, the following comparative statics hold.*

1. *As $k \to \infty$, $\beta^* \to 0$; as $k \to 0$, $\beta^* \to 1$. If $\rho \geq 0$, then β^* is strictly decreasing in k; if $\rho < 0$, then β^* is strictly quasi-concave in k, attaining a maximum at some point.*
2. *β^* is strictly increasing in ρ when $k > 3/4$, strictly decreasing in ρ when $k < 3/4$, and independent of ρ when $k = 3/4$.*
3. *When $\rho = 0$, $\beta^*/\beta^{\text{fp}}$ is strictly decreasing in k, approaching $\sqrt[3]{1/2} \approx 0.79$ as $k \to \infty$ and 1 as $k \to 0$.*

Part 1 of the proposition implies that when agents' characteristics are nonnegatively correlated, a designer faced with a more manipulable environment should put less weight on the agents' observable action. While such monotonicity is intuitive, it does not hold when there is negative correlation. Similarly, one might expect greater positive correlation to increase the optimum β^*; indeed, Frankel and Kartik (2019, Proposition 4) establish that it does have this effect on the (unique) positive fixed point $\beta^{\text{fp}} > 0$. But we see in part 2 of Proposition 2 that this holds for β^* only when the susceptibility-to-manipulation statistic k is large enough. Finally, part 3 implies that when the characteristics are uncorrelated, the ratio $\beta^{\text{fp}}/\beta^*$ decreases as the statistic k increases. As $k \to 0$, the fixed point fully reveals an agent's natural action ($\beta^{\text{fp}} \to 1$) and so the designer does not benefit from commitment power: the fixed point is optimal as it provides the minimum possible welfare loss. As $k \to \infty$, both β^* and β^{fp} tend to zero yet the ratio $\beta^*/\beta^{\text{fp}}$ stays bounded.

4. Discussion

4.1. Additional Designer Objectives

We have argued that the designer should make allocations less sensitive to data than in fixed points in order to improve information. Of course, information might only affect part of the designer's welfare. Recall that agents shift their action by $m\beta\gamma$ away from their natural action. If the designer seeks to induce higher actions—because this corresponds to socially valuable "effort", say—then the designer will want to increase β above β^*. If the designer instead wants to reduce signaling costs, he will further weaken manipulation incentives by attenuating β from β^* towards zero. Both shifts will harm allocation accuracy, the former by increasing the information loss from estimation and the latter by increasing misallocation loss given estimation.

Another possibility is that the designer seeks to match the allocation to an agent's gaming ability γ instead of her natural action η, perhaps because gaming ability is a skill the designer values. In that case one would expect the sensitivity to data under commitment to be larger than in fixed points; intuitively, increasing manipulation incentives provides less information about η but more information about γ (Frankel and Kartik, 2019).

4.2. Nonlinear Policies

That the designer uses linear allocation rules is generally restrictive. Gesche (2019) and Frankel and Kartik (2019) have shown that fixing any linear strategy for the agent, the designer's best response is linear if the agent's type distribution is bivariate elliptical (Gómez, Gómez-Villegas, and Marín, 2003), subsuming bivariate normal; see also Fischer and Verrecchia (2000) and Bénabou and Tirole (2006). Hence, under these joint distributions—and when agents optimally respond to linear allocation rules with linear strategies (see Subsection 2.1)—the linear fixed-point policies of the current paper correspond to equilibria of a signaling game. Ball (2020) extends these results to a multidimensional action space. A not implausible conjecture is that elliptical distributions also ensure optimality of linear allocation rules when the designer can commit.

4.3. Alternative Models of Information loss

We have developed our main point about flattening allocation rules in what we believe is a canonical model of information loss from manipulation, one used in a number of aforementioned papers. Information loss in this model stems from agents' heterogeneous responses to incentives. But we believe the underlying logic generalizes to other sources of information loss. For instance, even a model with a one-dimensional type (e.g., no heterogeneity on gaming ability γ) may have information loss from "pooling at the top" in a bounded action space. Appendix D establishes a version of our result for a simple model in that vein.

4.4. General Allocation Problems

We conclude by sketching a proposal—estimation with noise—for attenuating the impact of manipulable data in more general allocation problems. Consider an environment in which a designer estimates agent characteristic η from the observation of some x, then assigns allocation y based on both x and the estimate of η. The variable x need not be scalar. As such, the allocation rule need not have any easily interpreted coefficient measuring how "flat" or "steep" it is with respect to manipulable components of x.

To formalize our proposal, let a *data set* be a joint distribution over (x, η). Let ML be an *estimation procedure* (e.g., a machine learning algorithm) that takes as input an observable x and a data set d, and then outputs an allocation y. We interpret $ML(x; d)$ as first estimating η from x after being fit to the training data d, and then outputting the designer's preferred allocation given x and the estimate of η.

Estimation with noise. Recall the classical econometric result that measurement error on an independent variable leads to attenuation bias, i.e., to an estimated coefficient in a linear regression that is biased towards zero. Applying this concept, here is one approach for generating the optimal policy of Sections 2–3. First gather training data set $\tilde{d}$ from some linear policy $\tilde{Y}(x) = \beta_0 + \beta x$, where we take the coefficient β such that we expect the best response $\hat{\beta}(\beta)$ to be above β^*. Then add noise to the measurements of x in the data set $\tilde{d}$ to generate a new data set d'. For instance, replace each data point (x_i, η_i) in $\tilde{d}$ with data point (x'_i, η_i) in d', where the new regressor x' is defined as $x'_i = x_i + c + \varepsilon_i$ for $c \in \mathbb{R}$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. When we linearly regress η on x' in the data set d', attenuation bias establishes that we find a smaller coefficient than $\hat{\beta}(\beta)$: increasing the variance of the noise σ_ε^2 from 0 to infinity reduces the estimated coefficient of η on x' from $\hat{\beta}(\beta)$ to 0. For an appropriate level of noise, we hit the optimal coefficient β^*. Finally the constant c, added to or subtracted from all points x', can be adjusted so that the average allocation is equal to μ_η and thus the constant term in the regression is optimal.

We can generalize this estimation with noise to arbitrary estimation procedures on arbitrary data sets. Start with the training data set $\tilde{d}$ induced by some original policy $\tilde{Y}$. To generate the new data set d', add noise—perhaps with nonzero mean—to any manipulable components of x to get x', while keeping η unchanged. Now define the estimation with noise policy Y^{ews} as

$$Y^{\text{ews}}(x) = ML(x; d').$$

Crucially, when determining the allocation for an agent with observable x, we do not add noise to this agent's x. The noise is only added to the data set on which the algorithm is trained.[11] In other words, Y^{ews} sets each agent's allocation based on an estimate of η, where η is estimated using artificially noised up data. The logic of attenuation bias suggests that Y^{ews} is in some sense "flatter" with respect to the manipulable components of x, or "puts less weight" on those components, relative to the best response policy that does not add noise.

We hope future research will explore this proposal systematically and study its benefits in improving information from manipulable data in complex environments.

[11] Note that adding noise to the data here does not necessarily mean that the policy function Y^{ews} will be stochastic; indeed, by estimating on resampled data points with independent noise draws, the function can be made essentially deterministic conditional on the true data. In contrast, mechanisms designed to keep agent characteristics hidden from an observer may require stochastic output conditional on the underlying data. See the literature on differential privacy, surveyed in Dwork (2011).

A. Appendix: Proofs

A.1. Proof of Proposition 1

From Subsection 2.2, (β^*, β_0^*) solves

$$\min_{(\beta,\beta_0)\in\mathbb{R}^2} \mathbb{E}[(m\beta^2\gamma + \beta_0 - (1-\beta)\eta)^2].$$

The first-order condition with respect to β_0 implies

$$\beta_0^* = (1-\beta)\mu_\eta - m\beta^2\mu_\gamma.$$

Substituting β_0^* into the objective, the designer chooses β to minimize

$$\begin{aligned}
&\mathbb{E}[(m\beta^2(\gamma-\mu_\gamma) - (1-\beta)(\eta-\mu_\eta))^2]\\
=&(1-\beta)^2\sigma_\eta^2 + m^2\beta^4\sigma_\gamma^2 - 2(1-\beta)m\beta^2\rho\sigma_\eta\sigma_\gamma\\
=&\sigma_\eta^2\left[\left((1-\beta)-k\beta^2\right)^2 + 2(1-\rho)\beta^2(1-\beta)k\right],
\end{aligned}$$

where

$$k \equiv m\sigma_\gamma/\sigma_\eta > 0.$$

Equivalently, for $k > 0$ and $\rho \in (-1,1)$, β^* minimizes

$$L(\beta,k,\rho) \equiv \left(k\beta^2+\beta-1\right)^2 + 2(1-\rho)\beta^2(1-\beta)k. \tag{A.1}$$

Differentiating,

$$L_\beta(\beta,k,\rho) = -2(1-\beta) + 4k^2\beta^3 + 2\rho k\beta(3\beta-2). \tag{A.2}$$

Note that $L_\beta(0,k,\rho) < 0$, i.e., there is a first-order benefit from putting some positive weight on the agent's action.

The last statement of Proposition 1 follows from the second because, from Equation 5, $\hat{\beta}(\cdot)$ is continuous, $\hat{\beta}(0) > 0$, and $\hat{\beta}(\beta^{\text{fp}}) = \beta^{\text{fp}}$ for any β^{fp}. Proposition 1 is thus implied by Lemma 1 and the following result. We abuse notation hereafter and drop the arguments k and ρ from $L(\cdot)$ when those are held fixed. So, for example, $L(\beta)$ means that both k and ρ

are fixed.

Lemma A.1. *There exists $\beta^* \in (0,2)$ such that:*

1. *The loss function $L(\beta)$ from (A.1) is uniquely minimized over $\beta \in \mathbb{R}$ at β^*.*
2. *$\beta^* = \min_{\beta \geq 0}\{\beta : L'(\beta) \geq 0\}$.*
3. *$L''(\beta^*) > 0$.*

Proof. The proof has five steps below. Steps 1–3 are building blocks to Step 4, which establishes that all minimizers of $L(\beta)$ are in $(0,2)$. Step 5 then establishes there is in fact a unique minimizer, and it has the requisite properties. It is useful in this proof to extend the domain of the function L defined in (A.1) to include $\rho = -1$ and $\rho = 1$.

Step 1: We first establish two useful properties of $L(\beta, \rho = 1)$. Simplifying (A.1),

$$L(\beta, \rho = 1) = \left(k\beta^2 + \beta - 1\right)^2$$

is the square of a quadratic. The quadratic $k\beta^2 + \beta - 1$ is strictly convex in β, minimized at

$$\beta = \beta^m \equiv -1/(2k) < 0, \tag{A.3}$$

and, because it has one negative and one positive root, it is negative and strictly increasing on $[\beta^m, 0]$. It follows that $L(\cdot, \rho = 1)$ is strictly decreasing on $[\beta^m, 0]$ and symmetric around β^m (i.e., for any x, $L(\beta^m + x, \rho = 1) = L(\beta^m - x, \rho = 1)$).

Step 2: We claim that for any $\beta < 0$ and $\rho < 1$, there is $\tilde{\beta} \geq 0$ such that $L(\tilde{\beta}) < L(\beta)$. Since $L'(0) < 0$, it follows that for $\rho < 1$, $\arg\min L(\beta, \rho) \subset \mathbb{R}_{++}$.

To prove the claim, we first establish that for any $x > 0$ and $\beta = \beta^m - x$ (where β^m is defined in (A.3)), the symmetric point $\beta^m + x$ has a lower loss when $\rho < 1$; note that $\beta^m + x$ may also be negative. The argument is as follows:

$$\begin{aligned}
L(\beta^m - x, \rho) - L(\beta^m + x, \rho) &= L(\beta^m - x, \rho = 1) + 2(1-\rho)(\beta^m - x)^2(1 - \beta^m + x)k \\
&\quad - \left[L(\beta^m + x, \rho = 1) + 2(1-\rho)(\beta^m + x)^2(1 - \beta^m - x)k\right] \\
&= 2(1-\rho)k\left[(\beta^m - x)^2(1 - \beta^m + x) - (\beta^m + x)^2(1 - \beta^m - x)\right] \\
&= 4(1-\rho)kx\left(\beta^m(3\beta^m - 2) + x^2\right) \\
&\geq 0,
\end{aligned}$$

where the first equality is from the definition in (A.1), the second is because Step 1 established that $L(\beta^m + x, \rho = 1) = L(\beta^m - x, \rho = 1)$, the third equality is from algebraic simplification, and the inequality is because $\beta^m < 0$, $x > 0$, and $\rho < 1$.

It now suffices to establish $L(0, \rho) < L(\beta, \rho)$ for all $\beta \in [\beta^m, 0)$. Differentiating (A.2) yields $L_{\beta\rho}(\beta, \rho) = 2k\beta(3\beta - 2) > 0$ when $\beta < 0$. Hence for $\beta \in [\beta^m, 0)$, $L(0, \rho) - L(\beta, \rho) \leq L(0, \rho = 1) - L(\beta, \rho = 1) < 0$, where the strict inequality is from Step 1.

Step 3: $\arg\min_\beta L(\beta, \rho = -1) \cap (0, 2] \neq \emptyset$.

To prove this, simplify (A.1) to get

$$L(\beta, \rho = -1) = \left(k\beta^2 - \beta + 1\right)^2.$$

The quadratic $k\beta^2 - \beta + 1$ is strictly convex in β and minimized at $\beta = 1/(2k)$; moreover, if $k \geq 1/4$ then that quadratic is nonnegative, and otherwise it is equal to zero at $\beta = \frac{1 \pm \sqrt{1-4k}}{2k}$. It follows that if $k \geq 1/4$, $\arg\min L(\beta, \rho = -1) = \{1/(2k)\} \subset (0, 2]$. If $k \in (0, 1/4)$, $\min \arg\min L(\beta, \rho = -1) = \frac{1 - \sqrt{1-4k}}{2k} \in (0, 2)$.

Step 4: For $\rho \in (-1, 1)$, $\arg\min_\beta L(\beta, \rho) \subset (0, 2)$.

To prove this, note that $L_{\beta\rho}(\beta, \rho) = 2k\beta(3\beta - 2) > 0$ when $\beta > 2/3$. Monotone comparative statics (see Fact 1 in the Supplementary Appendix) imply that on the domain $(2/3, \infty)$ every minimizer of $L(\cdot, \rho)$ when $\rho > -1$ is smaller than every minimizer of $L(\cdot, \rho = -1)$. Step 3 then implies that all minimizers when $\rho > -1$ are less than 2; Step 2 established that when $\rho < 1$, all minimizers are larger than 0.

Step 5: Finally, we claim that for $\rho \in (-1, 1)$, $L'(\beta)$ has only one root in $(0, 2)$; moreover, $L''(\beta) > 0$ at that root. The lemma follows because $L'(\beta)$ is continuous and $L'(0) < 0$.

To prove the claim, first observe from Equation A.2 that $L'(\beta)$ is a cubic function that is initially strictly concave and then strictly convex, with inflection point $\beta = -\rho/(2k)$. For the rest of the proof, view L' or L'' as a function of β only.

1. If $\rho \geq 0$, then the inflection point is negative, and thus L' is strictly convex on $\beta > 0$. Since $L'(0) < 0$, L' has only one positive root, and $L'' > 0$ at that root.

2. Consider $\rho \in (-1, 0)$. L'' is minimized at the inflection point of L'. Differentiating Equation A.2 and evaluating at the inflection point,

$$L''\left(\frac{-\rho}{2k}\right) = 2 + 12k^2\left(\frac{-\rho}{2k}\right)^2 + 4\rho k\left(3\left(\frac{-\rho}{2k}\right) - 1\right) = 2 - 3\rho^2 - 4k\rho.$$

If this expression is positive, then $L''(\beta) > 0$ for all β, i.e., L' is strictly increasing and hence has a unique root.

So suppose instead $2 - 3\rho^2 - 4k\rho \leq 0$. Equivalently, since $\rho < 0$, suppose

$$k \leq \frac{2 - 3\rho^2}{4\rho}.$$

The right-hand side of this inequality is less than $-\rho/4$ because $\rho \in (-1, 0)$, and hence $k < -\rho/4$. Consequently, the inflection point, $\beta = -\rho/(2k)$, is larger than 2, and therefore $L'(\beta)$ is concave over $\beta \in (0, 2)$. Moreover, recall that $L'(0) < 0$, and also observe that $L'(2) = 32k^2 + 16k\rho + 2 > 0$ because $k < -\rho/4$ and $\rho \in (-1, 0)$. It follows that L' has only one root on $(0, 2)$, and $L'' > 0$ at that root. □

A.2. Proof of Lemma 1

It holds that

$$\mathbb{E}[(\hat{\eta}_\beta(x) - \eta)^2] = \sigma_\eta^2 \left(1 - R_{\eta x}^2\right) = \sigma_\eta^2 - (\hat{\beta}(\beta))^2 \mathrm{Var}(x),$$

where the first equality was noted after Equation 6, and the second equality holds because $R_{\eta x}^2 = (\mathrm{Cov}(x, \eta))^2 / (\mathrm{Var}(\eta)\mathrm{Var}(x))$ and $\hat{\beta}(\beta) = \mathrm{Cov}(x, \eta)/\mathrm{Var}(x)$.

We also have

$$\begin{aligned}
\mathbb{E}[(Y(x) - \hat{\eta}_\beta(x))^2] &= \mathbb{E}[(\beta x + \beta_0 - \hat{\beta}(\beta)x - \hat{\beta}_0(\beta))^2] && \text{from definitions} \\
&= \mathbb{E}\left[\left((\beta - \hat{\beta}(\beta))(x - \mathbb{E}[x])\right)^2\right] \\
&= (\beta - \hat{\beta}(\beta))^2 \mathrm{Var}(x),
\end{aligned}$$

where the second line is because $\beta\mathbb{E}[x] + \beta_0 = \mu_\eta = \hat{\beta}(\beta)\mathbb{E}[x] + \hat{\beta}_0(\beta)$ (the second equality here is standard; for the first, see the beginning of the proof of Proposition 1) and hence $\beta_0 - \hat{\beta}_0(\beta) = (\hat{\beta}(\beta) - \beta)\mathbb{E}[x]$.

Substituting these formulae into Equation 6 yields

$$\mathcal{L}(\beta) = \underbrace{\sigma_\eta^2 - (\hat{\beta}(\beta))^2 \mathrm{Var}(x)}_{\text{Info loss}} + \underbrace{(\beta - \hat{\beta}(\beta))^2 \mathrm{Var}(x)}_{\text{Misallocation loss}}.$$

Differentiating,

$$\mathcal{L}'(\beta) = \overbrace{\left(-2\hat{\beta}(\beta)\hat{\beta}'(\beta)\text{Var}(x) - (\hat{\beta}(\beta))^2 \frac{\mathrm{d}}{\mathrm{d}\beta}\text{Var}(x)\right)}^{\text{Marginal change in info loss}}$$
$$+ \underbrace{\left(-2(\beta - \hat{\beta}(\beta))\hat{\beta}'(\beta)\text{Var}(x) + (\beta - \hat{\beta}(\beta))^2 \frac{\mathrm{d}}{\mathrm{d}\beta}\text{Var}(x)\right)}_{\text{Marginal change in misallocation loss}}.$$

When $\beta = \beta^{\text{fp}} = \hat{\beta}(\beta^{\text{fp}})$, the marginal change in misallocation loss is evidently zero. Thus,

$$\mathcal{L}'(\beta^{\text{fp}}) = -2\beta^{\text{fp}}\hat{\beta}'(\beta^{\text{fp}})\text{Var}(x) - (\beta^{\text{fp}})^2 \frac{\mathrm{d}}{\mathrm{d}\beta}\text{Var}(x).$$

Using $\text{Var}(x) = \text{Cov}(x, \beta)/\hat{\beta}(\beta)$, $\text{Cov}(x, \beta) = \sigma_\eta^2 + m\rho\sigma_\eta\sigma_\gamma\beta$, $\text{Var}(x) = \text{Cov}(x, \beta) + m\rho\sigma_\eta\sigma_\gamma\beta + m^2\sigma_\gamma^2\beta^2$, and $\beta^{\text{fp}} = \hat{\beta}(\beta^{\text{fp}})$, some algebra yields[12]

$$\mathcal{L}'(\beta^{\text{fp}}) = \frac{2m^2}{\text{Var}(x)}(\beta^{\text{fp}})^2\sigma_\eta^2\sigma_\gamma^2(1-\rho^2).$$

Hence, $\mathcal{L}'(\beta^{\text{fp}}) > 0$ because $\beta^{\text{fp}} \neq 0$ (as $\hat{\beta}(0) = 1$ from Equation 5) and $\rho \in (-1, 1)$.

A.3. Proof of Proposition 2

The proof is via the following claims. Applying Lemma A.1, we without loss restrict attention to $\beta \in (0, 2)$ in all the claims.

Claim A.1. *β^* is continuously differentiable in ρ and k.*

Proof. Lemma A.1 established that $\text{sign}[L''(\beta^*)] > 0$. Thus, the implicit function theorem

[12] Letting C and V be shorthand for $\text{Cov}(x, \beta)$ and $\text{Var}(x)$ respectively, a prime denote the derivative with respect to β, suppressing arguments, evaluating all functions at β^{fp}, and using the properties noted:

$$\begin{aligned}
\mathcal{L}' &= -2\beta^{\text{fp}}\hat{\beta}'V - (C/V)^2V' = -2C\hat{\beta}' - (C/V)^2V' = (-2CVC' + C^2V')/V^2 \\
&= (C/V^2)\left[-2CC'(C + m\rho\sigma_\eta\sigma_\gamma\beta^{\text{fp}} + m^2\sigma_\gamma^2(\beta^{\text{fp}})^2) + C^2(2C' + 2m^2\sigma_\gamma^2\beta^{\text{fp}})\right] \\
&= (2\beta^{\text{fp}}C/V^2)\left[-C'(m\rho\sigma_\eta\sigma_\gamma + m^2\sigma_\gamma^2\beta^{\text{fp}}) + Cm^2\sigma_\gamma^2\right] \\
&= (2\beta^{\text{fp}}C/V^2)\left[-(m\rho\sigma_\eta\sigma_\gamma)^2 - (m\rho\sigma_\eta\sigma_\gamma)m^2\sigma_\gamma^2\beta^{\text{fp}}) + (\sigma_\eta^2 + m\rho\sigma_\eta\sigma_\gamma\beta^{\text{fp}})m^2\sigma_\gamma^2\right] \\
&= (2\beta^{\text{fp}}C/V^2)m^2(\sigma_\eta\sigma_\gamma)^2(1-\rho^2) \\
&= 2(\beta^{\text{fp}})^2(1/V)m^2\sigma_\eta^2\sigma_\gamma^2(1-\rho^2).
\end{aligned}$$

guarantees the existence of $\frac{\mathrm{d}\beta^*}{\mathrm{d}k} = -\frac{L_{\beta k}}{L_{\beta\beta}}$ and $\frac{\mathrm{d}\beta^*}{\mathrm{d}\rho} = -\frac{L_{\beta\rho}}{L_{\beta\beta}}$. $\square$

Claim A.2. *If $k > 3/4$ then $\beta^* < 2/3$ and is strictly increasing in ρ. If $k < 3/4$ then $\beta^* > 2/3$ and is strictly decreasing in ρ. If $k = 3/4$ then $\beta^* = 2/3$ independent of ρ.*

Proof. From Equation A.2 compute the cross partial

$$L_{\beta\rho} = 2k\beta(3\beta - 2).$$

Hence $L_{\beta\rho} < 0$ when $\beta < 2/3$, while $L_{\beta\rho} > 0$ when $\beta > 2/3$. Moreover, it follows from Equation A.2 that when $\beta = 2/3$, $\text{sign}[L_\beta] = \text{sign}[k - 3/4]$ independent of ρ.

1. Consider $k = 3/4$. Routine algebra verifies that L_β is strictly increasing in β, and hence $L_\beta = 0 \implies \beta = 2/3$, i.e., $\beta^* = 2/3$ independent of ρ.

2. Consider $k > 3/4$. Since $L_\beta > 0$ when $\beta = 2/3$, it follows that $\beta^* < 2/3$. (Recall $L_\beta < 0$ when $\beta = 0$, and Lemma A.1 implies that $\beta^* = \min\{\beta > 0 : L_\beta = 0\}$.) Since $L_{\beta\rho} < 0$ on the domain $\beta < 2/3$, monotone comparative statics (see Fact 1 in the Supplementary Appendix) imply β^* is strictly increasing in ρ.

3. Consider $k < 3/4$. For $\rho = 0$, we have $L_{\beta k} = 8k\beta^3 > 0$ and hence $\beta^* > 2/3$ using $\beta^* = 2/3$ when $k = 3/4$ and monotone comparative statics. It follows that $\beta^* > 2/3$ for all ρ because β^* is continuous in ρ and $L_\beta < 0$ when $\beta = 2/3$ whereas $L_\beta = 0$ when $\beta = \beta^*$. Since $L_{\beta\rho} > 0$ on the domain $\beta > 2/3$, monotone comparative statics imply β^* is strictly decreasing in ρ. $\square$

Claim A.3. *As $k \to \infty$, $\beta^* \to 0$; as $k \to 0$, $\beta^* \to 1$. If $\rho \geq 0$ then β^* is strictly decreasing in k. If $\rho < 0$ then β^* is strictly quasi-concave in k, attaining a maximum at some point.*

Proof. The first statement about limits is evident from inspecting Equation A.2. For the comparative statics, compute the cross partials

$$L_{\beta k} = 8k\beta^3 + 2\rho\beta(3\beta - 2) \quad \text{and} \quad L_{\beta kk} = 8\beta^3 > 0.$$

Since $\frac{\mathrm{d}\beta^*}{\mathrm{d}k} = -\frac{L_{\beta k}}{L_{\beta\beta}}$ and, from Lemma A.1, $L_{\beta\beta} > 0$ at $\beta = \beta^*$, the sign of $\frac{\mathrm{d}\beta^*}{\mathrm{d}k}$ is the sign of $-L_{\beta k}$. Using $\beta^* \to 1$ as $k \to 0$, we see that for small k and at $\beta = \beta^*$, $L_{\beta k}$ is larger than but arbitrarily close to 2ρ.

1. It follows that $L_{\beta k} > 0$ for all k and $\beta = \beta^*$ when $\rho \geq 0$. That is, $\frac{\mathrm{d}\beta^*}{\mathrm{d}k} < 0$ when $\rho \geq 0$.

2. Consider $\rho < 0$. Plainly $L_{\beta k} < 0$ for small k and $\beta = \beta^*$, while for some k it becomes positive (since $\beta^* \to 0$ as $k \to \infty$). Since $L_{\beta k}$ is strictly increasing in k, it follows that $\frac{\mathrm{d}\beta^*}{\mathrm{d}k}$ is strictly decreasing in k, initially positive and eventually negative. □

Claim A.4. *Assume $\rho = 0$. There is a unique β^{fp}, which is positive. Both β^{fp} and $\beta^*/\beta^{\text{fp}}$ are strictly decreasing in k. Moreover, $\beta^*/\beta^{\text{fp}} \to 1$ as $k \to \infty$ and $\beta^*/\beta^{\text{fp}} \to \sqrt[3]{1/2}$ as $k \to 0$.*

Proof. Assume $\rho = 0$. Equation C.1 simplifies to

$$k^2(\beta^{\text{fp}})^3 + \beta^{\text{fp}} - 1 = 0, \tag{A.4}$$

which has a unique solution, with $\beta^{\text{fp}} \in (0,1)$ strictly decreasing in k with range $(0,1)$.

The first-order condition for β^* simplifies to

$$2k^2(\beta^*)^3 + \beta^* - 1 = 0, \tag{A.5}$$

which has a unique solution, also in $(0,1)$ and strictly decreasing in k with range $(0,1)$.

Hence, $\beta^*/\beta^{\text{fp}} \to 1$ as $k \to 0$. Moroever, Equation A.4 and Equation A.5 imply that as $k \to \infty$, $k^2(\beta^{\text{fp}})^3 \to 1$ and $2k^2(\beta^*)^3 \to 1$, and hence $(\beta^*/\beta^{\text{fp}}) \to \sqrt[3]{1/2}$.

It remains to prove that $\beta^*/\beta^{\text{fp}}$ is strictly decreasing in k. Applying the implicit function theorem to Equation A.4 and Equation A.5 (which is indeed valid) and doing some algebra,

$$\begin{aligned}\frac{\mathrm{d}\beta^*}{\mathrm{d}k} &= -\frac{4k(\beta^*)^3}{6k^2(\beta^*)^2+1},\\ \frac{\mathrm{d}\beta^{\text{fp}}}{\mathrm{d}k} &= -\frac{2k(\beta^{\text{fp}})^3}{3k^2(\beta^{\text{fp}})^2+1}.\end{aligned}$$

$\beta^*/\beta^{\text{fp}}$ is strictly decreasing in k if and only if $\beta^{\text{fp}}\frac{\mathrm{d}\beta^*}{\mathrm{d}k} - \beta^*\frac{\mathrm{d}\beta^{\text{fp}}}{\mathrm{d}k} < 0$. Substituting in the formulae above, this inequality is equivalent to

$$\begin{aligned}&\frac{2k(\beta^{\text{fp}})^3\beta^*}{3k^2(\beta^{\text{fp}})^2+1} < \frac{4k(\beta^*)^3\beta^{\text{fp}}}{6k^2(\beta^*)^2+1}\\ \iff& \left(6k^2(\beta^*)^2+1\right)(\beta^{\text{fp}})^2 < \left(3k^2(\beta^{\text{fp}})^2+1\right)2(\beta^*)^2\\ \iff& \beta^{\text{fp}} < \beta^*\sqrt{2}.\end{aligned}$$

Plainly, the last inequality holds as $k \to 0$ because both $\beta^{\text{fp}} \to 1$ and $\beta^* \to 1$ as $k \to 0$. By continuity, we are done if there is no k at which $\beta^{\text{fp}} = \beta^* \sqrt{2}$. Indeed there is not because then Equation A.4 would become equivalent to

$$2k^2(\beta^*)^3 + \beta^* - 1/\sqrt{2} = 0,$$

contradicting Equation A.5. $\square$

References

Ali, S. N. and R. Bénabou (2019): "Image Versus Information: Changing Societal Norms and Optimal Privacy," Forthcoming in *American Economic Journal: Microeconomics*.

Ball, I. (2020): "Scoring Strategic Agents," Working paper.

Bénabou, R. and J. Tirole (2006): "Incentives and Prosocial Behavior," *American Economic Review*, 96, 1652–1678.

Blackwell, D. (1951): "Comparison of Experiments," in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, vol. 1, 93–102.

Boleslavsky, R., D. L. Kelly, and C. R. Taylor (2017): "Selloffs, Bailouts, and Feedback: Can Asset Markets Inform Policy?" *Journal of Economic Theory*, 169, 294–343.

Bonatti, A. and G. Cisternas (2019): "Consumer Scores and Price Discrimination," Forthcoming in *Review of Economic Studies*.

Bond, P. and I. Goldstein (2015): "Government Intervention and Information Aggregation by Prices," *Journal of Finance*, 70, 2777–2812.

Braverman, M. and S. Garg (2019): "The Role of Randomness and Noise in Strategic Classification," ACM EC 2019 Workshop on Learning in Presence of Strategic Behavior.

DWORK, C. (2011): "Differential Privacy," *Encyclopedia of Cryptography and Security*, 338–340.

EDERER, F., R. HOLDEN, AND M. MEYER (2018): "Gaming and Strategic Opacity in Incentive Provision," *RAND Journal of Economics*, 49, 819–854.

ELIAZ, K. AND R. SPIEGLER (2019): "The Model Selection Curse," *American Economic Review: Insights*, 1, 127–40.

FISCHER, P. E. AND R. E. VERRECCHIA (2000): "Reporting Bias," *The Accounting Review*, 75, 229–245.

FRANKEL, A. AND N. KARTIK (2019): "Muddled Information," *Journal of Political Economy*, 129, 1739–1776.

GESCHE, T. (2019): "De-biasing Strategic Communication," Unpublished.

GÓMEZ, E., M. A. GÓMEZ-VILLEGAS, AND J. M. MARÍN (2003): "A Survey on Continuous Elliptical Vector Distributions," *Revista matemática complutense*, 16, 345–361.

HARBAUGH, R. AND E. RASMUSEN (2018): "Coarse Grades: Informing the Public by Withholding Information," *American Economic Journal: Microeconomics*, 10, 210–35.

HARDT, M., N. MEGIDDO, C. PAPADIMITRIOU, AND M. WOOTTERS (2016): "Strategic Classification," in *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, ACM, 111–122.

HU, L., N. IMMORLICA, AND J. W. VAUGHAN (2019): "The Disparate Effects of Strategic Manipulation," in *ACM Conference on Fairness, Accountability, and Transparency*, Atlanta, Georgia.

JANN, O. AND C. SCHOTTMÜLLER (2018): "An Informational Theory of Privacy," Unpublished.

KLEINBERG, J. AND M. RAGHAVAN (2019): "How Do Classifiers Induce Agents to Invest Effort Strategically?" in *Proceedings of the 2019 ACM Conference on Economics and Computation*, ACM, 825–844.

KREPS, D. AND R. WILSON (1982): "Sequential Equilibria," *Econometrica*, 50, 863–894.

LIANG, A. AND E. MADSEN (2020): "Data Linkages and Incentives," unpublished.

Martinez-Gorricho, S. and C. Oyarzun (2019): "Hypothesis Testing with Endogenous Information," Unpublished.

Milli, S., J. Miller, A. D. Dragan, and M. Hardt (2018): "The Social Cost of Strategic Classification," ArXiv:1808.08460v2 [cs.LG].

Mussa, M. and S. Rosen (1978): "Monopoly and Product Quality," *Journal of Economic Theory*, 18, 301–317.

Perez-Richet, E. and V. Skreta (2018): "Test Design Under Falsification," Unpublished.

Prendergast, C. and R. H. Topel (1996): "Favoritism in Organizations," *Journal of Political Economy*, 104, 958–978.

Spence, M. (1973): "Job Market Signaling," *Quarterly Journal of Economics*, 87, 355–374.

Supplementary (Online) Appendices

B. Monotone Comparative Statics

The following fact on monotone comparative statics is used in the proof of Proposition 1 and in the proof of Proposition 2. Although it is well known, we include a proof.

Fact 1. Let $T \subseteq \mathbb{R}$, $Z \subseteq \mathbb{R}$ be open, and $f : Z \times T \to \mathbb{R}$ be continuously differentiable in z with for all $t \in T$, $\arg\min_{z\in Z} f(z,t) \neq \emptyset$. Define $M(t) \equiv \arg\min_{z\in Z} f(z,t)$. For any $\bar{t} \in T$ and $\underline{t} \in T$ with $\bar{t} > \underline{t}$, it holds that:

1. If $f_z(z,\bar{t}) > f_z(z,\underline{t})$ for all $z \in Z$, then for any $\overline{m} \in M(\bar{t})$ and any $\underline{m} \in M(\underline{t})$ it holds that $\overline{m} < \underline{m}$.

   Proof: For any $\hat{z} > \underline{m}$,

   $$f(\hat{z},\bar{t}) - f(\underline{m},\bar{t}) = \int_{\underline{m}}^{\hat{z}} f_z(z,\bar{t})\mathrm{d}z > \int_{\underline{m}}^{\hat{z}} f_z(z,\underline{t})\mathrm{d}z = f(\hat{z},\underline{t}) - f(\underline{m},\underline{t}) \geq 0.$$

   Hence $\overline{m} \leq \underline{m}$. The inequality must be strict because otherwise the first-order conditions yield $0 = f_z(\overline{m},\bar{t}) = f_z(\underline{m},\bar{t}) > f_z(\underline{m},\underline{t}) = 0$.

2. If $f_z(z,\bar{t}) < f_z(z,\underline{t})$ for all $z \in Z$, then for any $\overline{m} \in M(\bar{t})$ and any $\underline{m} \in M(\underline{t})$ it holds that $\overline{m} > \underline{m}$. (We omit a proof, as it is analogous to that above.) $\square$

C. Additional Results

Proposition C.1. *There exists $\beta^{\mathrm{fp}} > 0$ satisfying $\hat{\beta}(\beta^{\mathrm{fp}}) = \beta^{\mathrm{fp}}$. If $\rho \geq 0$ there is only one $\beta^{\mathrm{fp}} \geq 0$, and it satisfies $\beta^{\mathrm{fp}} \in (0,1)$.*

That there is only one positive fixed point under nonnegative correlation has been noted in different form in Frankel and Kartik (2019, Proposition 4).

Proof of Proposition C.1. For $\beta \geq 0$, Equation 5 can be rewritten as the cubic equation

$$m^2\sigma_\gamma^2\beta^3 + 2m\rho\sigma_\eta\sigma_\gamma\beta^2 + (\sigma_\eta^2 - m\rho\sigma_\eta\sigma_\gamma)\beta - \sigma_\eta^2 = 0. \tag{C.1}$$

The left-hand side of (C.1) is continuous, negative at $\beta = 0$ and tends to ∞ as $\beta \to \infty$. There is a positive solution to (C.1) by the intermediate value theorem.

For the second statement of the proposition, differentiate $\hat{\beta}(\cdot)$ from Equation 5 to obtain

$$\hat{\beta}'(\beta) = -\frac{m\sigma_\eta\sigma_\gamma(2\beta m\sigma_\eta\sigma_\gamma + \rho\sigma_\eta^2 + \rho\beta^2 m^2\sigma_\gamma^2)}{(\sigma_\eta^2 + 2\beta m\rho\sigma_\eta\sigma_\gamma + \beta^2 m^2\sigma_\gamma^2)^2}.$$

When $\rho \geq 0$, this derivative is negative for all $\beta > 0$. The result follows from the fact that $\hat{\beta}(0) = 1$ and, when $\rho \geq 0$, $\hat{\beta}(1) < 1$. □

D. Alternative Model of Information Loss

Our paper finds that a designer improves information, and thereby allocation accuracy, by flattening a fixed point rule. We developed this point in what we believe is a canonical model of information loss from manipulation, one used in a number of other papers. But we think the point applies more broadly, including in other models of information loss. For instance, even a model with a one-dimensional type (such as the model in this paper with no heterogeneity on the gaming ability γ) can lead to information loss when there is a bounded action space and strong manipulation incentives. The reason is "pooling at the top". We establish below a version of our main result for a simple model in this vein.

Let the agent take action $x \in \{0, 1\}$ with natural action $\eta \in \{0, 1\}$. The agent's type η is her private information, drawn with ex-ante probability $\pi \in (0, 1)$ that $\eta = 1$. After observing x, the designer chooses allocation $y \in \mathbb{R}$ with payoff $-(y - \eta)^2$. We assume, for simplicity, that the agent of type $\eta = 1$ must choose $x = 1$.[13] The payoff for type $\eta = 0$ is $y - cx$, where $c > 0$ is a commonly known parameter. To streamline the analysis, we assume $c \in (0, \pi)$.

A pure allocation rule or policy is $Y : \{0, 1\} \to \mathbb{R}$. Due to the designer's quadratic loss payoff, it is without loss to focus on pure policies. Given a policy Y, let $\Delta \equiv Y(1) - Y(0)$ be the difference in allocations across the two actions of the agent. We focus, without loss, on policies with $\Delta \geq 0$. A policy with a smaller Δ is a "flatter" policy, i.e., it is less sensitive to the agent's action. The naive policy Y^{n} sets $Y^{\mathrm{n}}(1) = 1$ and $Y^{\mathrm{n}}(0) = 0$, corresponding to a naive allocation difference of $\Delta^n = 1$. Let Δ^{fp} and Δ^* denote the corresponding differences from fixed point and commitment policies.

[13] Our main point goes through so long as action $x = 1$ is no more costly than $x = 0$ for type $\eta = 1$, as this will ensure it is optimal for type $\eta = 1$ to choose $x = 1$.

D.1. Naive Policy

Take any policy with $\Delta = 1$. Since we assume $c < \pi < 1$, even the agent with $\eta = 0$ will then choose $x = 1$. So welfare—the designer's ex-ante expected payoff—from the naive policy is

$$-\pi(0-0)^2 - (1-\pi)(1-0)^2 = -(1-\pi).$$

D.2. Fixed Point

At a Bayesian Nash equilibrium (of either the simultaneous move game, or when the agent moves first), $Y(x) = \mathbb{E}[\eta|x]$ for any x on the equilibrium path. If $x = 0$ is on the equilibrium path, $Y(0) = 0$ because type $\eta = 1$ does not play $x = 0$.

There is a fully-pooling equilibrium with both types playing $x = 1$: the designer plays $Y(1) = \pi$ and $Y(0) = 0$, and it is optimal for type $\eta = 0$ to play $x = 1$ because $c < \pi$. The corresponding welfare is

$$-\pi(\pi-1)^2 - (1-\pi)(\pi-0)^2 = -\pi(1-\pi).$$

There is no equilibrium in which the agent of type $\eta = 0$ puts positive probability on action $x = 0$, because that would imply $Y(1) > \pi$ and $Y(1) = 0$, against which the agent's unique best response is to play $x = 1$.

Therefore, we have identified the (essentially unique, up to the off-path allocation following $x = 0$) fixed point policy: $Y^{\text{fp}}(1) = \pi$, $Y^{\text{fp}}(0) = 0$, and therefore $\Delta^{\text{fp}} = \pi$. The agent pools on $x = 1$, and welfare is $-\pi(1-\pi)$.[14] This welfare is larger than that of the naive policy.

D.3. Commitment

Now suppose the designer's commits to a policy before the agent moves. From the earlier analysis, if $\Delta > c$ the agent will pool at $x = 1$ and so an optimal such policy is the fixed point policy Y^{fp}. For any $\Delta < c$, there is full separation: the agent's best response is $x = \eta$. Indeed, full separation is also a best response for the agent when $\Delta = c$. Given that the designer wants to match the agent's type, it follows that the optimal way to induce full separation is to set $\Delta = c$ (or $\Delta = c^-$), i.e., have $Y^*(1) = Y^*(0) + c$.

[14] The choice of $Y^{\text{fp}}(0) = 0$ can be justified from the perspective of the agent "trembling". In particular, in the signaling game where the agent moves before the designer, any sequential equilibrium (Kreps and Wilson, 1982) has $Y(0) = 0$, as only type $\eta = 0$ can play $x = 0$. But note that no matter how $Y(0)$ is specified, it must hold in a fixed point that $\Delta \leq c$; otherwise the agent will not pool at $x = 1$.

At such an optimum, quadratic loss utility implies that the designer sets an average action of $(1-\pi)Y^*(0) + \pi Y^*(1)$ equal to $\mathbb{E}[\eta] = \pi$. Plugging in $Y^*(1) = Y^*(0) + c$ yields

$$(1-\pi)Y^*(0) + \pi(Y^*(0) + c) = \pi,$$

and hence the solution

$$Y^*(0) = \pi(1-c), \quad Y^*(1) = \pi(1-c) + c.$$

The corresponding welfare is

$$-(1-\pi)(\pi(1-c) - 0)^2 - \pi(\pi(1-c) + c - 1)^2 = -(1-c)^2(1-\pi)\pi.$$

This welfare is larger than that under the fixed point. Moreover, the optimal policy has $\Delta^* = c$ while the fixed point has $\Delta^{\mathrm{fp}} = \pi$ and the naive policy has $\Delta^{\mathrm{n}} = 1$. Thus the optimal policy is flatter than the fixed point, which in turn is flatter than the naive policy:

$$\Delta^* < \Delta^{\mathrm{fp}} < \Delta^{\mathrm{n}}.$$

Note that the designer obtains no benefit from reducing Δ from $\Delta^{\mathrm{fp}} = \pi$ until reaching $\Delta^* = c$; this is an artifact of the assumption that there is no heterogeneity in the manipulation cost c. In a model with such heterogeneity, there would be a more continuous benefit of reducing Δ from the fixed point.