Bayesian Persuasion for Algorithmic Recourse

Keegan Harris  
Carnegie Mellon University  
keeganh@cmu.edu

Valerie Chen  
Carnegie Mellon University  
valerichen@cmu.edu

Joon Sik Kim  
Carnegie Mellon University  
joonkim@cmu.edu

Ameet Talwalkar  
Carnegie Mellon University  
talwalkar@cmu.edu

Hoda Heidari  
Carnegie Mellon University  
hheidari@cmu.edu

Zhiwei Steven Wu  
Carnegie Mellon University  
zstevenwu@cmu.edu

Abstract

When subjected to automated decision-making, decision subjects may strategically modify their observable features in ways they believe will maximize their chances of receiving a favorable decision. In many practical situations, the underlying assessment rule is deliberately kept secret to avoid gaming and maintain competitive advantage. The resulting opacity forces the decision subjects to rely on incomplete information when making strategic feature modifications. We capture such settings as a game of Bayesian persuasion, in which the decision maker offers a form of recourse to the decision subject by providing them with an action recommendation (or signal) to incentivize them to modify their features in desirable ways. We show that when using persuasion, the decision maker and decision subject are never worse off in expectation, while the decision maker can be significantly better off. While the decision maker’s problem of finding the optimal Bayesian incentive-compatible (BIC) signaling policy takes the form of optimization over infinitely-many variables, we show that this optimization can be cast as a linear program over finitely-many regions of the space of possible assessment rules. While this reformulation simplifies the problem dramatically, solving the linear program requires reasoning about exponentially-many variables, even in relatively simple cases. Motivated by this observation, we provide a polynomial-time approximation scheme that recovers a near-optimal signaling policy. Finally, our numerical simulations on semi-synthetic data empirically demonstrate the benefits of using persuasion in the algorithmic recourse setting.

1 Introduction

High-stakes decision-making systems increasingly utilize data-driven algorithms to assess individuals in such domains as education [31], employment [5][36], and lending [24]. Individuals subjected to these assessments (henceforth, decision subjects) may strategically modify their observable features in ways they believe maximize their chances of receiving favorable decisions [21][9]. The decision subject often has a set of actions/interventions available to them. Each of these actions leads to some measurable effect on their observable features, and subsequently, the decision they receive. From the decision maker’s perspective, some of these actions may be more desirable than others. Consider credit scoring as an example: Credit scores predict how likely an individual applicant...
We study a model of strategic learning in which the underlying assessment rule is not revealed to decision subjects. Before the assessment rule is trained, the decision maker and decision subject have some prior knowledge about the realization of the assessment rule. This prior represents the “common knowledge” about what the rule may be. Secondly, when the decision maker provides recommendations to decision subjects on which action to take, the recommendations should be compatible with the subjects’ incentives to ensure they will follow the recommendation. Finally, our model assumes the decision maker discloses how they generate recommendations for recourse—an increasingly relevant requirement under recent regulations (e.g., [10]).

Utilizing our model, we aim to design a mechanism for a decision maker to provide recourse to a decision subject who has incomplete information about the underlying assessment rule. We assume the assessment rule makes predictions about some future outcome of the decision subject (e.g., whether they will pay back a loan in time if granted one). Before the assessment rule is trained (i.e., before the model parameters are fit), the decision maker and decision subject have some prior belief about the realization of the assessment rule. This prior represents the “common knowledge” about the importance of various observable features for making accurate predictions. After training, the assessment rule is revealed to the decision maker, who then recommends an action for the decision subject to take, based on their pre-determined signaling policy. Upon receiving this action recommendation, the decision subject updates their belief about the underlying assessment rule. They then take the action which they believe will maximize their utility (i.e., the benefit from the decision they receive, minus the cost of taking their selected action) in expectation. Finally, the decision maker uses the assessment rule to make a prediction about the decision subject.

The interaction described above is an instance of Bayesian persuasion, a game-theoretic model of information revelation originally due to Kamenica and Gentzkow. The specific instance of Bayesian persuasion we consider in this work is summarized in Figure \[1\].

---

2Similar logic applies to other examples of strategic settings including: college admissions, in which someone associated with the university may have the ability to offer advice to applicants, but does not have the ability to unilaterally change the underlying assessment rule, or hiring, where a recruiter for a company may have knowledge of the factors the company uses to make hiring decisions, but may not be able to change this criteria or reveal it to job applicants.
Interaction protocol between the decision maker and decision subject

1. The decision maker and decision subject initially have some prior/belief about the assessment rule that will be trained.
2. Before training, the decision maker commits to a signaling policy. After training, the assessment rule is revealed to the decision maker.
3. The decision maker then uses their signaling policy and knowledge of the assessment rule to recommend an action for the decision subject to take.
4. The decision subject updates their belief given the recommendation, and chooses an action that they believe maximizes their utility.
5. The decision subject receives a prediction through the assessment rule.

Figure 1: Summary of the setting we consider.

Our contributions. Our central conceptual contribution is to cast the problem of offering recourse under partial transparency as a game of Bayesian persuasion. Our key technical contributions consist of comparing optimal action recommendation policies in this new setup with two natural alternatives: (1) fully revealing the assessment rule to the decision subjects, or (2) revealing no information at all about the assessment rule. We provide new insights about the potentially significant advantages of action recommendation over these baselines, and offer efficient formulations to derive the optimal recommendations. More specifically, our analysis offers the following takeaways:

1. Using tools from Bayesian persuasion, we show that it is possible for the decision maker to provide incentive-compatible action recommendations that encourage rational decision subjects to modify their features through beneficial interventions. While the decision maker and decision subjects are never worse off in expectation from using optimal incentive-compatible recommendations, we show that situations exist in which the decision maker is significantly better off in expectation utilizing the optimal signaling policy (as opposed to the two baselines, Section 3).
2. We derive the optimal signaling policy for the decision maker. While the decision maker’s optimal signaling policy initially appears challenging to compute (as it involves optimizing over continuously-many variables), we show that the problem can naturally be cast as a linear program defined in terms of a finite set of variables. However, solving this linear program may require reasoning about exponentially-many variables. Motivated by this observation, we provide a polynomial-time algorithm to approximate the optimal signaling policy (Section 4).
3. We empirically evaluate our persuasion mechanism on semi-synthetic data based on the Home Equity Line of Credit (HELOC) dataset, and find that the optimal signaling policy performs significantly better than the two natural alternatives across a wide range of instances (Section 5).

1.1 Related work

Strategic responses to unknown predictive models. To the best of our knowledge, our work is the first to use tools from persuasion to model the strategic interaction between a decision maker and strategic decision subjects when the underlying predictive model is not public knowledge. Several prior works have addressed similar problems through different models and techniques. For example, Akyol et al. quantify the “price of transparency”, a quantity which compares the decision maker’s utility when the predictive model is fully known with their utility when the model is not revealed to the decision subjects. Tsirtsis and Rodriguez study the effects of counterfactual explanations on strategic behavior. Ghalme et al. compare the prediction error of a classifier when it is public knowledge with the error when decision subjects must learn a version of it, and label this difference the “price of opacity”. They show that small errors in decision subjects’ estimates of the true underlying model may lead to large errors in the performance of the model. The authors argue that their work provides formal incentives for decision makers to adopt full transparency as a policy. Our work, in contrast, is based on the observation that even if decision makers are willing to reveal their models, legal requirements, privacy concerns, and intellectual property restrictions may prohibit full transparency. So we instead study the consequences of partial transparency—a common condition in real-world domains.

Bechavod et al. study the effects of information discrepancy across different sub-populations of decision subjects on their ability to improve their observable features in strategic learning settings.
Like us, they do not assume the predictive model is fully known to the decision subjects. Instead, the authors model decision subjects as trying to infer the underlying predictive model by learning from their social circle of family and friends, which naturally causes different groups to form within the population. In contrast to this line of work, we study a setting in which the decision maker provides customized feedback to each decision subject individually. Additionally, while the models proposed by \cite{17, 2} circumvent the assumption of full information about the deployed model, they restrict the decision subjects’ knowledge to be obtained only through past data.

**Algorithmic recourse.** Our work is closely related to recent work on algorithmic recourse \cite{28}. Algorithmic recourse is concerned with providing explanations and recommendations to individuals who have received unfavorable automated decisions. A line of algorithmic recourse methods including \cite{4, 13, 25} focus on suggesting actionable or realistic changes to underlying qualifications to decision subjects interested in improving their decisions. Our action recommendations are “actionable” in the sense that they are interventions which promote long-term desirable behaviors while ensuring that the decision subject is not worse off in expectation.

**Transparency.** Recent legal and regulatory frameworks, such as the General Data Protection Regulation (GDPR) \cite{10}, motivate the development of forms of algorithmic transparency suitable for real-world deployment. While this work can be thought of as providing additional transparency into the decision-making process, it does not naturally fall into the existing organizations of explanation methods (e.g., as outlined in \cite{7}), as our policy does not simply recommend actions based on the decision rule. Rather, our goal is to incentivize actionable interventions on the decision subjects’ observable features which are desirable to the decision maker, and we leverage persuasion techniques to ensure compliance.

**Bayesian persuasion.** There has been growing interest in Bayesian persuasion \cite{27} in the computer science and machine learning communities in recent years. Dughmi and Xu \cite{12, 13} characterize the computational complexity of computing the optimal signaling policy for several popular models of persuasion. Castiglioni et al. \cite{4} study the problem of learning the receiver’s utilities through repeated interactions. Work in the multi-armed bandit literature \cite{34, 35, 32, 36, 39} leverages techniques from Bayesian persuasion to incentivize agents to perform bandit exploration. Finally, linear programming-based approaches to Bayesian persuasion have been studied in the economics literature \cite{30, 14}, although the persuasion setting we study differs considerably.

**Other strategic learning settings.** The strategic learning literature \cite{18, 17, 38, 32, 23, 14, 30, 29, 16} broadly studies machine learning questions in the presence of strategic decision subjects. There has been a long line of work in strategic learning that focuses on how strategic decision subjects adapt their input to a machine learning algorithm in order to receive a more desirable prediction, although most prior work in this literature assumes that the underlying assessment rule is fully revealed to the decision subjects, which is typically not true in reality.

## 2 Setting and background

Consider a setting in which a decision maker assigns a predicted label $\hat{y} \in \{-1, +1\}$ (e.g., whether or not someone will repay a loan if granted one) to a decision subject with initial observable features $x_0 = (x_{0,1}, \ldots, x_{0,d-1}, 1) \in \mathbb{R}^d$ (e.g., amount of current debt, bank account balance, etc.)\footnote{We append a 1 to the decision subject’s feature vector for notational convenience.}. We assume the decision maker uses a fixed linear decision rule to make predictions, i.e., $\hat{y} = \text{sign}\{x_0^\top \theta\}$, where the assessment rule $\theta \in \Theta \subseteq \mathbb{R}^d$. The goal of the decision subject is to receive a positive classification (e.g., get approved for a loan). Given this goal, the decision subject may choose to take some action $a$ from some set of possible actions $\mathcal{A}$ to modify their observable features (for example, they may decide to pay off a certain amount of existing debt, or redistribute their debt to game the credit score). We assume that the decision subject has $m$ actions $\{a_1, a_2, \ldots, a_m\} \in \mathcal{A}$ at their disposal in order to improve their outcomes. For convenience, we add $a_0 \in \mathcal{A}$ to denote taking “no action”. By taking action $a$, the decision subject incurs some cost $c(a) \in \mathbb{R}$. This could be an actual monetary cost, but it can also represent non-monetary notions of cost such as opportunity cost or the time/effort the decision subject may have to exert to take the action. We assume taking an action $a$ changes a decision subject’s observable feature values from $x_0$ to $x_0 + \Delta x(a)$, where $\Delta x(a) \in \mathbb{R}^d$, and $\Delta x_j(a)$ specifies the change in the $j$th observable feature as the result of taking
action $a$. For the special case of $a_0$, we have $\Delta x(a_0) = 0$, $c(a_0) = 0$. As a result of taking action $a$, a decision subject, $ds$, receives utility $u_{ds}(a, \theta) = \text{sign}(x_0 + \Delta x(a)) \cdot (\theta) - c(a)$. In other words, the decision subject receives some positive (negative) utility for a positive (negative) classification, subject to a $\text{cost}$ for taking the action.

If the decision subject had exact knowledge of the assessment rule $\theta$ used by the decision maker, they could solve an optimization problem to determine the best action to take in order to maximize their utility. However, in many settings it is not realistic for a decision subject to have perfect knowledge of $\theta$. Instead, we model the decision subject’s information through a $\text{prior}$ $\Pi$ over $\theta$, which can be thought of as “common knowledge” about the relative importance of various observable features in predicting the outcome of interest. For example, the decision subject may believe that prior payment history would likely be highly correlated with future default. We will use $\pi(\cdot)$ to denote the probability density function of $\Pi$ (so that $\pi(\theta)$ denotes the probability of the deployed assessment rule being $\theta$). We assume the decision subject is rational and risk-neutral. So at any point during the interaction, if they hold a belief $\Pi'$ about the underlying assessment rule, they would pick an action $a^*$ that maximize their $\text{expected}$ utility with respect to that belief. More precisely, they solve $a^* \in \arg\max_{a \in \mathcal{A}} E_{\theta \sim \Pi'}[u_{ds}(a, \theta)]$.

From the decision maker’s perspective, some actions may be more desirable than others. For example, a bank may prefer that an applicant pay off more existing debt than less when applying for a loan. To formalize this notion of action preference, we say that the decision maker $\text{commits}$ to their signaling policy before training the assessment rule. This can be seen as a form of $\text{commitment}$, i.e., the decision subject must believe that the decision maker will select actions according to their signaling policy. In our setting, this will be the case since the decision maker commitment to their signaling policy before training the assessment rule. This can be seen as a form of transparency, as the decision maker is publicly announcing how they will use their assessment rule to provide action recommendations/recourse before they train the assessment rule. For simplicity, we assume that the decision maker shares the same prior beliefs $\Pi$ as the decision subject over the observable features before the model is trained. These assumptions are standard in the Bayesian persuasion literature (see, e.g., [27, 34, 33]).

### 2.1 Bayesian persuasion in the algorithmic recourse setting

The decision maker has an $\text{information advantage}$ over the decision subject, due to the fact that they know the true assessment rule $\theta$, whereas the decision subject does not. The decision maker may be able to leverage this information advantage to incentivize the decision subject to take a more favorable action (compared to the one they would have taken according to their prior) by recommending an action to the decision subject according to a commonly known $\text{signaling policy}$.

**Definition 2.1 (Signaling policy).** A signaling policy $S : \Theta \to \mathcal{A}$ is a (possibly stochastic) mapping from assessment rules to actions.

We use $\sigma \sim S(\theta)$ to denote the action recommendation sampled from signaling policy $S$, where $\sigma \in \mathcal{A}$ is the realized recommended action.

The decision maker’s signaling policy is assumed to be fixed and common knowledge. This is because in order for the decision subject to perform a Bayesian update based on the observed recommendation, they need to know the signaling policy. Additionally, the decision maker must have the $\text{power of commitment}$, i.e., the decision subject must believe that the decision maker will select actions according to their signaling policy. In our setting, this will be the case since the decision maker commits to their signaling policy before training the assessment rule. This can be seen as a form of transparency, as the decision maker is publicly announcing how they will use their assessment rule to provide action recommendations/recourse before they train the assessment rule. For simplicity, we assume that the decision maker shares the same prior beliefs $\Pi$ as the decision subject over the observable features before the model is trained. These assumptions are standard in the Bayesian persuasion literature (see, e.g., [27, 34, 33]).

## In order for the decision subject to be incentivized to follow the actions recommended by the decision maker, the signaling policy $S$ needs to be $\text{Bayesian incentive-compatible}$.  

**Definition 2.2 (Bayesian incentive-compatibility).** Consider a decision subject $ds$ with initial observable features $x_0$ and prior $\Pi$. A signaling policy $S$ is Bayesian incentive-compatible (BIC) for $ds$ if $\mathbb{E}_{\theta \sim \Pi}[u_{ds}(\sigma, \theta) | | \sigma = a] \geq \mathbb{E}_{\theta \sim \Pi}[u_{ds}(a', \theta) | | \sigma = a]$ for all actions $a, a' \in \mathcal{A}$ such that $S(\theta)$ has positive support on $\sigma = a$.
We refer to this region as. We observe that in this simple setting, we can bin values of
x desired action to make a difference in their classification. We refer to this region as
action on the outcome the decision subject would receive if
changes their observable feature from
subject to take. The decision subject then takes a possibly different action
Given the true value of
have some prior over it, denoted by
i.e., the decision subject receives a positive classification if
mutually desirable trait, and credit scores to be a good measure of credit-worthiness. We assume the
advantage may improve their expected utility under our setting, we study the following example.
As is generally the case in the persuasion literature (e.g., credit score) and two possible actions:
values are observed and the signaling policy is computed.
We remark that while for the ease of exposition our model focuses the interactions between the
decision maker and a single decision subject, our results can be extended to a heterogeneous pop-
ulation of decision subjects—as long as we assume their interactions with the decision-maker are
independent of one another (e.g., this assumption rules out one subject updating their belief based on
the outcome of another subject’s prior interaction with the decision-maker). Under such a setting,
the decision maker would publicly commit to a method of computing the signaling policy, given a
decision subject’s initial observable features as input. Once a decision subject arrives, their feature
values are observed and the signaling policy is computed.

### 3 Characterizing the utility gains of persuasion

As is generally the case in the persuasion literature, the decision maker can achieve higher
expected utility with an optimized signaling policy compared to if they provided no recommendation
or fully disclosed the model. To characterize how much leveraging the decision maker’s information
advantage may improve their expected utility under our setting, we study the following example.
Consider a simple setting under which a single decision subject has one observable feature \( x_0 \)
(e.g., credit score) and two possible actions: \( a_0 = \) “do nothing” (i.e., \( \Delta x(a_0) = 0, c(a_0) = 0, u_{dm}(a_0) = 0 \)) and \( a_1 = \) “pay off existing debt” (i.e., \( \Delta x(a_1) > 0, c(a_1) > 0, u_{dm}(a_1) = 1 \)), which
in turn raises their credit score. For the sake of our illustration, we assume credit-worthiness to be a
mutually desirable trait, and credit scores to be a good measure of credit-worthiness. We assume the
decision maker would like to design a signaling policy to maximize the chance of the decision subject
taking action \( a_1 \), regardless of whether or not the applicant will receive the loan. In this simple
setting, the decision maker’s decision rule can be characterized by a single threshold parameter \( \theta \),
i.e., the decision subject receives a positive classification if \( x + \theta \geq 0 \) and a negative classification
otherwise. Note that while the decision subject does not know the exact value of \( \theta \), they instead
have some prior over it, denoted by \( \Pi \).

Given the true value of \( \theta \), the decision maker recommends an action \( \sigma \in \{a_0, a_1\} \) for the decision
subject to take. The decision subject then takes a possibly different action \( a \in \{a_0, a_1\} \), which
changes their observable feature from \( x_0 \) to \( x = x_0 + \Delta x(a) \). Recall that the decision subject’s
utility takes the form \( u_{ds}(a, \theta) = \text{sign}\{(x_0 + \Delta x(a)) + \theta\} - c(a) \). Note that if \( c(a_1) > 2 \), then
\( u_{ds}(a_0, \theta) > u_{ds}(a_1, \theta) \) holds for any value of \( \theta \), meaning that it is impossible to incentivize any
differential subjective to play action \( a_1 \). Therefore, in order to enable the decision maker to
incentivize action \( a_1 \), we assume \( c(a_1) < 2 \).

We observe that in this simple setting, we can bin values of \( \theta \) into three different “regions”, based
on the outcome the decision subject would receive if \( \theta \) were actually in that region. First, if 
\( x_0 + \Delta x(a_1) + \theta < 0 \), the decision subject will not receive a positive classification, even if they take
action \( a_1 \). In this region, the decision subject’s initial feature value \( x_0 \) is “too low” for taking the
desired action to make a difference in their classification. We refer to this region as \( L \). Second, if
\( x_0 + \theta \geq 0 \), the decision subject will receive a positive classification no matter what action they take.
In this region, \( x_0 \) is “too high” for the action they take to make any difference on their classification.
We refer to this region as \( H \). Third, if \( x_0 + \theta < 0 \) and \( x_0 + \Delta x(a_1) + \theta \geq 0 \), the decision subject
will receive a positive classification if they take action \( a_1 \) and a negative classification if they take
action \( a_0 \). We refer to this region as \( M \). Consider the signaling policy in Figure 2.

**Example signaling policy \( S(\theta) \)**

| Case 1: | \( \theta \in L \). Recommend action \( a_1 \) with probability \( q \) and action \( a_0 \) with probability \( 1 - q \) |
| Case 2: | \( \theta \in M \). Recommend action \( a_1 \) with probability \( 1 \) |
| Case 3: | \( \theta \in H \). Recommend action \( a_1 \) with probability \( q \) and action \( a_0 \) with probability \( 1 - q \) |

Figure 2: Signaling policy for the example of Section 3.
where we omit the valid probability constraints over \( \pi(M) \) that the decision subject would have taken had they known the true \( \theta \), with probability 1. However, in Case 1 and Case 3, the decision maker recommends, with probability \( q \), an action \( (a_1) \) that the decision subject would not have taken knowing \( \theta \), leveraging the fact that the decision subject does not know exactly which case they are currently in. If the decision subject follows the decision maker’s recommendation from \( S \), then the decision maker expected utility will increase from 0 to \( q \) if the realized \( \theta \in L \) or \( \theta \in H \), and will remain the same otherwise. Intuitively, if \( q \) is “small enough” (where the precise definition of “small” depends on the prior over \( \theta \) and the cost of taking action \( a_1 \)), then it will be in the decision subject’s best interest to follow the decision maker’s recommendation, even though they know that the decision maker may sometimes recommend taking action \( a_1 \) when it is not in their best interest to take that action. That is, the decision maker may recommend that a decision subject pay off existing debt with probability \( q \) when it is unnecessary for them to do so in order to secure a loan. We now give a criteria on \( q \) which ensures the signaling policy \( S \) is BIC.

**Proposition 3.1.** Signaling policy \( S \) is Bayesian incentive-compatible if \( q = \min\{\pi(M)(2 - \epsilon(a_1)), 1\} \), where \( \pi(M) := \mathbb{P}_{\theta \sim \Pi}(x_0 + \theta < 0 \text{ and } x_0 + \Delta x(a_1) + \theta \geq 0) \).

**Proof Sketch.** We show that \( \mathbb{E}_{\theta \sim \Pi}[u_{ds}(a_\theta, \theta) | \sigma = a_\theta] \geq \mathbb{E}_{\theta \sim \Pi}[u_{ds}(a_1, \theta) | \sigma = a_1] \) and \( \mathbb{E}_{\theta \sim \Pi}[u_{ds}(a_1, \theta) | \sigma = a_1] \geq \mathbb{E}_{\theta \sim \Pi}[u_{ds}(a_\theta, \theta) | \sigma = a_\theta] \). Since these conditions are satisfied, \( S \) is BIC. The full proof may be found in Appendix [C].

Next, we show that the decision maker’s expected utility when recommending actions according to the optimal signaling policy can be arbitrarily higher than their expected utility from revealing full information or no information. We prove the following result in Appendix [D].

**Proposition 3.2.** For any \( \epsilon > 0 \), there exists a problem instance such that the expected decision maker utility from recommending actions according to the optimal signaling policy is \( 1 - \epsilon \) and the expected decision maker utility for revealing full information or revealing no information is at most \( \epsilon \).

## 4 Computing the optimal signaling policy

In Section [D], we show a one-dimensional example, where a signaling policy can obtain arbitrarily better utilities compared to revealing full information and revealing no information. We now derive the decision maker’s optimal signaling policy for the general setting with arbitrary numbers of observable features and actions described in Section [C]. Under the general setting, the decision maker’s optimal signaling policy can be described by the following optimization:

\[
\max_{p(\sigma = a | \theta), a \in A} \mathbb{E}_{\theta \sim S(\theta), \theta \sim \Pi}[u_{dm}(\sigma)]
\]

s.t. \( \mathbb{E}_{\theta \sim \Pi}[u_{ds}(a, \theta) - u_{ds}(a', \theta) | \sigma = a] \geq 0, \forall a, a' \in A \),

where we omit the valid probability constraints over \( p(\sigma = a | \theta), a \in A \) for brevity. In words, the decision maker wants to design a signaling policy \( S \) in order to maximize their expected utility, subject to the constraint that the signaling policy is BIC. At first glance, the optimization may initially seem hopeless as there are infinitely many values of \( p(\sigma = a | \theta) \) (one for every possible \( \theta \in \Theta \)) that the decision maker’s optimal policy must optimize over. However, we will show that the decision maker’s optimal policy can actually be recovered by optimizing over finitely many variables.

By rewriting the BIC constraints as integrals over \( \Theta \) and applying Bayes’ rule, our optimization over \( p(\sigma = a | \theta), a \in A \) takes the following form

\[
\max_{p(\sigma = a | \theta), a \in A} \mathbb{E}_{\theta \sim S(\theta), \theta \sim \Pi}[u_{dm}(\sigma)]
\]

s.t. \( \int_{\Theta} p(\sigma = a | \theta) \pi(\theta)(u_{ds}(a, \theta) - u_{ds}(a', \theta))d\theta \geq 0, \forall a, a' \in A \).

Note that if \( u_{ds}(a, \theta) - u_{ds}(a', \theta) \) is the same for some “equivalence region” \( R \subseteq \Theta \) (which we formally define below), we can pull \( u_{ds}(a, \theta) - u_{ds}(a', \theta) \) out of the integral and instead sum over the different equivalence regions. Intuitively, an equivalence region can be thought of as the set of all \( \theta \in \Theta \) pairs that are indistinguishable from a decision subject’s perspective because they lead to the exact same utility for any possible action the decision subject could take. Based on this idea, we formally define a region of \( \Theta \) as follows.
We leave open the question of whether there are classes of succinctly represented priors that permit

We adapt the sampling-based approximation algorithm of Dughmi and Xu

While the problem of determining the decision maker’s optimal signaling policy can be transformed

which does not require the full prior

Appendix F for more details). This is perhaps unsurprising as without any assumptions on

Theorem 4.3. Algorithm \textbullet \textbullet \textbullet runs in poly(m, \frac{1}{\delta}) time (where m = |A|), and implements an \epsilon-BIC signaling policy that is \epsilon-optimal with probability at least 1 − \delta.

We leave open the question of whether there are classes of succinctly represented priors that permit efficient algorithms for computing the exact optimal policy in time polynomial in \(d \cdot m\). It is also plausible to design efficient algorithms that only require some form of query access to the prior distribution. However, information-theoretic lower bounds of \(\Omega(d \cdot m)\) rule out query access through sampling.

5 Experiments

In this section, we provide experimental results using a semi-synthetic setting where decision subjects are based on individuals in the Home Equity Line of Credit (HELOC) dataset \(\frac{1}{13}\). The HELOC dataset contains information about 9,282 customers who received a Home Equity Line of Credit. Each individual in the dataset has 23 observable features related to an applicant’s financial history (e.g., percentage of previous payments that were delinquent) and a label which characterizes their loan repayment status (repaid/defaulted). We compare the decision maker utility for different models of information revelation: our optimal signaling policy, revealing full information about the model, revealing no information about the model. As our theory predicts, the expected decision maker utility when recommending actions according to the optimal signaling policy either matches or exceeds the
expected utility from revealing full information or no information about the assessment rule across all problem instances. Moreover, the expected decision maker utility from signaling is significantly higher on average. Next, we explore how the decision maker’s expected utility changes when action costs and changes in observable features are varied jointly. Our results are summarized in Figures 3 and 4.

In order to adapt the HELOC dataset to our strategic setting, we select four features and define five hypothetical actions \( \mathcal{A} = \{a_0, a_1, a_2, a_3, a_4\} \) that decision subjects may take in order to improve their observable features. Actions \( \{a_1, a_2, a_3, a_4\} \) result in changes to each of the decision subject’s four observable features, whereas action \( a_0 \) does not. For simplicity, we view actions \( \{a_1, a_2, a_3, a_4\} \) as equally desirable to the decision maker, and assume they are all more desirable than \( a_0 \). Using these four features, we train a logistic regression model that predicts whether an individual is likely to pay back a loan if given one, which will serve as the decision maker’s realized assessment rule. For more information on how we constructed our experiments, see Appendix I.

Given a \( \{(c(a_i), \Delta x(a_i))\}_{i=1} \) instance and information revelation scheme, we calculate the decision maker’s total expected utility by summing their expected utility for each applicant. Figure 3 shows the average total expected decision maker utility across different \( \Delta x(a) \) and cost configurations for priors with varying amounts of uncertainty. See Figure 9 in Appendix I for plots of all instances which were used to generate Figure 3. Across all instances, the optimal signaling policy (red) achieves higher average total utility compared to the other information revelation schemes (blue and green). The difference is further amplified whenever the decision subjects are less certain about the true assessment rule (i.e., when \( \sigma^2 \) is large). Intuitively, this is because the decision maker leverages the decision subjects’ uncertainty about the true assessment rule in order to incentivize them to take desirable actions, and as the uncertainty increases, so does their ability of persuasion.

To better understand how the decision maker’s expected utility changes as a function of \( c(a) \) and \( \Delta x(a) \), we sweep through multiple \( \{(c(a_i), \Delta x(a_i))\}_{i=1}^4 \) tuples on a grid of \( (c(a_i), \Delta x(a_i)) \in \{0, 0.25, 0.5\} \times \{0, 0.5, 1.0\} \) for \( i \in \{1, 2, 3, 4\} \) and measure the effectiveness of the three information revelation schemes. Figure 4 shows the surface of the decision maker utility as a function of \( (c(a_i), \Delta x(a_i)) \) for the optimal signaling policy (red), revealing full information (blue), and revealing nothing (green). This gap increases when the decision subject is less certain about the model being used (higher \( \sigma^2 \)).
6 Conclusion

We investigate the problem of offering algorithmic recourse without requiring full transparency (i.e., revealing the assessment rule). We cast this problem as a game of Bayesian persuasion, and offer several new insights regarding how a decision maker can leverage their information advantage over decision subjects to incentivize mutually beneficial actions. Our stylized model relies on several simplifying assumptions, which suggest important directions for future work:

**Public persuasion.** We assume that the recommendations received by each decision subject are private. However, if a decision subject is given access to recommendations for multiple individuals, it may be possible for them to reconstruct the underlying model. While out of the scope of this work, it would be interesting to study models of public persuasion in the algorithmic recourse setting.

**Beyond linear decision rules.** We focus on settings with linear decision rules and assume all decision subject parameters (e.g., cost function, initial observable features, etc.) are known to the decision maker. We leave it for future work to extend our findings to non-linear decision rules, or settings in which some of the decision subjects’ parameters are unknown to the decision maker.

7 Acknowledgements

KH is supported by a NDSEG Fellowship. ZSW and KH were supported in part by the NSF FAI Award #1939606, a Google Faculty Research Award, a J.P. Morgan Faculty Award, a Facebook Research Award, and a Mozilla Research Grant. AT was supported in part by the National Science Foundation grants IIS1705121, IIS1838017, IIS2046613, IIS2112471, a funding from Meta, Morgan Stanley and Amazon. HH acknowledges support from NSF IIS2040929, a CyLab 2021 grant, and a Meta (Facebook) research award. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any of these funding agencies. KH would like to thank Haifeng Xu for insightful conversations about Dughmi and Xu [13]. KH would also like to thank James Best, Yatong Chen, Jeremy Cohen, Daniel Ngo, Chara Podimata, and Logan Stapleton for helpful suggestions and conversations.
References

[1] E. Akyol, C. Langbort, and T. Basar. Price of transparency in strategic machine learning. CoRR, abs/1610.08210, 2016. URL http://arxiv.org/abs/1610.08210.

[2] Y. Bechavod, C. Podimata, Z. S. Wu, and J. Ziani. Information discrepancy in strategic learning. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, editors, International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, volume 162 of Proceedings of Machine Learning Research, pages 1691–1715. PMLR, 2022. URL https://proceedings.mlr.press/v162/bechavod22a.html.

[3] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika, 39(3/4):324–345, 1952. ISSN 00063444. URL http://www.jstor.org/stable/2334029.

[4] M. Castiglioni, A. Celli, A. Marchesi, and N. Gatti. Online bayesian persuasion. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/ba5451d3c91a0f982f103cdebe249bc78-Abstract.html.

[5] A. Chalfin, O. Danieli, A. Hillis, Z. Jelveh, M. Luca, J. Ludvig, and S. Mullainathan. Productivity and selection of human capital with machine learning. American Economic Review, 106(5):124–27, May 2016. doi: 10.1257/aer.p20161029. URL https://www.aeaweb.org/articles?id=10.1257/aer.p20161029.

[6] B. Chen, P. I. Frazier, and D. Kempe. Incentivizing exploration by heterogeneous users. In S. Bubeck, V. Perchet, and P. Rigollet, editors, Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018, volume 75 of Proceedings of Machine Learning Research, pages 798–818. PMLR, 2018. URL http://proceedings.mlr.press/v75/chen18a.html.

[7] V. Chen, J. Li, J. S. Kim, G. Plumb, and A. Talwalkar. Interpretable machine learning: moving from mythos to diagnostics. Commun. ACM, 65(8):43–50, 2022. doi: 10.1145/3546036. URL https://doi.org/10.1145/3546036.

[8] Y. Chen, J. Wang, and Y. Liu. Strategic classification with a light touch: Learning classifiers that incentivize constructive adaptation, 2021.

[9] D. Citron and F. Pasquale. The scored society: Due process for automated predictions. Washington Law Review, 89:1–33, 03 2014.

[10] Council of European Union. Council regulation (EU) no 679/2016, 2016. URL https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

[11] J. Dong, A. Roth, Z. Schutzman, B. Waggoner, and Z. S. Wu. Strategic classification from revealed preferences. In É. Tardos, E. Elkind, and R. Vohra, editors, Proceedings of the 2018 ACM Conference on Economics and Computation, Ithaca, NY, USA, June 18-22, 2018, pages 55–70. ACM, 2018. doi: 10.1145/3219166.3219193. URL https://doi.org/10.1145/3219166.3219193.

[12] S. Dughmi and H. Xu. Algorithmic persuasion with no externalities. In C. Daskalakis, M. Babai, and H. Moulin, editors, Proceedings of the 2017 ACM Conference on Economics and Computation, EC '17, Cambridge, MA, USA, June 26-30, 2017, pages 351–368. ACM, 2017. doi: 10.1145/3033274.3085152. URL https://doi.org/10.1145/3033274.3085152.

[13] S. Dughmi and H. Xu. Algorithmic bayesian persuasion. SIAM J. Comput., 50(3), 2021. doi: 10.1137/16M1098334. URL https://doi.org/10.1137/16M1098334.

[14] P. Dworczak and G. Martini. The simple economics of optimal persuasion. Journal of Political Economy, 127(5):1993–2048, 2019. doi: 10.1086/701813. URL https://doi.org/10.1086/701813.

[15] FICO. Explainable machine learning challenge. https://community.fico.com/s/explainable-machine-learning-challenge, 2018.

[16] A. Frankel and N. Kartik. Improving Information from Manipulable Data. Journal of the European Economic Association, 20(1):79–115, 06 2021. ISSN 1542-4766. doi: 10.1093/jeea/jvab017. URL https://doi.org/10.1093/jeea/jvab017.
[17] G. Ghalme, V. Nair, I. Eilat, I. Talgam-Cohen, and N. Rosenfeld. Strategic classification in the dark. In M. Meila and T. Zhang, editors, Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of Proceedings of Machine Learning Research, pages 3672–3681. PMLR, 2021. URL http://proceedings.mlr.press/v139/ghalme21a.html

[18] M. Hardt, N. Megiddo, C. H. Papadimitriou, and M. Wootters. Strategic classification. In M. Sudan, editor, Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science, Cambridge, MA, USA, January 14-16, 2016, pages 111–122. ACM, 2016. doi: 10.1145/2840728.2840730. URL https://doi.org/10.1145/2840728.2840730

[19] K. Harris, H. Heidari, and Z. S. Wu. Stateful strategic regression. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, pages 28728–28741, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f1404c2624fa7f2507ba04fd9df5f61-Abstract.html

[20] K. Harris, D. D. T. Ngo, L. Stapleton, H. Heidari, and S. Wu. Strategic instrumental variable regression: Recovering causal relationships from strategic responses. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, editors, International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, volume 162 of Proceedings of Machine Learning Research, pages 8502–8522. PMLR, 2022. URL https://proceedings.mlr.press/v162/harris22a.html

[21] T. Homanoff, R. O’Brien, and A. B. Sussman. Does knowing your fico score change financial behavior? evidence from a field experiment with student loan borrowers. Review of Economics and Statistics, 103(2):236–250, 2021.

[22] N. Immorlica, J. Mao, A. Slivkins, and Z. S. Wu. Bayesian exploration with heterogeneous agents. In L. Liu, R. W. White, A. Mantrach, F. Silvestri, J. J. McAuley, R. Baeza-Yates, and L. Zia, editors, The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019, pages 751–761. ACM, 2019. doi: 10.1145/3308558.3313649. URL https://doi.org/10.1145/3308558.3313649

[23] M. Jagadeesan, C. Mendler-Dünner, and M. Hardt. Alternative microfoundations for strategic classification. In M. Meila and T. Zhang, editors, Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of Proceedings of Machine Learning Research, pages 4687–4697. PMLR, 2021. URL http://proceedings.mlr.press/v139/jagadeesan21a.html

[24] J. Jagtiani and C. Lemieux. The roles of alternative data and machine learning in fintech lending: Evidence from the lendingclub consumer platform. Financial Management, 48(4):1009–1029, 2019. doi: https://doi.org/10.1111/fima.12295. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/fima.12295

[25] S. Joshi, O. Koyejo, W. Vijithbenjaronk, B. Kim, and J. Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. CoRR, abs/1907.09615, 2019. URL http://arxiv.org/abs/1907.09615

[26] E. Kamenica. Bayesian persuasion and information design. Annual Review of Economics, 11(1):249–272, 2019. doi: 10.1146/annurev-economics-080218-025739. URL https://doi.org/10.1146/annurev-economics-080218-025739

[27] E. Kamenica and M. Gentzkow. Bayesian persuasion. American Economic Review, 101(6):2590–2615, October 2011. doi: 10.1257/aer.101.6.2590. URL https://www.aeaweb.org/articles?id=10.1257/aer.101.6.2590

[28] A. Karimi, G. Barthe, B. Schölkopf, and I. Valera. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. CoRR, abs/2010.04050, 2020. URL http://arxiv.org/abs/2010.04050

[29] J. M. Kleinberg and M. Raghavan. How do classifiers induce agents to invest effort strategically? ACM Trans. Economics and Comput., 8(4):19:1–19:23, 2020. doi: 10.1145/3417742. URL https://doi.org/10.1145/3417742
[30] A. Kolotilin. Optimal information disclosure: A linear programming approach. *Theoretical Economics*, 13(2):607–635, 2018. doi: https://doi.org/10.3982/TE1805. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/TE1805.

[31] D. Kučak, V. Juričić, and G. Dambić. Machine learning in education—a survey of current research trends. *Annals of DAAAM & Proceedings*, 29, 2018.

[32] S. Levanon and N. Rosenfeld. Strategic classification made practical. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 6243–6253. PMLR, 2021. URL http://proceedings.mlr.press/v139/levanon21a.html

[33] Y. Mansour, A. Slivkins, V. Syrgkanis, and Z. S. Wu. Bayesian exploration: Incentivizing exploration in bayesian games. In V. Conitzer, D. Bergemann, and Y. Chen, editors, *Proceedings of the 2016 ACM Conference on Economics and Computation, EC ’16, Maastricht, The Netherlands, July 24-28, 2016*, page 661. ACM, 2016. doi: 10.1145/2940716.2940755. URL https://doi.org/10.1145/2940716.2940755.

[34] Y. Mansour, A. Slivkins, and V. Syrgkanis. Bayesian incentive-compatible bandit exploration. *Oper. Res.*, 68(4):1132–1161, 2020. doi: 10.1287/opre.2019.1949. URL https://doi.org/10.1287/opre.2019.1949.

[35] C. McDiarmid. *On the method of bounded differences*, page 148–188. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989. doi: 10.1017/CBO9781107359949.008.

[36] M. Raghavan, S. Barocas, J. M. Kleinberg, and K. Levy. Mitigating bias in algorithmic hiring: evaluating claims and practices. In M. Hildebrandt, C. Castillo, L. E. Celis, S. Ruggieri, L. Taylor, and G. Zanfir-Fortuna, editors, *FAIT* ’20: *Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 469–481. ACM, 2020. doi: 10.1145/3351095.3372828. URL https://doi.org/10.1145/3351095.3372828.

[37] K. Rawal and H. Lakkaraju. Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3ee7730e97c67473a424ccf4f49ab20-Abstract.html.

[38] J. Sánchez-Monedero, L. Dencik, and L. Edwards. What does it mean to ‘solve’ the problem of discrimination in hiring?: social, technical and legal perspectives from the UK on automated hiring systems. In M. Hildebrandt, C. Castillo, L. E. Celis, S. Ruggieri, L. Taylor, and G. Zanfir-Fortuna, editors, *FAIT* ’20: *Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 458–468. ACM, 2020. doi: 10.1145/3351095.3372849. URL https://doi.org/10.1145/3351095.3372849.

[39] M. Sellke and A. Slivkins. The price of incentivizing exploration: A characterization via Thompson sampling and sample complexity. In P. Biró, S. Chawla, and F. Echenique, editors, *EC ’21: The 22nd ACM Conference on Economics and Computation, Budapest, Hungary, July 18-23, 2021*, pages 795–796. ACM, 2021. doi: 10.1145/3465456.3467549. URL https://doi.org/10.1145/3465456.3467549.

[40] Y. Shavit, B. L. Edelman, and B. Axelrod. Causal strategic linear regression. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 8676–8686. PMLR, 2020. URL http://proceedings.mlr.press/v119/shavit20a.html.

[41] S. Tsirtsis and M. G. Rodriguez. Decisions, counterfactual explanations and strategic behavior. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c2ba1bc5a4b239208c8b37b901cd03b363-Abstract.html.

[42] B. Ustun, A. Spangher, and Y. Liu. Actionable recourse in linear classification. In danah boyd and J. H. Morgenstern, editors, *Proceedings of the Conference on Fairness, Accountability, and"
[43] S. Wachter, B. D. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *CoRR*, abs/1711.00399, 2017. URL http://arxiv.org/abs/1711.00399