Incentivizing Compliance with Algorithmic Instruments

Daniel Ngo [*1] **Logan Stapleton** [*1] **Vasilis Syrgkanis** [2] **Zhiwei Steven Wu** [3]

Abstract

Randomized experiments can be susceptible to selection bias due to potential non-compliance by the participants. While much of the existing work has studied compliance as a static behavior, we propose a game-theoretic model to study compliance as dynamic behavior that may change over time. In rounds, a social planner interacts with a sequence of heterogeneous agents who arrive with their unobserved private type that determines both their prior preferences across the actions (e.g., control and treatment) and their baseline rewards without taking any treatment. The planner provides each agent with a randomized recommendation that may alter their beliefs and their action selection. We develop a novel recommendation mechanism that views the planner's recommendation as a form of instrumental variable (IV) that only affects an agents' action selection, but not the observed rewards. We construct such IVs by carefully mapping the history –the interactions between the planner and the previous agents– to a random recommendation. Even though the initial agents may be completely non-compliant, our mechanism can incentivize compliance over time, thereby enabling the estimation of the treatment effect of each treatment, and minimizing the cumulative regret of the planner whose goal is to identify the optimal treatment.

1. Introduction

In many applications, estimating the causal effect of a treatment or intervention is at the heart of a decision-making process. Examples include a study on the effect of a vaccine on immunity, an assessment of the effect of a training program on workers' efficiency, and a evaluation of the effect of a sales campaign on a company's profit. Many studies on causal effects rely on randomized experiments, which randomly assign each individual in a population to a treatment group or a control group and then estimate the causal effects by comparing the outcomes across groups. However, in many real-world domains, participation is voluntary, which can be susceptible to *non-compliance*. For example, people may turn down a vaccine or a drug when they are assigned to receive the treatment (Wright, 1993). Another example is a randomized evaluation of the Job Training Partnership Act (JTPA) training program (Bloom et al., 1997), where only 60 percent of the workers assigned to be trained chose to receive training, while roughly 2 percent of those assigned to the control group chose to receive training. In many cases, non-compliance can cause selection bias: for example, those who choose to receive the drug or vaccine in a randomized trial tend to be healthier, and those who join the training program might may more productive to begin with.

Although non-compliance in randomized experiments has been well studied in many observational studies (see e.g. Angrist and Pischke (2008)), there has been little work that studies and models how compliance varies over time. In reality, however, participants' compliance behaviors may not be static: they may change according to their time-varying beliefs about the treatments. If the outcomes from the previous trials suggest that the treatments are effective, then the participants may become more willing to accept the recommendation. For example, those initially weary about a new vaccine may change their mind once they see others take it without experiencing negative symptoms.[1] Motivated by this observation, this paper studies the design of dynamic trial mechanisms that map history–the observations from previous trials–to a treatment recommendation and gradually incentivize compliance over time.

In this paper, we introduce a game theoretic model to study the dynamic (non)-compliance behavior due to changing beliefs. In our model, there is a collection of treatments such that each treatment j is associated with an unknown treatment effect θ_j. We study an online learning game, in which a set of T myopic agents arrive sequentially over T rounds. Each agent t has a private unobserved *type* u_t,

[*]Equal contribution [1]University of Minnesota [2]Microsoft Research [3]Carnegie Mellon University. ZSW is supported by an NSF FAI Award #1939606. Correspondence to: Logan Stapleton <lstaple99@gmail.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

[1]A recent survey shows that many Americans choose to wait before deciding to receive a COVID-19 vaccine (Hamel et al., 2021).

which determines their heterogeneous prior beliefs about the treatment effects. Each agent's goal is to select a treatment j that maximizes the reward: $\theta_j + g_t^{(u_t)}$, where $g_t^{(u_t)}$ denotes the type-dependent baseline reward (without taking any treatment). We introduce a social planner whose goal is to estimate the effects of underlying treatments and incentive the agents to select the treatment that maximize long-term cumulative reward. Upon the arrival of each agent t, the planner provides the agent with a random treatment recommendation, which is computed by a policy that maps the history of interactions with the previous $(t-1)$ agents. While agent t does not observe the previous history, they form a posterior belief over the treatment effects based on the recommendation and then select the action that maximizes their expected utility.

Under this model, we provide dynamic trial mechanisms that incentivize compliance over time and accurately estimate the treatment effects. The key technical insight is that the planner's random recommendation at each round can be viewed as an *instrument* that is independent of the agent's private type and only influences the observed rewards through the agent's choice of action. By leveraging this observation, we can perform instrumental variable (IV) regression to recover the treatment effects, as long as some of the agents are compliant with the recommendations. To create compliance incentives, our mechanisms leverage techniques from the literature of *incentivizing exploration* (Mansour et al., 2015; Slivkins, 2019). The key idea is *information asymmetry*: since each agent does not directly observe the previous history, the planner has more information. By strategically mapping previous history to instruments, the planner can incentivize agents to explore treatments that are less preferred a-priori.

We first focus on the binary action setting, where each agent can select treatment or control. Then we will extend our results to the k treatments setting in section 6. In the binary setting, we first provide two mechanisms that works with two initial non-compliance situations.

Complete non-compliance. In Section 3, we consider a setting where the planner initially has no information about the treatment effect θ, so all agents are initially non-compliant with the planner's recommendations. We provide Algorithm 1 which first lets initial agents choose their preferred arms, then constructs recommendations that incentivize compliance for some later agents. This enables treatment effect estimation through IV regression.

Partial compliance. In Section 4, we consider a setting where the planner has an initial estimate of the treatment effect θ (that may be obtained by running Algorithm 1), so they can incentivize some agents to comply. We provide Algorithm 2, which can be viewed as the bandit algorithm *active arm elimination* (Even-Dar et al., 2006) which uses IV estimates to compare treatments. Samples collected by Algorithm 1 provide an increasingly accurate estimate $\hat{\theta}$ and incentivize more agents to comply over time.

Regret minimization. In Section 5, we show that if the planner first runs Algorithm 1 to obtain an initial treatment effect estimate $\hat{\theta}$ and then runs Algorithm 2 to amplify compliance, then he can achieve $\tilde{O}(\sqrt{T})$ regret w.r.t. the cumulative reward given by enforcing the optimal action for all agents. We then extend such a regret minimization policy to the setting with k different treatments in Section 6.

Experiments. Lastly, in Section 7, we complement our theoretical results with numerical simulations, which allow us to examine how parameters in agents' prior beliefs influence the convergence rate of our recommendation algorithm.

1.1. Related Work

We design mechanisms which strategically select instruments to incentivize compliance over time, so that we can apply tools from IV regression (Angrist and Krueger, 2001; Angrist and Imbens, 1995; Imbens et al., 1996) to estimate causal effects. Although IV regression is an established tool to estimate causal effects where there is non-compliance in observational studies (see e.g. Bloom et al. (1997); Angrist (2005)), our results deviate significantly from previous works, due to the dynamic nature of our model. In particular, even if all agents are initially non-compliant, our mechanism can still incentivize compliance over time and estimate treatment effects —whereas directly applying standard IV regression at the onset cannot.

Our work draws on techniques from the growing literature of incentivizing exploration (IE) (Kremer et al., 2013; Mansour et al., 2015; 2016; Immorlica et al., 2019; Sellke and Slivkins, 2020), where the goal is also to incentivize myopic agents to explore arms in a multi-armed bandit setting (Auer et al., 2002) using information asymmetry techniques from Bayesian persuasion (Kamenica and Gentzkow, 2011). While our mechanisms are technically similar to those in Mansour et al. (2015), our work differs in several key aspects. First, prior work in IE —including Mansour et al. (2015)— does not capture selection bias and cannot be directly applied in our setting to recover causal effects. The mechanism in Mansour et al. (2015) aims to enforce full compliance (also called *Bayesian incentive-compatibility*) that requires all agents to follow the planner's recommendations: as a result, the mechanism needs to cater to the type of agents that are most difficult to convince. By contrast, our mechanism relies only on the compliance of a partial subset of agents in order to obtain accurate estimates.

There has also been a line of work on mechanisms that

incentivize exploration via payments (Frazier et al., 2014; Chen et al., 2018; Kannan et al., 2017). There are several known disadvantages of such payment mechanisms, including potential high costs and ethical concerns (Groth, 2010). See Slivkins (2017) for a detailed discussion.

Thematically, our work relates to work on "instrument-armed bandits" by Kallus (2018), which also views arm recommendations as instruments. However, the compliance behavior (modeled as a fixed stochastic mapping from instrument to treatments) is static in Kallus (2018): it does not change over time —even if the planner has obtained accurate estimate(s) of the treatment effect(s). By comparison, since all agents eventually become compliant in our setting, we can achieve sublinear regret w.r.t. the best treatment, which is not achievable in a static compliance model.

2. Treatment-Control Model

We study a sequential game between a *social planner* and a sequence of *agents* over T rounds, where T is known to the social planner. We will first focus on the binary setting with a single treatment, and study the more general setting of k treatments in Section 6. In the binary setting, the treatment of interest has unknown effect $\theta \in [-1, 1]$. In each round t, a new agent indexed by t arrives with their *private type* u_t drawn independently from a distribution $\mathcal{U}$ over the set of all private types U. Each agent t has two actions to choose from: taking the treatment (denoted as $x_t = 1$) and not taking the treatment, i.e. the control (denoted as $x_t = 0$). Upon arrival, agent t also receives an action recommendation $z_t \in \{0, 1\}$ from the planner. After selecting an action $x_t \in \{0, 1\}$, agent t receives a reward $y_t \in \mathbb{R}$, given by

$$y_t = \theta x_t + g_t^{(u_t)} \tag{1}$$

where $g_t^{(u_t)}$ denotes the confounding *baseline reward* which depends on the agent's private type u_t; each is drawn from a sub-Gaussian distribution with a sub-Gaussian norm of σ_g. The social planner's goal is to estimate the treatment effect θ and maximize the total expected reward of all T agents.

History and recommendation policy. The interaction between the planner and the agent t is given by the tuple (z_t, x_t, y_t). For each t, let H_t denote the *history* from round 1 to t, i.e. the sequence of interactions between the social planner and the first t agents, such that $H_t := ((z_1, x_1, y_1), \ldots, (z_t, x_t, y_t))$. Before the game starts, the social planner commits to a recommendation policy $\pi = (\pi_t)_{t=1}^T$ where each $\pi_t \colon (\{0,1\}\times\{0,1\}\times\mathbb{R})^{t-1} \to \Delta(\{0,1\})$ is a randomized mapping from the history H_{t-1} to recommendation z_t. Policy π is fully known to all agents.

Beliefs, incentives, and action choices. Each agent t knows their place t in the sequential game, and their private type u_t determines a *prior belief* $\mathcal{P}^{(u_t)}$, which is a joint distribution over the treatment effect θ and noisy error term $g^{(u)}$. Agent t selects action x_t as such:

$$x_t := \mathbb{1}\left[\mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)},\pi_t}[\theta \mid z_t, t] > 0\right]. \tag{2}$$

An agent t is *compliant* with a recommendation z_t if the agent chooses the recommended action, i.e. $x_t = z_t$. We'll also say that a recommendation is *compliant* if $x_t = z_t$.

2.1. Recommendations as Instruments

Unlike the standard multi-armed bandit and previous models on incentivizing exploration (Mansour et al., 2015; 2016), the heterogeneous beliefs in our setting can lead to selection bias. For example, agents who are willing to take the treatment may also have higher baseline rewards. Thus, simply comparing rewards across the treatment group ($x = 1$) and the control group ($x = 0$) will lead to a biased estimate of θ. To overcome this selection bias, we will view the planner's recommendations as instruments and perform *instrumental variable (IV) regression* to estimate θ. There are two criteria for z_t to be a valid instrument: (1) z_t influences the selection x_t, and (2) z_t is independent from the noisy baseline reward $g^{(u)}$. Our goal is to design a recommendation policy to meet criterion (1). Criterion (2) follows because planner chooses z_t randomly, independent of the type u_t.

Wald Estimator. Our mechanism periodically solves the following IV regression problem: given a set S of n observations $(x_i, y_i, z_i)_{i=1}^n$, compute an estimate $\hat{\theta}_S$ of θ. We consider the following two-stage least square (2SLS) or Wald estimator (which are equivalent for binary treatments):

$$\hat{\theta}_S = \frac{\sum_{i=1}^n (y_i - \bar{y})(z_i - \bar{z})}{\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})}, \tag{3}$$

where $\bar{x}, \bar{y}, \bar{z}$ denote the empirical means of variables x_i, y_i, and z_i respectively.

While existing work on IV regression mostly focuses on asymptotic analyses, we provide a high-probability finite-sample error bound for $\hat{\theta}_S$, which is required by our regret analysis and may be of independent interest.

Theorem 2.1 (Finite-sample error bound for Wald estimator)**.** *Let $z_1, z_2, \ldots, z_n \in \{0, 1\}$ be a sequence of instruments. Suppose there is a sequence of n agents such that each agent i has their private type u_i drawn independently from $\mathcal{U}$, selects action x_i under instrument z_i, and receives reward y_i. Let sample set $S = (x_i, y_i, z_i)_{i=1}^n$. Let $A : (\{0,1\}^n \times \{0,1\}^n \times \mathbb{R}^n) \to \mathbb{R}$ denote the approximation bound for set S, such that*

$$A(S, \delta) := \frac{2\sigma_g\sqrt{2n\log(2/\delta)}}{|\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})|}$$

and the Wald estimator given by (3) *satisfies*

$$\left|\hat{\theta}_S - \theta\right| \leq A(S, \delta)$$

with probability at least $1 - \delta$*, for any* $\delta \in (0, 1)$.

Proof Sketch. See Appendix B for the full proof. The bound follows by substituting our expressions for y_t, x_t into the IV regression estimator, applying the Cauchy-Schwarz inequality to split the bound into two terms (one dependent on $\{(g_t, z_t)\}_{t=1}^{|S|}$ and one dependent on $\{(x_t, z_t)\}_{t=1}^{|S|}$), and bound the second term with high probability.

Note that the error rate above depends on the covariance between the instruments z and action choices x. In particular, when $\sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})$ is linear in n, the error rate becomes $\tilde{O}(1/\sqrt{n})$. In the following sections, we will provide mechanisms that incentivize compliance so that the instruments z become correlated with actions x, enabling us to achieve such an error rate.

3. Overcoming Complete Non-Compliance

In this section, we present a recommendation policy that incentivizes compliance to enable IV estimation. We focus on a setting where the agents are initially completely non-compliant: since the planner has no information about the treatment effect in the initial rounds, the recommendations have no influence on agents' action selections. For simplicity of exposition, we will present our policy in a setting where there are two types of agents who are initially "always-takers" and "never-takers." As we show later Section 6, this assumption can be relaxed to have arbitrarily many types and also allow all types to be "always-takers."

Formally, consider two types of agents $i \in \{0, 1\}$. For type i, let p_i be the fraction of agents in the population, $\mathcal{P}^{(i)}$ the prior beliefs, and $g^{(i)}$ the baseline reward random variables. Agents of type 1 initially prefer the treatment and type 0 agents prefer control: their prior means for θ satisfy $\mu^{(1)} = \mathbb{E}_{\mathcal{P}^{(1)}}[\theta] > 0$ and $\mu^{(0)} = \mathbb{E}_{\mathcal{P}^{(0)}}[\theta] < 0$.

Our policy (Algorithm 1) splits into two stages. In the first stage, agents take their preferred action according to their prior beliefs: type 0 agents choose control and type 1 treatment. This allows us to collect ℓ_0 and ℓ_1 observations of rewards for $x = 0$ and $x = 1$, respectively. Let $\bar{y}^0$ and $\bar{y}^1$ denote the empirical average rewards for the two actions, respectively. Note that since the baseline rewards $g^{(u)}$ are correlated with the selections x, the difference $(\bar{y}^1 - \bar{y}^0)$ is a biased estimate for θ.

In the second stage, we use this initial set of reward observations to construct valid instruments which incentivize agents of one of the two types to follow both control and treatment recommendations. Without loss of generality, we focus on incentivizing compliance among type 0 agents. Since they already prefer control, the primary difficulty here is to incentivize type 0 agents to comply with treatment recommendations.[2] We leverage the following observation: according to the prior $\mathcal{P}^{(0)}$ of type 0 agents, there is a non-zero probability that the biased estimate $(\bar{y}^1 - \bar{y}^0)$ is so large that θ must be positive.

Formally, consider the following event for the average rewards $\bar{y}^0$ and $\bar{y}^1$:

$$\xi = \left\{\bar{y}^1 > \bar{y}^0 + \sigma_g\left(\sqrt{\frac{2\log(2/\delta)}{\ell_0}} + \sqrt{\frac{2\log(2/\delta)}{\ell_1}}\right) + G^{(0)} + \frac{1}{2}\right\} \quad (4)$$

where $G^{(0)}$ is a constant such that $G^{(0)} > \mathbb{E}_{\mathcal{P}^{(0)}}[g^{(1)} - g^{(0)}]$ and σ_g is the variance parameter for $g^{(0)}$ and $g^{(1)}$.

Assumption 3.1 (Knowledge Assumption for Algorithm 1)**.** *Within Section 3, the following are common knowledge among agents and planner:*[3]

1. *Type 0 agents prefer control and type 1 agents prefer treatment. The fraction of agents of type 0 in the population is* $p_0 \geq 0$ *and the fraction of type 1 is* $p_1 > 0$.
2. *Type 0's prior treatment effect mean* $\mu^{(0)}$ *and the probability of event* ξ*, denoted* $\mathbb{P}_{\mathcal{P}^{(0)}}[\xi]$*, over the prior* $\mathcal{P}^{(0)}$ *of type 0.*[4]

We prove that Algorithm 1 is compliant for agents of type 0 as long as the exploration probability ρ is less than some constant that depends on prior $\mathcal{P}^{(0)}$. When an agent of type 0 is recommended treatment, they do not know whether this is due to exploration or exploitation. However, with small enough ρ, their expected gain from exploiting exceeds the expected loss from exploring. Hence, the agents comply with the recommendation and take treatment.

Lemma 3.2 (Type 0 compliance with Algorithm 1)**.** *Under Assumption 3.1, any type 0 agent who arrives in the last* ℓ *rounds of Algorithm 1 is compliant with any recommendation, as long as the exploration probability* ρ *satisfies*

$$\rho \leq 1 + \frac{4\mu^{(0)}}{\mathbb{P}_{\mathcal{P}^{(0)}}[\xi] - 4\mu^{(0)}} \quad (5)$$

where the event ξ *is defined above in Equation* (4)*.*

Proof Sketch. See Appendix C for the full proof. The proof follows by expressing the compliance condition for type 0

[2] We could instead incentivize type 1 agents to take control. This would require 1) rewriting event ξ so it indicates that the expectation of θ over $\mathcal{P}^{(1)}$ must be negative and 2) rewriting Algorithm 1 so that control is recommended when exploring. We cannot incentivize both types to comply at the same time.

[3] Assumptions do not hold elsewhere, unless explicitly stated.

[4] These assumptions (as well as Assumption 4.1 and Assumption 5.1) require only partial knowledge of the priors for compliant agents only. They are no more restrictive than the least restrictive (detail-free) assumptions of Mansour et al. (2015).

Algorithm 1 Overcoming complete non-compliance

Input: exploration probability $\rho \in (0,1)$, $\ell \in \mathbb{N}$ (assume w.l.o.g. $\rho\ell \in \mathbb{N}$), minimum first stage samples $\ell_0, \ell_1 \in \mathbb{N}$, and failure probability $\delta < \mathbb{P}_{\mathcal{P}_0}[\xi]/8$

1st stage: The first $2\max(\ell_0/p_0, \ell_1/p_1)$ agents are given no recommendation (they choose what they prefer)

2nd stage: Based on at least ℓ_0 control and ℓ_1 treatment samples collected in the first stage:

if $\bar{y}^1 > \bar{y}^0 + \sigma_g\left(\sqrt{\frac{2\log(2/\delta)}{\ell_0}} + \sqrt{\frac{2\log(2/\delta)}{\ell_1}}\right) + G^{(0)} + \frac{1}{2}$ **then**

  $a^* = 1$

else

  $a^* = 0$

end if

From the next ℓ agents, pick $\rho\ell$ agents uniformly at random to be in the explore set E

for the next ℓ rounds **do**

  if agent t is in explore set E **then**

    $z_t = 1$

  else

    $z_t = a^*$

  end if

end for

agents as different cases, depending on the recommendation. By keeping the exploration probability ρ small with regard to type 0 agent's prior-dependent probability $\mathbb{P}_{\mathcal{P}^{(0)}}[\xi]$ and the conditional expected treatment effect $\mathbb{E}_{\mathcal{P}^{(0)}}[\theta|\xi]$, the expected gain from exploiting is greater than the expected loss from exploring. Hence, type 0 agents would comply with the recommendation. We further simplify the condition on exploration probability ρ by applying high probability bound on the samples collected from the 1st stage (where no recommendations were given).

We also provide a separate accuracy guarantee for the treatment effect estimate $\hat{\theta}$ at the end of the Algorithm 1.

Theorem 3.3 (Treatment Effect Confidence Interval after Algorithm 1)**.** *With sample set $S_\ell = (x_i, y_i, z_i)_{i=1}^{\ell}$ of ℓ samples collected from the second stage of Algorithm 1 — run with exploration probability ρ small enough so that type 0 agents are compliant (see Lemma 3.2),— approximation bound $A(S_\ell, \delta)$ satisfies the following, with probability at least $1-\delta$:*

$$A(S_\ell, \delta) \leq \frac{2\sigma_g\sqrt{2\log(5/\delta)}}{\rho(1-\rho)p_0\sqrt{\ell} - (3-\rho)\sqrt{\frac{\rho\log(5/\delta)}{2(1-\rho)}}}$$

for any $\delta \in (0,1)$. Recall σ_g is the variance of $g^{(u_i)}$, p_0 is the fraction of compliant never-takers in the population of agents,[5] and $A(S_\ell, \delta)$ is defined as in Theorem 2.1.

Proof Sketch. See Appendix C for the full proof. Note that Theorem 2.1 applies, so we only have to bound the denominator term which is dependent on $\{(x_t, z_t)\}_{t=1}^{|S_\ell|}$. We assume that Algorithm 1 is initialized with parameters (see Lemma 3.2) such that type 0 agent is compliant. We bound the term dependent on $\{(x_t, z_t)\}_{t=1}^{|S_\ell|}$ with high probability.

3.1. Algorithm 1 Extensions

Algorithm 1 can be extended to handle more general settings:

1. There can be arbitrarily many types of agents that do not share the same prior. In this case, let $\mathbb{E}_{\mathcal{P}^{(u)}}[g^0]$ and $\mathbb{E}_{\mathcal{P}^{(u)}}[g^1]$ denote the expected baseline rewards for never-takers and always-takers, respectively, over the prior $\mathcal{P}^{(u)}$ of any type u and $G^{(u)} > \mathbb{E}_{\mathcal{P}^{(u)}}[g^1 - g^0]$. Then, Algorithm 1 can still incentivize any never-taker type u agents to comply as long as the planner has a lower bound on $\mathbb{P}_{\mathcal{P}^{(u)}}[\xi^{(u)}]$, where $\xi^{(u)}$ is defined just as ξ in Equation (4), except $G^{(0)}$ is replaced with $G^{(u)}$. Theorem 3.3 applies as is.
2. All types can be always-takers (who prefer the treatment). The algorithm can incentivize some of the agents to take control with an event ξ defined without $\bar{y}^0$ and flipped (i.e. the mean treatment reward is much lower than the expected baseline reward).[6]

By Theorem 3.3, samples collected from Algorithm 1 produce a confidence interval on the treatment effect θ which decreases proportionally to $1/\sqrt{t}$ by round t. However, it still decreases slowly because the exploration probability ρ is small (see roughly how small in Section 7). In Section 4, we give an algorithm for which this confidence interval improves quicker and works for arbitrarily many types.

4. Overcoming Partial Non-Compliance

In this section, we present a recommendation policy which (1) capitalizes on partial compliance, eventually incentivizing all agents to comply, and (2) determines whether the treatment effect is positive or not (with high probability). Algorithm 2 recommends control and treatment sequentially (one after the other). Lemma 4.2 gives conditions for partial compliance from the beginning of Algorithm 2, given access to initial samples which form a crude estimate of the treatment effect. Theorem 4.3 demonstrates how rapidly this estimate improves throughout Algorithm 2, which solely depends on the fraction of compliant agents (and not on some fraction like ρ with Algorithm 1). More (and eventually all) types of agents progressively become compliant throughout Algorithm 2.

Assumption 4.1 (Knowledge Assumption for Algorithm 2)**.**

[5] We redefine p_0 here to be applicable to more general settings.

[6] Also, Lemma C.3 can be proved sans clean event C_0.

Within Section 4, the following are common knowledge among agents and planner:

1. *The fraction of agents in the population who prefer control is $p_0 \geq 0$; that who prefer treatment is $p_1 \geq 0$.*
2. *For each type u and for some τ (which can differ per u), the probability $\tau\,\mathbb{P}_{\mathcal{P}^{(u)}}[\theta > \tau]$ is known if $\mathbb{E}_{\mathcal{P}^{(u)}}[\theta] < 0$; or $\tau\,\mathbb{P}_{\mathcal{P}^{(u)}}[\theta < -\tau]$ is known if $\mathbb{E}_{\mathcal{P}^{(u)}}[\theta] \geq 0$.*

Algorithm 2 Overcoming partial compliance

Input: samples $S_0 := (x_i, z_i, y_i)_{i=1}^{|S_0|}$ which meet Theorem 2.1 conditions and produce IV estimate $\hat{\theta}_{S_0}$;[7] time horizon T; number of recommendations of each action per phase h; approximation bound failure probability δ;
Split the remaining rounds (up to T) into consecutive phases of h rounds each, starting with $q = 1$;
Let $\hat{\theta}_0 = \hat{\theta}_{S_0}$ and $A_0 = A(S_0, \delta)$;
while $|\hat{\theta}_{q-1}| \leq A_{q-1}$ **do**
  The next $2h$ agents are recommended control and treatment sequentially (one after the other);
  Let S_q be samples up to and including phase q, i.e. $S_q := (x_i, z_i, y_i)_{i=1}^{|S_0|+hq} = S_{q-1} + \{\text{round } q \text{ samples}\}$
  Let S_q^{BEST} be the sample set with the smallest approximation bound so far (from phase 1 to q), i.e. $S_q^{\text{BEST}} = \text{argmin}_{S_r, 0 \leq r \leq q} A(S_r, \delta)$;
  Define $\hat{\theta}_q = \hat{\theta}_{S_q^{\text{BEST}}}$ and $A_q = A(S_q^{\text{BEST}}, \delta)$;
  $q = q + 1$;
end while
For all remaining agents recommend $a^* = \mathbb{1}\left[\hat{\theta}_q > 0\right]$.

We focus on a setting where agents are assumed to have been at least partially compliant in the past, such that we may form an IV estimate from the history. The social planner employs Algorithm 2, which is a modification of the *Active Arms Elimination* algorithm (Even-Dar et al., 2006). Treatment and control "race", i.e. are recommended sequentially, until the expected treatment effect is known to be negative or positive (with high probability). Then, the algorithm recommends the "winner" (the action with higher expected reward) for the remainder of the time horizon T.

The compliance incentive works as such: when an agent is given a recommendation, they do not know whether it is because the action is still in the "race" or if the action is the "winner". When the algorithm is initialized with samples that form an IV estimate which is sufficiently close to the true treatment effect (according to the agent's prior), then the probability that any recommended action has "won" is high enough such that the agent's expected gain from taking a "winning" action outweighs the expected loss from taking a "racing" one. We formalize this in Lemma 4.2.

[7] Operator $|\cdot|$ denotes the cardinality of a set.

Lemma 4.2 (Algorithm 2 Partial Compliance)**.** *Recall that Algorithm 2 is initialized with input samples $S_0 = (x_i, y_i, z_i)_{i=1}^{|S_0|}$. For any type u with the following prior preference (control or treatment), if S_0 satisfies the following condition, with probability at least $1 - \delta$, then all agents of type u will comply with recommendations of Algorithm 2:*

$$A(S_0, \delta) \leq \begin{cases} \tau\,\mathbb{P}_{\mathcal{P}^{(u)}}[\theta > \tau]/4 & \text{if } \mathbb{E}_{\mathcal{P}^{(u)}}[\theta] < 0; \\ \tau\,\mathbb{P}_{\mathcal{P}^{(u)}}[\theta < -\tau]/4 & \text{if } \mathbb{E}_{\mathcal{P}^{(u)}}[\theta] \geq 0, \end{cases}$$

for some $\tau \in (0, 1)$, where $A(S_0, \delta)$ is the approximation bound for S_0 and any $\delta \in (0, 1)$ (see Theorem 2.1).

Proof Sketch. See Appendix D.1 for the full proof. The proof follows by using a "clean event" analysis where the IV estimated treatment effect $\hat{\theta}$ is close to the true treatment effect θ. We split the conditional expected treatment effect $\mathbb{E}_{\mathcal{P}^{(u)}}[\theta]$ into different cases for the value of θ. With an IV estimate that is sufficiently close to the true treatment effect, the expected gain from exploiting (taking the "winning" action) is greater than the expected loss from exploring (taking a recommended action when the "race" is not over) and the agent of type u would comply with recommendation.

When a nonzero fraction of agents comply from the beginning, the samples gathered in Algorithm 2 provide treatment effect estimates $\hat{\theta}$ which become increasingly accurate over rounds. In the following Theorem 4.3, we provide a high probability guarantee on this accuracy.

Theorem 4.3 (Treatment Effect Confidence Interval from Algorithm 2 with Partial Compliance)**.** *With set $S = (x_i, y_i, z_i)_{i=1}^{|S|}$ of $|S|$ samples collected from Algorithm 2 where p_c is the fraction of compliant agents in the population, we form an estimate $\hat{\theta}_S$ of the treatment effect θ. With probability at least $1 - \delta$,*

$$\left|\hat{\theta}_S - \theta\right| \leq \frac{8\sigma_g\sqrt{2\log(5/\delta)}}{p_c\sqrt{|S|} - \sqrt{50\log(5/\delta)}}$$

for any $\delta \in (0, 1)$, where σ_g is the variance of $g^{(u_i)}$.

Proof Sketch. See Appendix D.3.1 for a full proof. Note that Theorem 2.1 applies, so we only to have to bound the denominator term which is dependent on $\{(x_t, z_t)\}_{t=1}^{|S|}$. We assume that Algorithm 2 is initialized with parameters such that $p_c > 0$ fraction of the population complies with all recommendations. We bound the term dependent on $\{(x_t, z_t)\}_{t=1}^{|S|}$ with high probability.

Agents become compliant during Algorithm 2 for the same reason others become compliant from the beginning: they expect that the estimate $\hat{\theta}$ is sufficiently accurate and it's likely they're getting recommended an action because it won the race. For large enough T, all agents will become compliant.[8] Note that the accuracy improvement in Theorem 4.3 relies solely on the proportion of agents p_c who

comply from the beginning of Algorithm 2, which relies on the accuracy of the approximation bound given by initial samples S_0. Thus, if the social planner can choose more accurate S_0, then the treatment effect estimate $\hat{\theta}$ given by samples from Algorithm 2 becomes more accurate quicker. In Section 5, we present a recommendation policy in which S_0 can be chosen by running Algorithm 1.

5. Combined Recommendation Policy

In this section, we present a recommendation policy π_c, which spans T rounds and runs Algorithms 1 and 2 in sequence. This policy achieves $\tilde{O}(\sqrt{T})$ regret for sufficiently large T and produces an estimate $\hat{\theta}$ which deviates from the true treatment effect θ by $O(1/\sqrt{T})$.[9]

Assumption 5.1 (Knowledge Assumption for Policy π_c)**.** *Within Sections 5.1 and 5.2, the following are common knowledge among agents and planner:*

1. *All prior-dependent constants given in Assumption 4.1*
2. *For each type u which prefers control, prior mean $\mu^{(u)}$ and a lower bound on the probability $\mathbb{P}_{\mathcal{P}^{(u)}}[\xi^{(u)}]$ (defined in Extension 1 of Algorithm 1 from Section 3.1)*

5.1. Recommendation Policy π_c

Recommendation policy π_c over T rounds is given as such:

1) Run Algorithm 1 with exploration probability ρ set to incentivize at least $p_{c_1} > 0$ fraction of agents of the population who initially prefer control to comply in Algorithm 1 and ℓ to make at least $p_{c_2} > 0$ fraction of agents comply in Algorithm 2 (see Lemma 5.2).

2) Initialize Algorithm 2 with samples from Algorithm 1. At least p_{c_2} fraction of agents comply in Algorithm 2.

We first provide conditions on ℓ to define policy π_c.

Lemma 5.2 (Lower bound on ℓ for Type u Compliance in Algorithm 2)**.** *Recall that S_ℓ denotes the samples collected from the second stage of Algorithm 1. Let S_ℓ be the input samples S_0 in Algorithm 2. Assume that p_{c_1} proportion of agents in the population are compliant with recommendations of Algorithm 1 and length ℓ satisfies:*

$$\ell \geq \begin{cases} \left(\frac{\kappa_1}{\tau\, \mathbb{P}_{\mathcal{P}^{(u)}}[\theta > \tau]} + \kappa_2\right)^2 & \text{if } \mathop{\mathbb{E}}_{\mathcal{P}^{(u)}}[\theta] < 0 \\ \left(\frac{\kappa_1}{\tau\, \mathbb{P}_{\mathcal{P}^{(u)}}[\theta < -\tau]} + \kappa_2\right)^2 & \text{if } \mathop{\mathbb{E}}_{\mathcal{P}^{(u)}}[\theta] \geq 0 \end{cases} \tag{6}$$

for some $\tau \in (0,1)$ and where $\kappa_1 := \frac{8\sigma_g\sqrt{2\log(5/\delta)}}{p_{c_1}\rho(1-\rho)}$ and $\kappa_2 := (3-\rho)\sqrt{\frac{\rho\log(5/\delta)}{2(1-\rho)}}$ for any $\delta \in (0,1)$. Then any agent of type u will comply with recommendations of Algorithm 2.

[8] See Lemma D.1 for details.

[9] We spare the reader the details of the exact bound. It can be deduced via Theorems 3.3, 4.3 and D.2 and Lemmas 4.2 and D.1.

Proof Sketch. See Appendix D.3 for the full proof. The proof follows by substituting the value of ℓ into the approximation bound Theorem 3.3 and simplifying. The compliance condition follows from Lemma 4.2.

Policy π_c shifts from Algorithm 1 to Algorithm 2 as soon as the condition on ℓ above is satisfied. This is because 1) the treatment effect estimate $\hat{\theta}$ get more accurate quicker and 2) less regret is accumulated in Algorithm 2 than Algorithm 1.

5.2. Regret Analysis

The goal of recommendation policy π_c is to maximize the cumulative reward of all agents. We measure the policy's performance through *regret*. We are interested in minimizing regret, which is specific to the treatment effect θ. Since agents' priors are not exactly known to the social planner, this pseudo-regret is correct for any realization of these priors and treatment effect θ.

Definition 5.3. [Pseudo-regret] The pseudo-regret of a recommendation policy is given as such:

$$R_\theta(T) = T\max(\theta, 0) - \sum_{t=1}^{T} \theta x_t \tag{7}$$

We present regret guarantees for recommendation policy π_c. First, policy π_c achieves sub-linear pseudo-regret.

Lemma 5.4 (Pseudo-regret)**.** *The pseudo-regret accumulated from policy π_c is bounded for any $\theta \in [-1,1]$ as follows, with probability at least $1-\delta$ for any $\delta \in (0,1)$:*

$$R_\theta(T) \leq L_1 + O(\sqrt{T\log(T/\delta)}) \tag{8}$$

for sufficiently large time horizon T, where the length of Algorithm 1 is $L_1 = \ell + 2\max\left(\frac{\ell_0}{p_0}, \frac{\ell_1}{p_1}\right)$.

Proof Sketch. See Appendix E.1 for the full proof. The proof follows by observing that Algorithm 2 must end after some $\log(T)$ phases. We can bound the regret of the policy π_c by at most that of Algorithm 2 plus θ per each round of Algorithm 1, or alternatively, we can upper bound it by θ per each round of the policy π_c.

Policy π_c also achieves sub-linear regret, where the expectation is over the randomness in the priors of the agents. Lemma 5.5 provides a basic performance guarantee of our recommendation policy.

Lemma 5.5 (Regret)**.** *Policy π_c achieves regret as follows:*

$$\mathbb{E}[R(T)] = O(\sqrt{T\log(T)}) \tag{9}$$

for sufficiently large time horizon T.

Proof Sketch. See Appendix E.2 for the full proof. The proof follows by observing that we can set the parameters in

Algorithm 1 and Algorithm 2 in terms of the time horizon T while maintaining compliance throughout policy π_c.

These results are comparable to the pseudo-regret of the classic multi-armed bandit problem, with some added constants factors for the compliance constraints (Even-Dar et al., 2006). The pseudo-regret of our policy π_c is asymptotically equivalent to an extension of the detail-free recommendation algorithm of (Mansour et al., 2015), which incentivizes full compliance for all types. However, our policy can finish in a more timely manner and has smaller prior-dependent constants in the asymptotic bound.

In Section 6, we provide an extension of our model and policy π_c to arbitrarily many treatments with unknown effects. We also provide similar regret guarantees.

6. Many Treatments with Unknown Effects

In this section, we introduce a setting which extends the previous binary treatment setting by considering k treatments (and no control). We now consider a treatment effect vector $\theta \in \mathbb{R}^k$; and $x, z \in \{0,1\}^k$ are one-hot encodings of the treatment choice and recommendation, respectively. We assume that $\mathbb{E}[g^{(u_i)}] = 0$.[10] All other terms are defined similar to those in Section 2. Here, the reward $y_i \in \mathbb{R}$ and action choice $x_i \in \{0,1\}^k$ at round i are given as such:[11]

$$\begin{cases} y_i = \langle \theta, x_i \rangle + g^{(u_i)} \\ x_i = \underset{1 \le j \le k}{\operatorname{argmax}} \left(\mathbb{E}_{\mathcal{P}^{(u_i)}, \pi_i}[\theta^j | z_i, i] \right) \end{cases} \tag{10}$$

Given sample set $S = (z_i, x_i, y_i)_{i=1}^n$, we compute IV estimate $\hat{\theta}_S$ of θ as such:

$$\hat{\theta}_S = \left(\sum_{i=1}^n z_i x_i^\intercal \right)^{-1} \sum_{i=1}^n z_i y_i \tag{11}$$

Next, we state finite sample approximation results which extend Theorem 2.1 to this general setting.

Theorem 6.1 (Many Treatments Effect Approximation Bound)**.** *Let $z_1, \dots, z_n \in \{0,1\}^k$ be a sequence of instruments. Suppose there is a sequence of n agents such that each agent i has private type u_i drawn independently from $\mathcal{U}$, selects x_i under instrument z_i and receives reward y_i. Let sample set $S = (x_i, y_i, z_i)_{i=1}^n$. The approximation bound $A(S, \delta)$ is given as such:*[12]

$$A(S,\delta) = \frac{\sigma_g \sqrt{2nk \log(k/\delta)}}{\sigma_{\min}\left(\sum_{i=1}^n z_i x_i^\intercal\right)},$$

and the IV estimator given by Equation (11) satisfies

$$\left\| \hat{\theta}_S - \theta \right\|_2 \le A(S,\delta)$$

with probability at least $1-\delta$ for any $\delta \in (0,1)$.

Proof Sketch. See Appendix F.3 for the full proof. The bound follows by substituting our expressions for y_t, x_t into the IV regression estimator, applying the Cauchy-Schwarz inequality to split the bound into two terms (one dependent on $\{(g_t^{(u_t)}, z_t)\}_{t=1}^{|S|}$ and one dependent on $\{(x_t, z_t)\}_{t=1}^{|S|}$). We bound the second term with high probability.

Next, we extend recommendation policy π_c to k treatments (see Definition F.5 in Appendix F for details).[13]

Assumption 6.2 (Knowledge Assumption for General Policy π_c)**.** *Within Section 6, the following are common knowledge among agents and planner:*

1. *All agents share a preference ordering over all k treatments, i.e. for any type u, the prior expected reward $\mathbb{E}_{\mathcal{P}^{(u)}}[\theta^1] > \mathbb{E}_{\mathcal{P}^{(u)}}[\theta^2] > \cdots > \mathbb{E}_{\mathcal{P}^{(u)}}[\theta^k]$.*[14]
2. *Prior-dependent constants $\mathbb{P}_{\mathcal{P}^{(u)}}[\xi^{(u)}]$ and $\mathbb{P}_{\mathcal{P}^{(u)}}[G^v > \tau]$ for some $\tau \in (0,1)$ (see Appendix F.4).*

In order to incentivize agents of any type u to comply with general extensions Algorithms 3 and 4, we (again) set exploration probability ρ and length ℓ to satisfy some compliance conditions relative to $\mathbb{P}_{\mathcal{P}^{(u)}}[\xi^{(u)}]$ and $\tau\, \mathbb{P}_{\mathcal{P}^{(u)}}[G^v > \tau]$, respectively (see Appendix F.4). We present the (expected) regret from the k treatment extension of policy π_c next.

Lemma 6.3 (Regret of Policy π_c for k Treatments)**.** *An extension of policy π_c achieves (expected) regret as follows:*

$$\mathbb{E}[R(T)] = O\left(k\sqrt{kT\log(kT)}\right) \tag{12}$$

for sufficiently large time horizon T.

Proof Sketch. See Appendix F.4 for the full proof. The proof follows the same structure as that of Lemma 5.5.

Though our analysis covers a more general k treatment setting than Mansour et al. (2015) (capturing non-compliance and selection bias), our policy π_c accumulates asymptotically comparable regret in terms of T. See Appendix F for all other results. Next, in Section 7, we implement Algorithm 1 experimentally.

7. Numerical Experiments

In this section, we present experiments to evaluate Algorithm 1. We mention previously in the paper that this approximation bound decreases slowly throughout Algorithm 1,

[10] Without this assumption, we run into identifiability issues: we cannot reconstruct the individual treatment effects $\theta^1, \dots, \theta^k$ without fixing some mean $\mathbb{E}[g^{(u)}]$. Yet, for purposes of regret minimization, assuming $\mathbb{E}[g^{(u)}] = 0$ does not change our results.

[11] We bastardize notation by writing $x_i = j$ instead of $x_i = \mathbf{e}_j$ (the k-dimensional unit vector along the jth dimension).

[12] The operator $\sigma_{\min}(\cdot)$ denotes the smallest singular value.

[13] Algorithm 3 extends Algorithm 1 and Algorithm 4 extends Algorithm 2.

[14] This ordering assumption is shared by Mansour et al. (2015).

because the exploration probability ρ is small (see Theorem 3.3). So, we are interested in (1) *how small* the exploration probability ρ in Algorithm 1 is and (2) *how slowly* the approximation bound on the absolute difference $|\theta - \hat{\theta}|$ decreases as Algorithm 1 progresses (where $\hat{\theta}$ is based on samples from Algorithm 1). These are important to study, because this slow improvement in accuracy is the primary source of inefficiency (in terms of sample size) for policy π_c, which accumulates linear regret during Algorithm 1 (see Lemma 5.4) for marginal improvements in estimation accuracy. This motivates the social planner to move to Algorithm 2 —where the estimation accuracy increases much quicker— as soon as possible in policy π_c. Yet, there is also a tradeoff for moving to Algorithm 2 too quickly: if Algorithm 1 is not run for long enough, then only a small portion of agents may comply in Algorithm 2. In order to better inform the choice of hyperparameters in policy π_c (specifically, the compliance paramters p_{c_1} and p_{c_2}), we empirically estimate these quantities experimentally. We defer experiments on Algorithm 2 to the appendix.[15]

Experimental Description. We consider a setting with two types of agents: type 0 who are initially never-takers and type 1 who are initially always-takers. We let each agent's prior on the treatment effect be a truncated Gaussian distribution between -1 and 1. The noisy baseline reward $g_t^{(u_t)}$ for each type u of agents is drawn from a Gaussian distribution $\mathcal{N}(\mu_{g^{(u)}}, 1)$, with its mean $\mu_{g^{(u)}}$ also drawn from a Gaussian prior. We let each type of agent have equal proportion in the population, i.e. $p_0 = p_1 = 0.5$. We are interested in finding the probability of event ξ (as defined in Equation (4)) and the exploration probability ρ (as defined in Equation (5)). Instead of deriving an explicit formula for $\mathbb{P}_{\mathcal{P}_0}[\xi]$ to calculate the exploration probability ρ, we estimate it using Monte Carlo simulation by running the first stage of Algorithm 1 for 1000 iterations and aggregating the results. After this, Algorithm 1 is run with the previously-found exploration probability ρ over an increasing number of rounds. We repeatedly calculate the IV estimate of the treatment effect and compare it to a naive OLS estimate (that regresses the treatment onto the reward) over the same samples as a benchmark.

Results. In Figure 1, we compare the approximation bound on $|\theta - \hat{\theta}|$ between IV estimate $\hat{\theta}$ versus via a naive estimate for a specific, chosen $\rho = 0.001$. In our experiments, the exploration probability ρ generally lies within $[0.001, 0.008]$. In Figure 1, we let hidden treatment effect $\theta = 0.5$, type 0 and type 1 agents' priors on the treatment effect be $\mathcal{N}(-0.5, 1)$ and $\mathcal{N}(0.9, 1)$ —each truncated onto $[-1, 1]$,— respectively. We also let the mean baseline reward for type 0 and type 1 agents be $\mu_{g^{(0)}} \sim \mathcal{N}(0, 1)$ and

[15]The code is available here.

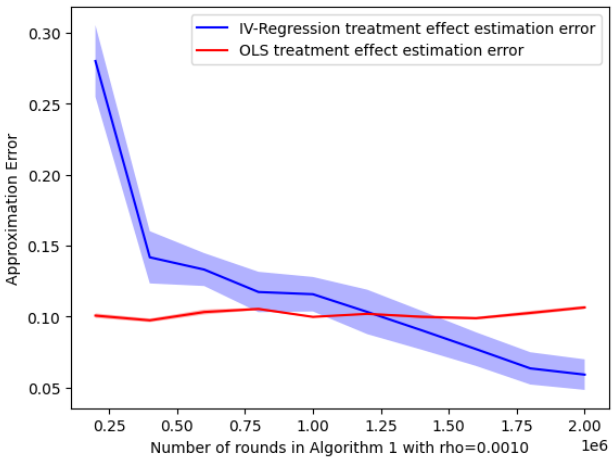

Figure 1. Approximation bound using IV regression and OLS during Algorithm 1 with $\rho = 0.001$. Results are averaged over 5 runs; light blue error bars represent one standard error.

$\mu_{g^{(1)}} \sim \mathcal{N}(0.1, 1)$, respectively. These priors allow us to set the exploration probability $\rho = 0.001$ for Figure 1, where IV regression consistently outperforms OLS for any reasonably long run of Algorithm 1.

To our knowledge, these experiments are the first empirical evaluation of an algorithm in the incentivizing exploration literature. Figure 1 shows the effect of small exploration probability $\rho = 0.001$: we need to run Algorithm 1 for a while to get a decently-accurate causal effect estimate.[16]

8. Conclusion

In this paper, we presented a model for how (non)-compliance changes over time based on beliefs and new information. We provide novel mechanisms that view treatment recommendations as instrumental variables (IVs), which enable treatment effects estimation via IV regression, even in the presence of non-compliance and confounding.

Future Work. Here, we focused on a setting where the causal model is linear and there is no treatment modification by the private type (so all agents share the same treatment effect θ). Future work may extend our results to non-linear settings and settings with treatment effect heterogeneity. We may also relax the (somewhat unrealistic) assumptions 1) that the social planner knows key prior-dependent constants about all agents and 2) that agents fully know their prior, the recommendation mechanism, and can exactly update their posterior over treatment effects. Finally, our empirical results invite further work to improve the practicality of incentivizing exploration mechanisms to allow for more frequent exploration and lessen the number of samples needed.

[16]We suspect this weakness is likely endemic to previous works.

References

Joshua Angrist. Instrumental variables methods in experimental criminological research: What, why, and how? Working Paper 314, National Bureau of Economic Research, September 2005. URL http://www.nber.org/papers/t0314.

Joshua Angrist and Alan Krueger. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15(4):69–85, 2001.

Joshua D. Angrist and Guido W. Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442, 1995. doi: 10.1080/01621459.1995.10476535. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476535.

Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, December 2008. ISBN 0691120358.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002. doi: 10.1023/A:1013689704352.

Howard S. Bloom, Larry L. Orr, Stephen H. Bell, George Cave, Fred Doolittle, Winston Lin, and Johannes M. Bos. The benefits and costs of jtpa title ii-a programs: Key findings from the national job training partnership act study. *The Journal of Human Resources*, 32(3):549–576, 1997. ISSN 0022166X. URL http://www.jstor.org/stable/146183.

Bangrui Chen, Peter Frazier, and David Kempe. Incentivizing exploration by heterogeneous users. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 798–818. PMLR, 06–09 Jul 2018. URL http://proceedings.mlr.press/v75/chen18a.html.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research (JMLR)*, 7:1079–1105, 2006.

Peter Frazier, David Kempe, Jon Kleinberg, and Robert Kleinberg. Incentivizing exploration. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, EC '14, page 5–22, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450325653. doi: 10.1145/2600057.2602897. URL https://doi.org/10.1145/2600057.2602897.

Susan W Groth. Honorarium or coercion: use of incentives for participants in clinical research. *The Journal of the New York State Nurses' Association*, 41(1):11, 2010.

Liz Hamel, Ashley Kirzinger, Lunna Lopes, Grace Sparks, Audrey Kearney, Mellisha Stokes, and Mollyann Brodie. Kff covid-19 vaccine monitor: May 2021, 2021. URL https://www.kff.org/coronavirus-covid-19/poll-finding/kff-covid-19-vaccine-monitor-may-2021/.

Guido Imbens, Joshua Angrist, and Donald Rubin. Identification of causal effects using instrumental variables. *Journal of Econometrics*, 71(1-2):145–160, 1996.

Nicole Immorlica, Jieming Mao, Aleksandrs Slivkins, and Zhiwei Steven Wu. Bayesian exploration with heterogeneous agents. In *The World Wide Web Conference*, WWW '19, page 751–761, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313649. URL https://doi.org/10.1145/3308558.3313649.

Nathan Kallus. Instrument-armed bandits. *ArXiv*, abs/1705.07377, 2018.

Emir Kamenica and Matthew 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.

Sampath Kannan, Michael Kearns, Jamie Morgenstern, Mallesh Pai, Aaron Roth, Rakesh Vohra, and Zhiwei Steven Wu. Fairness incentives for myopic agents. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, EC '17, page 369–386, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450345279. doi: 10.1145/3033274.3085154. URL https://doi.org/10.1145/3033274.3085154.

Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the "wisdom of the crowd". In Michael J. Kearns, R. Preston McAfee, and Éva Tardos, editors, *Proceedings of the fourteenth ACM Conference on Electronic Commerce, EC 2013, Philadelphia, PA, USA, June 16-20, 2013*, pages 605–606. ACM, 2013. doi: 10.1145/2492002.2482542. URL https://doi.org/10.1145/2492002.2482542.

Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. *In 15th ACM Conf. on Economics and Computation (ACM EC)*, 2015.

Yishay Mansour, Aleksandrs Slivkins, Vasilis Syrgkanis, and Zhiwei Steven Wu. Bayesian exploration: Incentivizing exploration in bayesian games. In Vincent Conitzer, Dirk Bergemann, and Yiling 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.

Mark Sellke and Aleksandrs Slivkins. Sample complexity of incentivized exploration. *CoRR*, abs/2002.00558, 2020. URL https://arxiv.org/abs/2002.00558.

Aleksandrs Slivkins. Incentivizing exploration via information asymmetry. *XRDS*, 24(1):38–41, September 2017. ISSN 1528-4972. doi: 10.1145/3123744. URL https://doi.org/10.1145/3123744.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Found. Trends Mach. Learn.*, 12(1-2):1–286, 2019. doi: 10.1561/2200000068. URL https://doi.org/10.1561/2200000068.

EC Wright. Non-compliance–or how many aunts has matilda? *Lancet (London, England)*, 342(8876): 909—913, October 1993. ISSN 0140-6736. doi: 10.1016/0140-6736(93)91951-h. URL https://doi.org/10.1016/0140-6736(93)91951-h.