Designing Evaluation Rules That Are Robust to Strategic Behavior

Jon Kleinberg
Cornell University

Manish Raghavan
Cornell University

Abstract

Machine learning is often used to produce decision-making rules that classify or evaluate individuals. When these individuals have incentives to be classified a certain way, they may behave strategically to influence their outcomes. We develop a model for how strategic agents can invest effort to change the outcomes they receive, and we give a tight characterization of when such agents can be incentivized to invest specified forms of effort into improving their outcomes as opposed to "gaming" the classifier. We show that whenever any "reasonable" mechanism can do so, a simple linear mechanism suffices. This work is based on "How Do Classifiers Induce Agents To Invest Effort Strategically?" published in Economics and Computation 2019 (Kleinberg and Raghavan 2019).

Introduction

Algorithmic decision-making is becoming increasingly common in a number of social contexts, including hiring, education, lending, and criminal risk assessment. Policies in these domains are often implemented by automated systems, evaluating individuals based on *features*, which serve as proxies to measure a person's underlying attributes. Accompanying the growing prevalence of such automated decision-making tools has been a push for *transparency*: algorithmic systems should be open to examination, and simple enough that people affected by them can understand their decisions. Proponents of transparency argue that it acts as a safeguard to prevent algorithmic systems from introducing biases or other undesirable properties, and that "secret" decision-making rules can create inequalities between insiders who know how the system works and outsiders who don't.

On the other hand, there are concerns that knowing exactly how decisions are being made will simply lead individuals to "game" the rule by strategically manipulating their appearances so as to receive favorable outcomes. Such concerns are a necessary consequence of the fact that an evaluator can only observe an individual via features that serve as imperfect measurements of his or her true qualities. From the evaluator's perspective, this creates a basic tension between effort the agent invests to raise the true underlying attributes that the evaluator cares about, and effort that may serve to improve the proxy features without actually improving the underlying attributes. This tension underlies the formulation of *Goodhart's Law*, widely known in the economics literature, which states that once a proxy measure becomes a goal in itself, it is no longer a useful measure (Hardt et al. 2016). This principle also underpins concerns about strategic gaming of evaluations in search engine rankings (Davis 2006), credit scoring (Bambauer and Zarsky 2018; Foust and Pressman 2008), academic paper visibility (Beel, Gipp, and Wilde 2009), reputation management (Zarsky 2007), and many other domains.

Incentivizing a desired effort investment. Viewing strategic behavior as completely undesirable, however, implicitly takes the position that the true qualities we wish to measure cannot be improved. In many settings, this doesn't accord with the fact that much of the effort people invest leads to improvements that in fact benefit both themselves and the measured evaluation. For example, if a student seeks to improve their GPA by learning the material, we would typically view this as a productive form of strategic behavior, both for the student and the evaluation process. But if the student tried to improve their GPA by learning a set of highly specific test-taking heuristics, we might instead see this as changing their feature value (GPA) without improving their underlying mastery of the material. Thus, different forms of behavior can modify appearances in many ways, only some of which a decision-maker may wish to encourage.

These considerations are at the heart of the following class of design problems, illustrated schematically in Figure 1. An *evaluator* creates a decision rule to assess an *agent* in terms of a set of features, and this leads the agent to make choices about how to invest effort across their actions to improve these features. From the evaluator's point of view, some forms of agent effort are valuable (like learning educational material), while others are not (like learning test-taking heuristics, or cheating). Hence, some decision rules work better than others in creating appropriate incentives: the evaluator would like to create a decision rule whose incentives lead the agent to invest in forms of effort that the evaluator considers valuable.

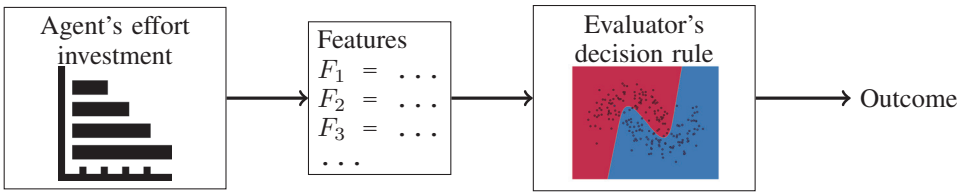


Figure 1: The basic framework: an agent chooses how to invest effort to improve the values of certain features, and an evaluator chooses a decision rule that creates indirect incentives favoring certain investments of effort over others.

The present work: Designing evaluation rules. We model this dynamic as a game between an *evaluator* who is performing an assessment, and an *agent* who wants to score well on this assessment. An instance of the problem consists of a set of actions in which the agent can invest *effort*, and a set of functions determining how the effort spent on these actions translates into *features* that the evaluator observes. The agent's goal is to achieve a high score by allocating their effort across actions. The evaluator's goal is to find an evaluation rule to induce a specific *effort profile* from the agent, which specifies a level of effort devoted to each action. Our main result tightly characterizes when a given effort profile can be incentivized and shows that a simple class of mechanisms suffices to do so.

Our work has close ties to the principal-agent literature from economics: an evaluator (the principal) wants to set a policy (the evaluation rule) that accounts for the agent's strategic responses. Our main result has some similarities, as well as some key differences, relative to a classical economic formulation in principal-agent models (Grossman and Hart 1983; Hermalin and Katz 1991; Holmstrom and Milgrom 1987; 1991). In particular, many of these results show the optimality of linear contracts, albeit under a different context and set of assumptions than the ones studied here.

Model and Motivating Example

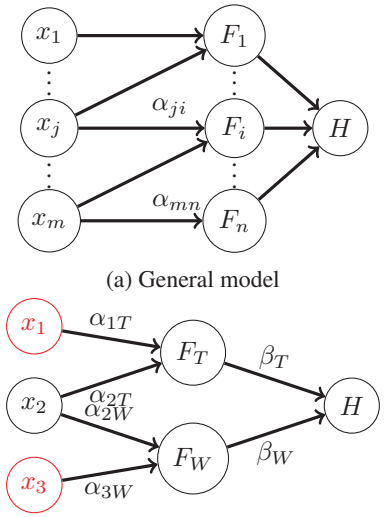
In this section, we develop a formal model of an agent's investment of effort. There are m actions the agent can take, and they must decide to allocate an amount of effort x_j to each activity j. We'll assume the agent has some budget B of effort to invest, so $\sum_{j=1}^{m} x_j \leq B$, and we'll call this investment of effort $x = (x_1, x_2, \ldots, x_m)$ an *effort profile*.

The evaluator cannot directly observe the agent's effort profile, but instead observes features $F_1, \ldots, F_n$ derived from the agent's effort profile. The value of each F_i grows monotonically in the effort the agent invests in certain actions according to an *effort conversion function* $f_i(\cdot)$:

$$F_i = f_i\left(\sum_{j=1}^{m} \alpha_{ji} x_j\right), \tag{1}$$

where each $f_i(\cdot)$ is nonnegative, smooth, weakly concave, and strictly increasing.

We represent these parameters of the problem using a bipartite graph with the actions $x_1, x_2, \ldots, x_m$ on the left, the features $F_1, \ldots, F_n$ on the right, and an edge of weight α_{ji} whenever $\alpha_{ji} > 0$, so that effort on action x_j contributes to the value of feature F_i. We call this graph, along with the associated parameters (the matrix $\alpha \in \mathbb{R}^{m \times n}$ with entries α_{ji}; functions $f_i : \mathbb{R} \to \mathbb{R}$ for $i \in \{1, ..., n\}$; and a budget B), the *effort graph* G. Figure 2 shows some examples of what G might look like.



(a) General model

(b) The classroom setting

Figure 2: The conversion of effort to feature values can be represented using a weighted bipartite graph, where effort x_j spent on action j has an edge of weight α_{ji} to feature F_i.

The evaluator combines the features F_i using some mechanism M (a function of the n feature values) to produce an output H, which is the agent's utility. We assume M is known to the agents. Because all features are increasing in the amount of effort invested by the agent — in particular, including the kinds of effort we want to incentivize — we'll restrict our attention to the class of monotone mechanisms, meaning that if agent X has larger values in all features than agent Y, then X's outcome should be at least as good as that of Y. Formally, we write this as follows:

Definition 1. A *monotone mechanism* M on features F_i is a mapping $\mathbb{R}^n \to \mathbb{R}$ such that for $F, F' \in \mathbb{R}^n$ with $F'_i \geq F_i$ for all $i \in \{1, ..., n\}$, $M(F') \geq M(F)$. Also, for any F, there exists $i \in \{1, ..., n\}$ such that strictly increasing F_i strictly increases $M(F)$ (meaning it is strictly optimal for the agent to invest its entire budget).

The agent's utility is simply its outcome H. Thus, for a

mechanism M, the agent's optimal strategy is given by the following optimization problem:

$$x^* = \arg\max_{x \in \mathbb{R}^m} M(F) \quad \text{s.t.} \sum_{i=1}^{n} x_i \leq B, \quad x \geq \mathbf{0} \qquad (2)$$

where each component F_i of F is defined as in (1). Throughout this work, we'll assume that agents behave rationally and optimally, though it would be an interesting subject for future work to consider extensions of this model where agents suffer from behavioral biases.

An extended example. To make this concrete, we proceed with an example demonstrating the subtle and somewhat counterintuitive effects at play here. We consider a classroom setting, where the strategic investment of effort has long been considered (Koretz et al. 1991; Koretz 2008). The effort graph shown in Figure 2b depicts this setting, where the teacher is the evaluator and the student is the agent. There are two pieces of graded work for the class (a test F_T and homework F_W), and the student can study the material (x_2) to improve their scores on both of these. They can also cheat on the test (x_1) and look up homework answers on-line (x_3). Their combined effort $\alpha_{1T}x_1 + \alpha_{2T}x_2$ contributes to their score on the test, and their combined effort $\alpha_{2W}x_2 + \alpha_{3W}x_3$ contributes to their score on the homework. We leave the budget B and effort conversion functions f_T and f_W uninstantiated for purposes of this example, as our main conclusions will not depend on them. From these scores, the teacher must decide on a student's final grade H. For simplicity, we'll assume the grading scheme is simply a linear combination, meaning $H = \beta_T F_T + \beta_W F_W$ for some real numbers $\beta_T, \beta_W \geq 0$.

The teacher's objective is to incentivize the student to learn the material; thus, they want induce the student to invest their entire budget into x_2. Of course, this may not be possible. For example, if α_{1T} and α_{3W} are significantly larger than α_{2T} and α_{2W} respectively, so that it is much easier to cheat on the test and copy homework answers than to study, the student would maximize their utility by investing all of their effort into these undesirable activities.

In fact, we can make this precise as follows. For any unit of effort invested in x_2, the student could instead invest $\frac{\alpha_{2T}}{\alpha_{1T}}$ and $\frac{\alpha_{2W}}{\alpha_{3W}}$ units of effort into x_1 and x_3 respectively without changing the values of F_T and F_W. Moreover, if $\frac{\alpha_{2T}}{\alpha_{1T}} + \frac{\alpha_{2W}}{\alpha_{3W}} < 1$, then this substitution strictly reduces the sum $x_1 + x_2 + x_3$, leaving additional effort available (relative to the budget constraint) for raising the values of F_T and F_W. It follows that any solution with $x_2 > 0$ can be strictly improved through this substitution. Thus, under this condition, the teacher cannot incentivize the student to study.

When $\frac{\alpha_{2T}}{\alpha_{1T}} + \frac{\alpha_{2W}}{\alpha_{3W}} \geq 1$, on the other hand, a consequence of our results is that no matter what f_T, f_W and B are, there exist some β_T, β_W that the teacher can choose to incentivize the student to invest all their effort into studying. This may be somewhat surprising – for instance, consider the case where $\alpha_{1T} = \alpha_{3W} = 3$ and $\alpha_{2T} = \alpha_{2W} = 2$, meaning that the best way for the student to maximize their score on each piece of graded work individually is to invest undesirable effort instead of studying. Even so, it turns out that the student can still be incentivized to put all of their effort into studying by appropriately balancing the weight placed on the two pieces of graded work.

Stating the main result. In order to state the main result, we must formalize the notion of linear mechanisms.

Definition 2. A *linear mechanism* $M : \mathbb{R}^n \to \mathbb{R}$ is the mapping $M(F) = \beta^\top F = \sum_{i=1}^n \beta_i F_i$ for some $\beta \in \mathbb{R}^n$ such that $\beta_i \geq 0$ for all $i \in \{1, ..., n\}$ and $\sum_{i=1}^n \beta_i > 0$.

We rule out the mechanism in which all β_i are equal to 0, as it is not a monotone mechanism.

We will say that a mechanism M *incentivizes* effort profile x if x is an optimal response to M. Our main result is the following theorem, characterizing when a given effort profile can be incentivized.

Theorem 3. *For an effort graph G and an effort profile x, let $S(x) = \{j \mid x_j > 0\}$, i.e., the support of x. Then, the following are equivalent:*

1. *There exists a linear mechanism that incentivizes x.*
2. *There exists a monotone mechanism that incentivizes x.*
3. *For all x' such that $S(x') \subseteq S(x)$, there exists a linear mechanism that incentivizes x.*

Furthermore, there is a polynomial time algorithm that decides the incentivizability of x and provides β to incentivize x whenever such β exists.

When there exists a monotone mechanism incentivizing x, we'll call both x and $S(x)$ *incentivizable*. When x is not incentivizable, this algorithm finds a succinct "obstacle" to $S(x)$, meaning no x' such that $S(x') = S(x)$ is incentivizable. The following corollary is a direct consequence of Theorem 3.

Corollary 4. *For a set $S \subseteq \{1, 2, \ldots, m\}$, some x such that $S(x) = S$ is incentivizable if and only if all x with $S(x) = S$ are incentivizable.*

As a result, whether or not an effort profile is incentivizable depends only on its *support*, or the set of actions we wish to incentivize.

Conclusion

Although Transparency is widely viewed as a desirable property of automated decision-making systems, one might worry that transparency makes decision-making susceptible to strategic behavior. In this work, we have developed a framework for reasoning about how agents may respond strategically to a publicly-known evaluation rule, showing that a simple class of linear mechanisms suffices to incentivize desired behavior.

It is interesting to consider connections between our approach and existing work on strategic behavior in classification. In the computer science literature, concerns about on-line spam motivated models of *adversarial classification* (Dalvi et al. 2004); recent formulations have considered a broader range of settings in which an evaluator publishes a rule and an agent then manipulates their features (Hardt et

al. 2016; Brückner and Scheffer 2011; Dong et al. 2018; Milli et al. 2019; Hu, Immorlica, and Vaughan 2019). In contrast with our work, these papers tend to assume that all strategic effort from the agent are undesirable. Work on strategyproof linear regression (Dekel, Fischer, and Procaccia 2010; Chen et al. 2018; Cummings, Ioannidis, and Ligett 2015) considers a different model, in which strategic agents submit (x, y) pairs, and an evaluator seeks a regression function that incentives truthful reporting of y.

There are also interesting potential links to the large literature on principal-agent problems in economics (Arrow 1963; Pauly 1968; Arrow 1968; Cheung 1969; Kerr 1975; Stiglitz 1974; Jensen and Meckling 1976; Ross 1973), including the notion of *moral hazard*. In these models, a principal wants to incentive agent actions that they cannot directly observe. Insurance markets are the canonical examples: agents reduce their liability by purchasing insurance, leading them to act more recklessly. As in our model, the agent's actions are often formalized as "effort variables" which, at some cost to the agent, increase the agent's level of "production" (Laffont and Martimort 2009). Qualitatively, there are two primary differences between these models and ours: in the insurance setting, the agent's utility is an exogenous function based on aversion to risk; and the principal and agent generally have aligned incentives. Despite these differences, it would be interesting to explore whether there may be insights that transfer between our model and this body of results in economics.

References

Arrow, K. J. 1963. Uncertainty and the welfare economics of medical care. *American Economic Review*.

Arrow, K. J. 1968. The economics of moral hazard: further comment. *American Economic Review* 58.

Bambauer, J., and Zarsky, T. 2018. The algorithm game. *Notre Dame L. Rev.* 94:1.

Beel, J.; Gipp, B.; and Wilde, E. 2009. Academic search engine optimization (ASEO) optimizing scholarly literature for Google Scholar & co. *Journal of scholarly publishing* 41(2):176–190.

Brückner, M., and Scheffer, T. 2011. Stackelberg games for adversarial prediction problems. In *Proc. 17th ACM Knowledge discovery and data mining*.

Chen, Y.; Podimata, C.; Procaccia, A. D.; and Shah, N. 2018. Strategyproof linear regression in high dimensions. In *ACM Conf. on Economics and Computation*.

Cheung, S. N. 1969. *The theory of share tenancy*. Arcadia Press Ltd.

Cummings, R.; Ioannidis, S.; and Ligett, K. 2015. Truthful linear regression. In *Conf. Learning Theory*.

Dalvi, N.; Domingos, P.; Sanghai, S.; Verma, D.; et al. 2004. Adversarial classification. In *Proc. 10th ACM Knowledge discovery and data mining*.

Davis, H. 2006. *Search engine optimization*. "O'Reilly Media, Inc.".

Dekel, O.; Fischer, F.; and Procaccia, A. D. 2010. Incentive compatible regression learning. *Journal of Computer and System Sciences*.

Dong, J.; Roth, A.; Schutzman, Z.; Waggoner, B.; and Wu, Z. S. 2018. Strategic classification from revealed preferences. In *ACM Economics and Computation*.

Foust, D., and Pressman, A. 2008. Credit scores: Not-so-magic numbers. *Business Week* 7.

Grossman, S. J., and Hart, O. D. 1983. An analysis of the principal-agent problem. *Econometrica*.

Hardt, M.; Megiddo, N.; Papadimitriou, C.; and Wootters, M. 2016. Strategic classification. In *Proc. 2016 ACM Innovations in Theoretical Computer Science*.

Hermalin, B. E., and Katz, M. L. 1991. Moral hazard and verifiability: The effects of renegotiation in agency. *Econometrica*.

Holmstrom, B., and Milgrom, P. 1987. Aggregation and linearity in the provision of intertemporal incentives. *Econometrica*.

Holmstrom, B., and Milgrom, P. 1991. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *JL Econ. & Org.*

Hu, L.; Immorlica, N.; and Vaughan, J. W. 2019. The disparate effects of strategic manipulation. In *Fairness, Accountability, and Transparency*.

Jensen, M. C., and Meckling, W. H. 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. *J. financ. econ.*

Kerr, S. 1975. On the folly of rewarding A, while hoping for B. *Academy of Management journal* 18.

Kleinberg, J., and Raghavan, M. 2019. How do classifiers induce agents to invest effort strategically? In *ACM Conf. on Economics and Computation*.

Koretz, D.; Linn, R.; Dunbar, S.; and Shepard, L. 1991. The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests. In *American Educational Research Association and the National Council on Measurement in Education*.

Koretz, D. M. 2008. *Measuring up*. Harvard University Press.

Laffont, J.-J., and Martimort, D. 2009. *The Theory of Incentives: The Principal-Agent Model*. Princeton University Press.

Milli, S.; Miller, J.; Dragan, A. D.; and Hardt, M. 2019. The social cost of strategic classification. In *Fairness, Accountability, and Transparency (FAT*)*.

Pauly, M. V. 1968. The economics of moral hazard: comment. *American Economic Review*.

Ross, S. A. 1973. The economic theory of agency: The principal's problem. *Am. Econ. Rev.*

Stiglitz, J. E. 1974. Incentives and risk sharing in sharecropping. *The Review of Economic Studies* 41.

Zarsky, T. Z. 2007. Law and online social networks: Mapping the challenges and promises of user-generated information flows. *Fordham Intell. Prop. Media & Ent. LJ*.