The welfare impact of recommendation algorithms

LAURA DOVAL
Columbia Business School and CEPR
and
ALEX SMOLIN
Toulouse School of Economics and CEPR

In this letter, we summarize our recent work on the welfare impact of recommendation algorithms and propose questions for further study. We model recommendation algorithms as an information structure, which shapes how a third party takes actions that affect the welfare of different individuals in a population. Each recommendation algorithm thus induces a welfare profile, describing the expected payoffs of different individuals when the third party takes actions following the algorithm. Our framework allows us to characterize and compute the set of all such profiles, which we dub the Bayes welfare set. The Bayes welfare set allows us to reduce society's choice of an algorithm to the choice of a Bayes welfare profile. Our framework complements that of the algorithmic fairness literature which remains agnostic about the population's payoffs, focusing instead on statistical properties of algorithms, such as accuracy, parity, or fairness.

Categories and Subject Descriptors: [**Social and Behavioral Sciences**]: Economics—

General Terms: Economics, Theory
Additional Key Words and Phrases: Recommendation algorithms, fairness, persuasion, information structures

Authors' addresses: laura.doval@columbia.edu, alexey.v.smolin@gmail.com

1. INTRODUCTION

Information has increasingly become a tool for shaping society's choices in high-stakes domains. While this phenomenon is not exclusive to the "big data" economy—consider the role of scores and ratings in school placement, promotions, and credit allocation—the rise of algorithmic recommendations has highlighted society's growing reliance on information in policy-relevant domains. Consider, for instance, the role of algorithms in recommending who obtains bail [Angwin et al. 2016], credit [Jagtiani and Lemieux 2019], who is hired [Raghavan et al. 2020; Li et al. 2020], and which health treatments to prescribe [Obermeyer et al. 2019], and more recently, generative AI models, which users can leverage as *virtual consultants* [Immorlica et al. 2024].

The ever-increasing role of recommendation algorithms in high-stakes domains and their obvious welfare impact has caught the attention of the Computer Science and Economics communities. For instance, [Kearns and Roth 2019; 2020] underscore the importance of having portable definitions of privacy or fairness, which can be coupled with the training model's objective, to produce algorithms with desirable outcomes. Whereas the literature on differential privacy and algorithmic fairness is agnostic about how to measure individuals well-being or the objective of the algorithm designer, [Mullainathan 2018; Kleinberg et al. 2018; Rambachan et al. 2020] argue for letting the social planner's objective determine the properties of algorithms.

In [Doval and Smolin 2021; 2024], we provide a framework to study the welfare impact of recommendation algorithms on a population of heterogeneous individuals. Our framework marries welfare economics and information design. It integrates welfare economics because a primitive of our environment is a measure of individual welfare, which could represent the actual utility function of individuals in the society, or the social planner's perception of this utility. It also draws from information design because recommendation algorithms fundamentally operate as information structures, which provide noisy signals about an underlying state of the world to a decision maker who ultimately takes actions on behalf of the individuals. As such, recommendation algorithms are inherently bounded in their ability to generate welfare, the same way an information designer is bounded in their ability to persuade a receiver to take a given action [Dughmi 2017].

Our primary goal in this note is to introduce the readers to our framework, based on illustrations of the results in [Doval and Smolin 2021] and [Doval and Smolin 2024]. Section 2 introduces the simplest version of our framework to lay down the concepts in the simplest terms. In Section 3, we extend the framework so that it is closest to that in algorithmic fairness. We conclude by pointing out applications of our framework to information design and directions for future research at the intersection of Computer Science and Economics.

2. BASIC FRAMEWORK

In the basic model, a unit mass population of individuals have types in a finite set $\Theta = \{\theta_1, \ldots, \theta_N\}$, drawn from a full support prior distribution, $\mu_0 \in \Delta(\Theta)$. Each individual's welfare depends on her type θ and an (unmodeled) outside observer's

belief about her type. We represent this by a welfare function $w : \Delta(\Theta) \times \Theta \to \mathbb{R}$, representing for each belief μ and type θ, the welfare of individuals of type θ when the outside observer's belief is μ. For instance, if individuals' welfare depends on the actions of the outside observer, the welfare function captures in reduced-form how the outside observer's action, and hence welfare, changes as the outside observer's beliefs about Θ changes. Alternatively, the welfare function may capture that the population's welfare may be driven by image or reputation concerns, like in [Bénabou and Tirole 2006], or psychological motives, as in [Lipnowski and Mathevet 2018].

We model algorithms as *information structures*. An information structure $\Pi = (\pi, S)$ consists of a countable set of labels S, and a mapping π, which associates to each type θ a distribution over signals $\pi(\cdot|\theta) \in \Delta(S)$. Let μ_s denote the posterior belief given signal $s \in S$. An information structure induces two kinds of distribution over posterior beliefs $\{\mu_s : s \in S\}$. First, for each θ, the signal distribution $\pi(\cdot|\theta)$ induces a distribution over posterior beliefs conditional on an individual's type being θ. Second, the prior μ_0 and the signal distribution induce an *unconditional* distribution over posterior beliefs. We denote them by $\langle \Pi|\theta \rangle$ and $\langle \Pi \rangle$, respectively.

The welfare function w together with an information structure, Π, defines a welfare profile, $w_\Pi : \Theta \mapsto \mathbb{R}$, as

$$w_\Pi(\theta) = \mathbb{E}_{\langle \Pi|\theta \rangle} \left[w(\mu, \theta) \right] = \sum_{s \in S} \pi(s|\theta) w(\mu_s, \theta). \tag{1}$$

We denote such a profile, a Bayes welfare profile, and the set of all Bayes welfare profiles, the Bayes welfare set. Formally, the Bayes welfare set is defined as:

$$\mathrm{W} \equiv \left\{ \mathrm{w} \in \mathbb{R}^N : \exists\, \Pi \text{ s.t. } \mathrm{w}_i = w_\Pi(\theta_i)\ \forall i \in \{1, \ldots, N\} \right\}. \tag{2}$$

From the point of view of welfare economics, the Bayes welfare set admits a classical interpretation: It is the utility possibility set in an economy in which information structures take the role of allocations.

An apparent difficulty when characterizing the Bayes welfare set is that the Bayes welfare profiles depend on the *conditional* distributions over posterior beliefs induced by the information structure (cf. Equation (1)). However, we show any Bayes welfare profile satisfies the following:

$$w_\Pi(\theta) = \mathbb{E}_{\langle \Pi|\theta \rangle} \left[w(\mu, \theta) \right] = \mathbb{E}_{\langle \Pi \rangle} \left[\frac{\mu(\theta)}{\mu_0(\theta)} w(\mu, \theta) \right] = \mathbb{E}_{\langle \Pi \rangle} \left[\hat{w}(\mu, \theta) \right]. \tag{3}$$

That is, the expectation of w under Π conditional on θ can be expressed as the unconditional expectation of the *truth-adjusted* welfare function, $\hat{w}$, under Π. The truth-adjusted welfare function, $\hat{w}$, is the welfare function w adjusted by the *truth-drift* $\mu(\theta)/\mu_0(\theta)$. For any given posterior belief μ, the likelihood ratio $\mu(\theta)/\mu_0(\theta)$ measures the representation of type θ under μ relative to its ex ante representation under μ_0.

It follows that the Bayes welfare set can be characterized by studying the convex hull of the graph of the *vector-valued* function, $\hat{\mathrm{w}} : \Delta(\Theta) \mapsto \mathbb{R}^N$, where for each $i \in \{1, \ldots, N\}$, $\hat{\mathrm{w}}_i(\mu) \equiv \hat{w}(\mu, \theta_i)$. Indeed, we have the following:

Theorem 2.1 [Doval and Smolin 2024, Theorem 1]. *The Bayes welfare set* W *satisfies the following:*

$$\mathrm{W} = \left\{ \mathrm{w} \in \mathbb{R}^N : (\mu_0, \mathrm{w}) \in \mathrm{co}\left(\text{graph } \hat{\mathrm{w}}\right) \right\}, \tag{4}$$

where co *denotes the convex hull operator.*

Theorem 2.1 provides a geometric characterization of the set W: it is the section at the prior of the convex hull of the graph of the truth-adjusted welfare function $\hat{\mathrm{w}}$. We illustrate Theorem 2.1 through an example:

Example 2.2 *Online marketplace.* An online marketplace wants to design a recommendation algorithm, directing consumers to buy from sellers in the platform. For simplicity, assume sellers may be of one of two equally likely types: low quality θ_1, and high quality θ_2. Consumers prefer to buy from high quality sellers. Thus, each seller's profit in the marketplace depends on the likelihood μ the consumer attaches to the seller being of high quality. In particular, we assume the sellers' profits as a function of consumers' beliefs are as follows:

$$w(\mu, \theta) = \begin{cases} 0 & \text{if } \mu \in [0, 1/3) \\ 1/2 & \text{if } \mu \in [1/3, 2/3) \\ 1 & \text{if } \mu \in [2/3, 1] \end{cases}. \tag{5}$$

In this example, the set W then represents the set of profit profiles sellers with different qualities can attain in the marketplace under some information structure.

Figure 1 illustrates the convex hull of the graph of $\hat{\mathrm{w}}$ (Figure 1a) and the Bayes welfare set (Figure 1b) for the online marketplace example. For instance, fully revealing or concealing the sellers' quality is always feasible, so that the full and no-disclosure profiles, w^{FD} and w^{ND} are feasible. We highlight some properties of the Bayes welfare set:

—Despite the welfare function being symmetric across seller types, the Bayes welfare set is not symmetric because the adjusted-welfare function is not symmetric. By Bayes rule, when consumers are optimistic about the seller's quality being high, it is more likely they are facing a high rather than a low quality seller.

—In particular, the Bayes welfare set lies above the 45° line: the only Bayes welfare profile equalizing seller profits is the no disclosure one, but it is not Pareto efficient. In other words, fairness—measured by welfare parity—may be at odds with Pareto efficiency.

—The Pareto frontier of the Bayes welfare set is given by its north-east boundary. In particular, the flat segment at the top shows the profits of low-quality sellers can be increased without decreasing those of high-quality sellers.

—The points on the decreasing part of the Pareto frontier can only be generated with at least three signals. By contrast, in standard Bayesian persuasion, two signals are enough in the case of two states. Formally, the analogue of the W in Bayesian persuasion has dimension N, whereas the W has dimension $2N - 1$.

Because the Bayes welfare set is convex, it can be alternatively described by its supporting hyperplanes. [Doval and Smolin 2024, Theorem 2] shows the frontier of

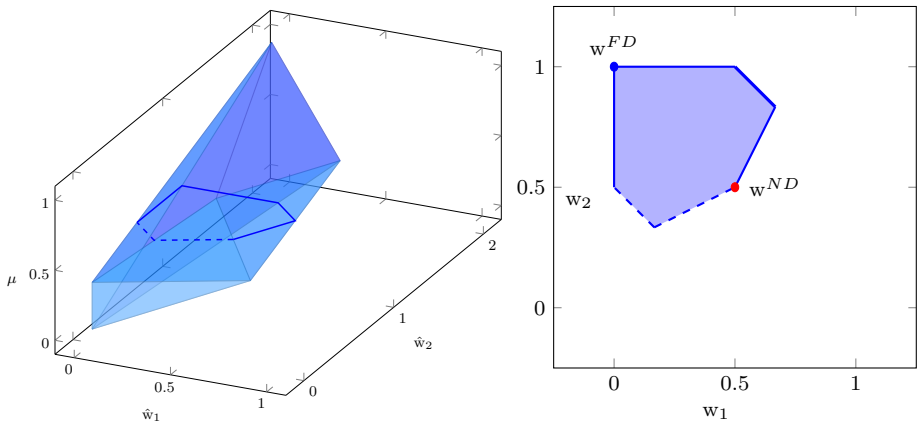

Figure (a) The convex hull of the graph of $\hat{w}$. Figure (b) The Bayes welfare set W.

Fig. 1: Constructing the Bayes welfare set in Example 2.2; w^{FD} and w^{ND} denote the profit profiles under full and no information, respectively.

the Bayes welfare set can be obtained as the solution to series of Bayesian persuasion problems as in [Kamenica and Gentzkow 2011], in which a utilitarian planner takes the role of the information designer. Concretely, consider the supporting hyperplane of the W in direction $\lambda \in \mathbb{R}^N \setminus \{0\}$. Then, the Bayes welfare profiles on the boundary of the W in direction λ can be obtained by solving the Bayesian persuasion problem of a sender with indirect utility

$$\hat{v}_\lambda(\mu) = \sum_{\theta \in \Theta} \mu(\theta) \frac{\lambda(\theta)}{\mu_0(\theta)} w(\mu, \theta).$$

We have found this result very useful in computing the Bayes welfare set in applications (see also [Corrao and Dai 2023] for an application to strategic communication).

3. BEYOND THE BASIC MODEL: GROUPS AND DATA

Two assumptions are implicit in the analysis so far. First, we assume the variable the unmodeled outside observer cares about is the same variable on which we condition the payoffs. Consider, however, an employer making hiring decisions based on a candidate's ability. If candidates belong to different groups, basing hiring recommendations on ability impacts the welfare of candidates across different groups. Second, we assume the information structure can arbitrarily condition on an individual's payoff-relevant type. However, because regulation may prevent the disclosure of protected characteristics, such as gender or race, considering algorithms that respect these restrictions is natural whenever θ includes such characteristics.

Formally, we extend the basic model as follows. We now distinguish between three random variables: an individual's group $g \in G$, the state $\omega \in \Omega$, and data $d \in D$. The first is the variable we condition payoffs on; the second is the variable of interest to the outside observer; the third allows us to capture limits on the information provided. We let $\mathbb{P} \in \Delta(G \times \Omega \times D)$ denote the joint distribution over group-state-

data pairs, and in a slight abuse of notation we denote by $\mathbb{P}(\cdot|g)$ and $\mathbb{P}(\cdot|d)$ the prior distribution conditional on the individual's group and the data realization, respectively. Below, we denote the marginal of $\mathbb{P}$ on D by $\eta_0 \in \Delta(D)$. In a slight abuse of notation, we define the welfare function as $w : \Delta(\Omega) \times \Omega \times G \mapsto \mathbb{R}$, with its first argument being the (unmodeled) outside observer's belief μ about the state ω, $\mu \in \Delta(\Omega)$. The basic model corresponds to the case in which $\Theta = G = \Omega = D$ and $\mathbb{P}(\omega, d|g) = \mathbb{1}[g = \omega = d]$.

To capture the limits data imposes on information provision, an information structure is now defined as a tuple (π, S), where $\pi : D \mapsto \Delta(S)$. Given the information policy, belief updating about (g, ω, d), and hence about ω, depends only on the updated belief about d. Specifically, letting η_s denote the updated belief starting from η_0, after observing signal $s \in S$, the updated belief on (g, ω, d) is given by $\mathbb{P}(g, \omega|d)\eta_s(d)$. Without loss of generality, we can write the welfare function as $w_\dagger(\eta, \omega, g) \equiv w(\mu(\eta), \omega, g)$.

Given an information structure (π, S), the welfare of individuals of group g is:

$$w_\Pi(g) = \sum_{\eta \in \text{supp}(\Pi)} \sum_{s \in S : \eta_s = \eta} \sum_{(\omega,d)} \mathbb{P}(\omega, d|g)\pi(s|d)w_\dagger(\eta, \omega, g), \tag{6}$$

and the Bayes welfare set continues to be defined as the set of Bayes welfare profiles.

The characterization of the Bayes welfare set in the basic model extends verbatim to the more general model, once we observe (the analogue of) the truth-adjusted welfare function now takes the form:

$$\hat{w}_\dagger(\eta, g) = \sum_{(\omega,d)} \mathbb{P}(\omega, d|g) \frac{\eta(d)}{\eta_0(d)} w_\dagger(\eta, \omega, g). \tag{7}$$

Equation (7) allows us to provide further insight into the adjusted welfare function in the basic model. The likelihood correction is now based on the variable d, highlighting that it corresponds to the variable on which information is being provided. In addition, the presence of additional uncertainty requires averaging over ω and d using weights $\mathbb{P}(\omega, d|g)$. Thus we can immediately extend Theorem 2.1 as:

Theorem 3.1 [Doval and Smolin 2021, Theorem 4]. *The Bayes welfare set can be calculated as:*

$$\mathrm{W} = \left\{ \mathrm{w} \in \mathbb{R}^{|G|} : (\eta_0, \mathrm{w}) \in \mathrm{co}\left(\mathrm{graph}\, \hat{\mathrm{w}}_\dagger\right) \right\}. \tag{8}$$

We note again that it is the prior on data, η_0, and not on the states which determines the constraint on how much information can be provided about the state of the world, and hence the limits on how much welfare can be generated via information.

Example 3.2 *Data Regulation in Hiring.* Consider two equally likely groups of workers, labeled A and B. Workers can have one of two ability levels $\Omega = \{0, 1\}$. In each group, half of the workers are high ability and half are low ability. Suppose these workers face a competitive job market: if the market's perceived likelihood that their ability is 1 equals $\mu \equiv \mu(1)$, they receive wage equal to μ. Equating workers' welfare to their wages, this means that $w(\mu, \omega, g) = \mathbb{E}_\mu[\omega]$.

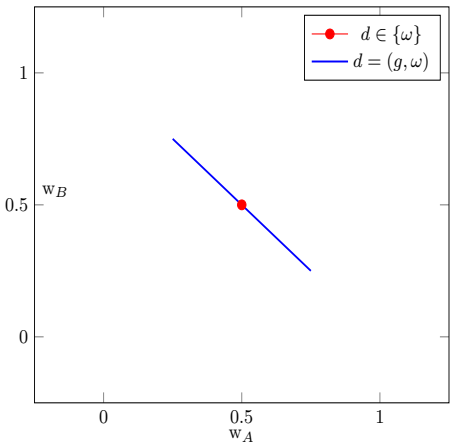

Fig. 2: Bayes welfare set under different data regimes in Example 3.2: The red circle corresponds to the Bayes welfare set under data regimes (i)–(iii); the blue line is the Bayes welfare set in regime (iv).

Rather than assuming a fixed data structure, we compare the Bayes welfare sets in this setting across two data regimes which can be interpreted as different data regulation policies that limit how much information can be revealed about a worker's ability: data reveals ability, but not group (i.e., $D = \Omega$), and data reveals both group and ability (i.e., $D = G \times \Omega$). Figure 2 illustrates the Bayes welfare set in each of these regimes.

Whereas in the first regime we can provide meaningful information about ability, that the distribution of ability is independent across groups together with the martingale property of beliefs implies that on average the posterior belief about the ability remains the same as under no information. It follows that in this case the Bayes welfare set consists of the no disclosure profile, $\mathrm{W} = \{(1/2, 1/2)\}$.

Consider now the second regime and an information structure that pools low-ability workers from group A with high-ability workers from group B and fully reveals all other workers. We can represent this as an information structure with signals $\{B0\}, \{A0, B1\}$, and $\{A1\}$, and induced posterior expectations of 0, $1/2$, and 1, respectively. Because different groups induce these signals with different probabilities, each group's welfare is given by:

$$\mathrm{w}_A = \frac{1}{2}\mathbb{E}[\mu \mid \{A0, B1\}] + \frac{1}{2}\mathbb{E}[\mu \mid \{A1\}] = \frac{1}{2}\frac{1}{2} + \frac{1}{2}1 = \frac{3}{4}, \tag{9}$$

$$\mathrm{w}_B = \frac{1}{2}\mathbb{E}[\mu \mid \{B0\}] + \frac{1}{2}\mathbb{E}[\mu \mid \{A0, B1\}] = \frac{1}{2}0 + \frac{1}{2}\frac{1}{2} = \frac{1}{4}. \tag{10}$$

In fact, this information structure achieves the maximal possible payoff for group A: It never pools workers from group A with the low ability workers of group B, it never pools the high ability workers from group A with workers from group B, and it pools all high ability workers from group B with the workers from group A. As such, the maximal welfare w_A is $3/4$.

We note, however, that the average payoff across groups is the same across all information structures:

$$\frac{1}{2}\mathrm{w}_A + \frac{1}{2}\mathrm{w}_B = \frac{1}{2}\mathbb{E}[\mu \mid g = A] + \frac{1}{2}\mathbb{E}[\mu \mid g = B] = \mathbb{E}[\mu] = \frac{1}{2}. \tag{11}$$

In other words, information merely *redistributes* welfare across the groups. Consequently, the information structure that maximizes the welfare of group A minimizes that of group B.

These observations together with the symmetry of the setting imply that in the fourth regime the Bayes welfare set is given by:

$$\mathrm{W} = \left\{(\mathrm{w}_A, \mathrm{w}_B) \in [1/4, 3/4]^2 : \mathrm{w}_A + \mathrm{w}_B = 1\right\}. \tag{12}$$

4. FINAL REMARKS

We conclude by describing alternative applications of our framework as well as some directions for further research.

4.1 Applications to information design

By interpreting our welfare function as an individual's type-dependent payoff function, the Bayes welfare set is also the object of interest in more standard information design applications. For instance, the types may represent the private information of an informed principal who can commit to an information structure only *after* observing her type, as in [Perez-Richet 2014] and [Koessler and Skreta 2023]. Similar constraints appear in the studies of information design without commitment, as in [Lipnowski and Ravid 2020], [Drakopoulos et al. 2022], and [Corrao and Dai 2023]. Thus, the Bayes welfare set can be viewed as a unifying concept that underlies the incentive constraints the equilibrium information structure must satisfy. As we show in our first working paper version, [Doval and Smolin 2021], our tools also open the door to the study of new problems in this literature such as communication equilibrium payoffs in Bayesian persuasion with transparent motives and Bayesian persuasion with an ambiguity averse sender.[1]

4.2 Further research

We conclude with three (non-exhaustive) suggestions for future research:

It is well-known that various statistical notions of fairness, such as equalized odds and calibration, are incompatible with each other (cf. [Chouldechova 2017; Kleinberg et al. 2016]). Furthermore, this incompatibility remains even when considering relaxations [Pleiss et al. 2017]. Yet, the Bayes welfare set provides another way to visualize the trade-offs among these competing notions. For instance, one could use the Bayes welfare set to understand which group is hurt the most when imposing either calibration or equalized odds. Similarly, one could consider information structures that preserve some form of privacy—e.g., the algorithm recommendations do not reveal information about group membership—and study the Bayes

[1] [Corrao and Dai 2023] fully characterize the set of communication equilibria with transparent motives.

welfare profiles consistent with such restrictions (cf. [Gopalan et al. 2021; Strack and Yang 2024]).

Since the seminal work of [Dughmi and Xu 2016], the computer science literature has made incredible progress in algorithmic Bayesian persuasion (see, e.g., [Babichenko and Barman 2016; Arieli and Babichenko 2019; Banerjee et al. 2024]). Most of this work is concerned with the computational aspects of achieving the sender's preferred payoff, whereas our work focuses on the cross-sectional implications of different information structures for which the sender's average payoff may not be a sufficient statistic.

In many applications, considering constraints on the information structures the planner has access to is natural. The model in Section 3 puts limits on how much information can be provided about the payoff-relevant state. Constraints such as those arising from differential privacy are relevant in many applications and understanding how they shape the choice out of the Bayes welfare set is of interest.

REFERENCES

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. 2016. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica 23*, 77–91.

Arieli, I. and Babichenko, Y. 2019. Private bayesian persuasion. *Journal of Economic Theory 182*, 185–217.

Babichenko, Y. and Barman, S. 2016. Computational aspects of private bayesian persuasion. *arXiv preprint arXiv:1603.01444*.

Banerjee, S., Munagala, K., Shen, Y., and Wang, K. 2024. Fair price discrimination. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 2679–2703.

Bénabou, R. and Tirole, J. 2006. Incentives and prosocial behavior. *American Economic Review 96,* 5, 1652–1678.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data 5,* 2, 153–163.

Corrao, R. and Dai, Y. 2023. The bounds of mediated communication. *arXiv preprint arXiv:2303.06244*.

Doval, L. and Smolin, A. 2021. Information payoffs: An interim perspective. *arXiv preprint arXiv:2109.03061*.

Doval, L. and Smolin, A. 2024. Persuasion and welfare. *Journal of Political Economy 132,* 7, 2451–2487.

Drakopoulos, K., Lo, I., and Mulvany, J. 2022. Blockchain mediated persuasion. *USC Marshall School of Business Research Paper Sponsored by iORB*.

Dughmi, S. 2017. Algorithmic information structure design: a survey. *ACM SIGecom Exchanges 15,* 2, 2–24.

Dughmi, S. and Xu, H. 2016. Algorithmic bayesian persuasion. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. 412–425.

Gopalan, P., Kalai, A. T., Reingold, O., Sharan, V., and Wieder, U. 2021. Omnipredictors. *arXiv preprint arXiv:2109.05389*.

Immorlica, N., Lucier, B., and Slivkins, A. 2024. Generative ai as economic agents. *ACM SIGecom Exchanges 22,* 1, 93–109.

Jagtiani, J. and Lemieux, C. 2019. The roles of alternative data and machine learning in fintech lending: Evidence from the lendingclub consumer platform. *Financial Management 48,* 4, 1009–1029.

Kamenica, E. and Gentzkow, M. 2011. Bayesian persuasion. *American Economic Review 101*, 2590–2615.

Kearns, M. and Roth, A. 2019. *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Kearns, M. and Roth, A. 2020. Ethical algorithm design. *ACM SIGecom Exchanges 18,* 1, 31–36.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Rambachan, A. 2018. Algorithmic fairness. In *AEA Papers and Proceedings*. Vol. 108. 22–27.

Kleinberg, J., Mullainathan, S., and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Koessler, F. and Skreta, V. 2023. Informed information design. *Journal of Political Economy 131,* 11, 3186–3232.

Li, D., Raymond, L. R., and Bergman, P. 2020. Hiring as exploration. *National Bureau of Economic Research*.

Lipnowski, E. and Mathevet, L. 2018. Disclosure to a psychological audience. *American Economic Journal: Microeconomics 10,* 4, 67–93.

Lipnowski, E. and Ravid, D. 2020. Cheap talk with transparent motives. *Econometrica 88,* 4, 1631–1660.

Mullainathan, S. 2018. Algorithmic fairness and the social welfare function. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. 1–1.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science 366,* 6464, 447–453.

Perez-Richet, E. 2014. Interim bayesian persuasion: First steps. *American Economic Review 104,* 5, 469–74.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. 2017. On fairness and calibration. *Advances in neural information processing systems 30*.

Raghavan, M., Barocas, S., Kleinberg, J., and Levy, K. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 469–481.

Rambachan, A., Kleinberg, J., Ludwig, J., and Mullainathan, S. 2020. An economic perspective on algorithmic fairness. In *AEA Papers and Proceedings*. Vol. 110. 91–95.

Strack, P. and Yang, K. H. 2024. Privacy preserving signals. *Available at SSRN 4467608*.