Privacy Preserving Signals*

Philipp Strack† Kai Hao Yang‡

September 18, 2023

Abstract

A signal is *privacy-preserving* with respect to a collection of *privacy sets*, if the posterior probability assigned to every privacy set remains unchanged conditional on any signal realization. We characterize the privacy-preserving signals for arbitrary state space and arbitrary privacy sets. A signal is privacy-preserving if and only if it is a garbling of a *reordered quantile signal*. These signals are equivalent to couplings, which in turn lead to a characterization of optimal privacy-preserving signals for a decision-maker. We demonstrate the applications of this characterization in the contexts of algorithmic fairness, price discrimination, and information design.

Keywords: Privacy-preserving signal, privacy sets, independence, reordered quantile signal, protected characteristics.

JEL classification: C11, D63, D42, D83,

*We thank Dirk Bergemann, Joyee Deb, Laura Doval, Mira Frick, Paul Heidhues, Ryota Iijima, Emir Kamenica, Elliot Lipnowski, Alessandro Lizzeri, Barry Nalebuff, Aniko Öry, Benjamin Polak, Aaron Roth, Ludvig Sinander, Nathan Yoder, Alexander Zentefis, and Jidong Zhou for their valuable comments and suggestions. We also thank Jialun (Neil) He for his research assistance. All errors are our own.

†Department of Economics, Yale University, Email: philipp.strack@yale.edu

‡School of Management, Yale University, Email: kaihao.yang@yale.edu

1 Introduction

In many economic settings, there are constraints on what information can be used or revealed: Characteristics such as race, gender, and sexual orientation are protected in many contexts, and the information that can be revealed about them is limited due to legal, regulatory, or social norms. Motivated by this, we study the set of signals (Blackwell experiments) which are constrained to not reveal certain information.

For example, consider the case of a bank determining whether to grant a loan to an individual. In making this decision, the bank benefits from using an individual's characteristics to predict whether they will default. However, the Equal Credit Opportunity Act prohibits discrimination against loan applicants on the basis of their "protected characteristics", such as race, gender, or age. As a result, the bank is legally required to ensure that its loan decisions are not influenced by these protected characteristics. In other words, the information used by the bank in making loan decisions cannot be based on these characteristics.

This paper presents a framework for understanding information in situations where certain aspects of the state of the world must be kept private, or equivalently, where decisions must be taken independent of certain aspects of the state of the world. Following Blackwell (1953), we model information as a signal about an abstract state of the world. To capture a notion of protected information, we introduce a collection of events called *privacy sets*, which represent the aspects of the state of the world that cannot be disclosed. For instance, in the context of a bank loan, the privacy sets would include all the protected characteristics. We define a signal as *privacy-preserving* if, for any signal realization, the posterior probability of any privacy set equals its prior probability. In other words, a privacy-preserving signal does not reveal any information about events that belong to the privacy sets.

We characterize all privacy-preserving signals. We first show that the privacy sets can always be represented as a component of an extended state space. Specifically, for arbitrary state of the world $\omega \in \Omega$, there exists an extended state space $(\omega, \theta) \in \Omega \times \Theta$ such that a signal is privacy-preserving if and only if it is independent of θ. This representation is convenient, as the privacy sets might otherwise overlap in arbitrarily complicated ways and are seemingly intractable. In the bank loan example, realizations of ω would be an applicant's default probability and observable attributes, and realizations of θ would be an applicant's protected characteristics. A privacy-preserving signal could reveal information about an applicant's

default probability ω, but must be independent of applicant's protected characteristics θ.

Our first main result (Theorem 1) shows that, when Ω is one-dimensional, a signal is privacy-preserving if and only if it is a garbling of some *reordered quantile signal*. The *quantile signal* is the signal that reveals, for each (ω, θ), the quantile $q = F(\omega \mid \theta)$ of ω given θ, plus potentially some noises when it has an atom. This signal is privacy-preserving as it is uniformly distributed on $[0, 1]$ for every value of θ. A reordered quantile signal is a signal obtained by further (randomly) reordering the unit interval using some measure preserving transformation Φ_θ. As we prove these transformations ensure that the reordered quantile signal remains privacy-preserving. Theorem 1 establishs that reordered quantile signals are exactly the frontier of all privacy-preserving signals in Blackwell's sense: Every privacy-preserving signal is Blackwell-less informative than some reordered quantile signal, while reordered quantile signals are Blackwell-undominated.

Although privacy-preserving signals do in general not have a Blackwell-maximum, there exists a privacy-preserving signal which induces a distribution of posterior *means* that is the most dispersed (Theorem 2). Consequently, in settings where the only economically relevant variables are the posterior means (e.g., when a decision-maker has a payoff that is affine in ω, or when ω is binary), *every* privacy-preserving signal is dominated by the generalized quantile signal.

While independence may seem to be a stringent requirement in some economic examples, our results extend to a setting where privacy-preserving is defined *conditional* on another given random variable. Specifically, consider a further extended state space $(\omega, \theta, y) \in \Omega \times \Theta \times Y$. A signal said to be *conditionally privacy-preserving* if it is independent of θ *conditional on* y. Thus, our methods can be used to analyze settings where signals are allowed to reveal information about the protected privacy sets, as long as it is through "materially relevant characteristics". For example, banks cannot directly base discriminatory loan decisions on race, but can make predictions using applicants' credit history, even though credit histories may be correlated with race.

Having characterized the set of privacy-preserving signals, we then explore how to optimize over this set. Consider a decision-maker who chooses an action $a \in A$ to maximize their payoff $u(\omega, \theta, a)$, after observing a privacy-preserving signal. Our results imply that it is without loss to restrict attention to reordered quantile signals. Proposition 2 shows that the resulting optimization problem is equivalent to a Kantorovich optimal transport problem

if the protected characteristic is binary (e.g. white/non-white) or an optimization problem over copulas more generally. Moreover, if the decision-maker's payoff is either single-crossing in (ω, a), or linear in ω, or if the state ω is binary then the generalized quantile signal is optimal for the decision-maker, *independent* of other details of the decision problem. Finally, we derive the optimal privacy-preserving signal for binary actions.

Next, we explore some economic applications of our main results. First, we apply our characterizations to fairness and algorithmic design. In this literature, one of the most commonly adopted notions of fairness is called *independence*. It requires decisions to be independent of protected characteristics (conditionally on materially relevant characteristics). Our results lead to a characterization of optimal fair algorithms, which generalizes existing results that apply only to binary-action decision problems with a specific payoff functions. Furthermore, we lay out an optimal and detail free procedure for regulating algorithm design.

As a second application, we consider price discrimination and market segmentation in the spirit of Bergemann, Brooks and Morris (2015). In a setting where a monopolist is able to segment consumers, we consider a situation where consumers with different protected characteristics (e.g., gender, race) or different sensitive information (e.g., genetic information), must face the same distribution of prices, even though the monopolist engages in third-degree price discrimination. Theorem 1 implies a characterization of market segmentations that would allow the monopolist to optimally price-discriminate without discriminating based on protected characteristics. Solving the optimal transport problem derived in Proposition 2 allows us compute the market segmentation that maximizes the monopolist's revenue while preventing price discrimination based on consumers' protected characteristics.

Third, we show that our results lead to generalizations of an elegant recent result by He, Sandomirskiy and Tamuz (2023), who study a setting where $n \geq 2$ agents are privately informed about a binary state by independent signals, which they refer to as a *private private information structure*. They show that for two agents, a private private information structure is undominated (i.e., no other private private information structure can induce a more dispersed distribution of posterior means for each agent) if and only if each agent's distribution of posterior is the conjugate of the other's. Our results generalize the characterization to the case of more than two agents and more than two states.

Lastly, we apply our results to Bayesian persuasion, where a sender chooses a privacy-preserving signal to inform a receiver, who then selects an action after observing the sig-

nal realization. Our results lead to a familiar "concavification" technique to solve for the sender's optimal privacy-preserving signal, as in Kamenica and Gentzkow (2011), except that the privacy constraints require an extension of underlying state space. When the state is one-dimensional, and the sender's indirect utility depends only on the posterior mean, the persuasion problem can be reduced to choosing a distribution subject to a mean-preserving contraction constraint (that is more demanding than the one in the case without privacy constraints derived in Gentzkow and Kamenica 2016).

Related Literature This paper is related to several strands of literature. We follow the canonical approach of Blackwell (1953) and model information as signals about an underlying state. We generalize Blackwell's characterization of feasible signals, by characterizing all feasible signals that are do not reveal information about a given collection of events. While Blackwell shows that a signal is feasible if and only if it is dominated by the signal which fully reveals the state, we show that a signal is feasible and privacy-preserving if and only if it is dominated by a "reordered quantile signal".

To illustrate the usefulness of our mathematical results, we apply them to various topics in economics and computer-science. In the literature on algorithmic fairness, it is well-documented that recommendations made by predictive algorithms could be discriminatory (see, for example, Angwin, Larson, Mattu and Kirchner 2016; Arnold, Dobbie and Hull 2022; Fuster, Goldsmith-Pinkham, Ramadorai and Walther forthcoming). Many studies, both theoretical and empirical, have suggests methods to regulate algorithm design to ensure fairness.

One of the most common criteria for fairness in this literature requires the decisions to be statistically independent of protected characteristics (see, e.g., Calders, Kamiran and Pechenizkiy 2009; Kamiran, Žilobaitė and Calders 2013; Zafar, Valera, Rodriguez and Gummadi 2015; Corbett-Davies, Pierson, Feller, Goel and Huq 2017; Kitagawa, Sakaguchi and Tetenov 2021; Gillis, McLaughlin and Spiess 2021). These papers characterize the optimal algorithms, which are fair in the aforementioned sense, for decision problems with binary actions, binary states and specific payoff structures. Applying our general characterization allows us to unify and generalize findings in this literature to more than two actions, more than two states, and general payoff functions. Our results implies a detail-free approach to regulating the design of predictive algorithms

In economics, recent work by Liang, Lu and Mu (2023) considers a different notion of fairness and characterizes the Pareto frontier in terms of payoffs of two groups and the difference in payoffs for a fixed binary-action decision problem with binary protected groups and discuss how the optimal policy garbles the inputs to the algorithm. Doval and Smolin (2023) characterize which vectors of group specific payoffs form the Pareto frontier in an information design problem and how points on the frontier can be reached by regulating the output of the algorithm.

A different notion of privacy used in the computer-science literature is differential privacy proposed by Dwork, McSherry, Nissim and Smith (2006). This notion considers signals as a function of the characteristics of a population of agents such that each individual agent affects the log-likelihood of each signal by at most ϵ. Schmutte and Yoder (2022) characterizes the distribution of posteriors that can be induced by signals satisfying ε-differential privacy and studies an information design problem subject to differential privacy.

In the literature on price discrimination, several studies explore the welfare implications of different market segmentations (see, for instance, Varian 1985; Aguirre, Cowan and Vickers 2010; Cowan 2016). Among these papers, Bergemann et al. (2015) interpret market segmentations as a signal about consumers' values and characterize the set of possible consumer and producer surplus in the case of unit demand. Haghpanah and Siegel (2022, Forthcoming) further explore the case when the monopolist sells multiple products. Our results provide a characterization of feasible market segmentations if the seller can not price discriminate based on protected characteristics.

Outline The rest of this paper is organized as follows: §2 introduces our framework for privacy-preserving signals. §3 characterizes privacy-preserving signals for the case of one-dimensional ω, followed by further results regarding optimal privacy-preserving signals in §4. §5 discusses economic applications of our results. §6 states the general version of our result and provides its proof, followed by further discussions in §7. §8 concludes. Proofs for the auxiliary lemmas are relegated to the Appendix, while proofs for the applications can be found in the Online Appendix.

2 Model

State and Signals Consider a standard Borel space $(\Omega, \mathcal{F})$ and a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$. The state ω is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$.[1] A signal $(S, \pi) = (S, (\pi_\omega)_{\omega \in \Omega})$ consists of a measurable set S of signal realizations, as well as a conditional distribution over signal realizations $\pi_\omega \in \Delta(S)$ for each state ω.[2] To simplify notation, we will refer to the signal (S, π) simply as π, and leave the set of signal realizations S implicit. We denote by $\mathbb{P}_\pi$ the induced probability measure on $\Omega \times S$ when s is distributed according to π_ω in state ω. Since $(\Omega, \mathcal{F}, \mathbb{P})$ is a standard probability space, the posterior posterior belief $\mathbb{P}_\pi[\cdot \mid s]$ given each realized $s \in S$ is well-defined.[3]

Privacy-Preserving Signals We are interested in signals about the state—or decisions that are made conditional on signal realizations—that do not reveal certain type of information. For example, in some economic context the signals might be required to preserve some notion of privacy; or an action taken by a firm might not be allowed to condition on protected characteristics such as race. For a collection of *privacy-sets*

$$\mathcal{P} \subseteq \mathcal{F}$$

that are closed under finite intersections,[4] each $P \in \mathcal{P}$ defines an event that has be to "kept private" and no information about it can be revealed. We do not impose any structures on $\mathcal{P}$ and allow for the number of privacy sets to be finite, or infinite and potentially uncountable.

Definition 1 (Privacy-Preserving Signals)**.** A signal π is *privacy-preserving* if the prior and

[1] Note that there are no restrictions on the state space besides being standard. For example, the state ω might be multidimensional and captures all relevant characteristics of an economic agent such as income, age, gender, race, address, etc.

[2] A signal is thus the same as a Blackwell experiment defined in Blackwell (1953).

[3] Formally, a signal π is a transition probability from Ω to $\Delta(S)$. Denote by Σ the σ-algebra of S, we consider the probability space consisting of the outcome space $\Omega \times S$ the product σ-algebra of $\mathcal{F}$ and Σ, and the probability measure $\mathbb{P}_\pi$ induced by drawing s from π_ω conditional on ω. The posterior belief $\mathbb{P}_\pi[\cdot \mid s]$ is a *regular version* of conditional expectation $\mathbb{E}[\mathbf{1}\{\omega \in \cdot\} \mid s]$ and thus is a transition probability that is consistent with the joint distribution $\mathbb{P}_\pi$ of ω and s.

[4] Namely, $\mathcal{P}$ is a π-system. This means that if two sets P_1 and P_2 are events that need to be kept private, then their intersection $P_1 \cap P_2$ has to be kept private too.

posterior probability of the state being in any privacy set coincide, i.e., for all $P \in \mathcal{P}, s \in S$,

$$\mathbb{P}[\omega \in P] = \mathbb{P}_\pi[\omega \in P \mid s]\,. \tag{1}$$

According to the definition, under any privacy-preserving signal, observing the signal realization s reveals no information about any privacy set $P \in \mathcal{P}$.

Characteristic-Specific Privacy Sets A particular instance of privacy sets is when the state is multidimensional, and the privacy sets are given by some components of the state. These components could capture protected characteristics of individuals such as race, gender, or age.

Definition 2 (Characteristic-Specific Privacy Sets)**.** The privacy sets $\mathcal{P}$ are *characteristic-specific* if $\omega = (\omega_1, \omega_2) \in \Omega = \Omega_1 \times \Omega_2$ and $\mathcal{P}$ is the σ-algebra generated by ω_2.

If privacy sets are characteristic-specific, then a signal π is privacy-preserving if and only if the signal realizations are independent of ω_2. In this case, we say that a signal is privacy-preserving with respect to ω_2.

Characteristic-specific privacy sets are tractable because the privacy sets are encoded in component ω_2. In fact, as we argue below, by possibly enlarging the state space, one can always represent arbitrary privacy sets as characteristic-specific privacy sets.

Proposition 1. *There exists a random variable $\theta : \Omega \to \Theta$ such that a signal is privacy-preserving with respect to $\mathcal{P}$ if and only if its signal realizations are independent of θ.*

Consequently, a signal is privacy-preserving if and only if it is privacy-preserving with respect to component θ of the extended state (ω, θ). As a result, we henceforth focus on a state space $\Omega \times \Theta$ and assume, without loss of generality, that the privacy sets are characteristic specific.

The formal proof of Proposition 1 can be found in the Appendix. To better understand the intuition, suppose that there are finitely many privacy sets: $|\mathcal{P}| < \infty$. As we prove in Lemma A.1 in the Appendix, a signal is privacy-preserving with respect to $\mathcal{P}$ if and only if it is privacy-preserving with respect to the σ-algebra generated by $\mathcal{P}$. Note that, the minimal elements $\{Q_k\}_{k=1}^n$ of the σ-algebra generated by $\mathcal{P}$ form a partition of Ω. Thus, the σ-algebra

generated by $\mathcal{P}$ is also generated by the random variable $\theta \in \{1, \ldots, n\}$, where, for all $\omega \in \Omega$,

$$\theta(\omega) := \sum_{k=1}^{n} k \, \mathbf{1}\{\omega \in Q_k\} \, .$$

The random variable θ indicates which element of the partition $\{Q_k\}_{k=1}^n$ the state ω falls into.

For example, if the privacy sets divide the population into female/male and white/non-white, then the minimal sets would be non-white females; non-white males; white females; and white males. In this case, θ would simply be a random variable indicating to which group a given person belongs to.

3 Characterization of Privacy-Preserving Signals

For the ease of exposition, we first state the result for the case where $\Omega \subseteq \mathbb{R}$.[5] The general result, which does not impose any assumptions on Ω, can be found in §6. To begin with, we first define a class of (privacy-preserving) signals that are crucial to our characterization.

Quantile Signals Denote by $F(\cdot \mid \theta)$ the cumulative distribution function (CDF) of ω conditional on θ. That is, $F(x \mid \theta) := \mathbb{P}[\omega \leq x \mid \theta]$ for all $x \in \mathbb{R}$ and $\theta \in \Theta$. If $F(\cdot \mid \theta)$ is continuous for all $\theta \in \Theta$, we define the random variable

$$q = F(\omega \mid \theta) \tag{2}$$

to be the empirical quantile of the conditional distribution. Whenever the conditional distributions $F(\cdot \mid \theta)$ are continuous, the empirical quantile $q = F(\omega \mid \theta)$ is uniformly distributed on $[0, 1]$ for all θ, and revealing it thus constitutes a privacy-preserving signal. If the conditional distributions are not all continuous, the empirical quantile is not uniformly distributed and additional randomization is needed to construct a privacy-preserving signal based on the quantile. In general, one can define the *generalized quantile* to be the quantile at those points where the CDF is continuous and uniformly randomizing over the associated quantiles at the

[5] This is equivalent to assuming that every privacy set $P \in \mathcal{P}$ is totally ordered. Note that this includes any instances when Ω is finite, as any finite set is totally ordered.

jump-points, i.e.

$$q \sim \begin{cases} \delta_{F(\omega|\theta)} & \text{if } F(\omega_- \mid \theta) = F(\omega \mid \theta) \\ \text{Unif}([F(\omega_- \mid \theta), F(\omega \mid \theta)]) & \text{else} \end{cases}, \tag{3}$$

where $F(\omega_- \mid \theta) = \lim_{\varepsilon \searrow 0} F(\omega - \varepsilon \mid \theta)$ denotes the left limit of $F(\cdot \mid \theta)$ at ω, δ_x denotes a Dirac measure at x, and $\text{Unif}([x_1, x_2])$ denotes the uniform distribution over the interval $[x_1, x_2]$. We refer to the signal above as the *generalized quantile signal*, and denote it by $\pi^\star$, as it reveals the empirical quantile with some potential additional randomization.

Reordered Quantile Signals We next use the generalized quantile signal to construct a family of privacy-preserving signals. Let $\Phi_\theta : [0,1] \to [0,1]$ be a function that preserve the Lebesgue measure, indexed by $\theta \in \Theta$.[6] For each state (ω, θ), we first draw a generalized quantile signal q according to (3). Then, conditional on the realization of q, randomly draw a signal realization s from the set

$$\Phi_\theta^{-1}(q) := \{s \in [0,1] : \Phi_\theta(s) = q\}.$$

Since Φ_θ is measure-preserving and since q is uniformly distributed for each θ, s can be drawn in a way so that s is uniformly distributed on $[0,1]$ for each θ.[7] This defines a privacy-preserving signal, which we denote by $\pi^\star_\Phi$, and refer to as a *reordered quantile signal*. Intuitively, $\pi^\star_\Phi$ is obtained by (randomly) reordering the generalized quantile signal in such a way that the uniform measure is preserved. Note that given any realizations of s, q, and θ, the signal s and the characteristic θ together reveal the generalized quantile as

$$\Phi_\theta(s) = q.$$

For example, if $\theta \in \{0,1\}$, then revealing $s = q$ when $\theta = 0$ and $s = 1 - q$ when $\theta = 1$ constitutes a reordered quantile signal, with $\Phi_0(s) = s$ and $\Phi_1(s) = 1 - s$. Note that a reordered quantile

[6] A function ϕ preserves the Lebesgue measure if $\int_0^1 \mathbf{1}\{\phi(r) \le x\}\, \mathrm{d}r = x$ for all $x \in [0,1]$.

[7] More precisely, let s be drawn from (any version of) the disintegration of the (Borel) probability measure: $C(A \times B) := \int_A \mathbf{1}\{\Phi_\theta(s) \in B\}\, \mathrm{d}s$ with respect to he Lebesgue measure. Then s is uniformly distributed since Φ_θ is measure-preserving.

signal might induce additional non-trivial randomization (even conditional on the realization of q), as demonstrated by the following example:[8]

Example 1. Consider the measure-preserving transformation $\Phi_\theta(s) = 2s - \mathbf{1}\{s \geq 1/2\}$. Then the signal realization s given by the reordered quantile signal conditional on q and θ is random and equals $q/2$ and $q/2 + 1/2$ with equal probability. As q is uniformly distributed on $[0, 1]$, s is uniformly distributed on $[0, 0.5]$ and $[0.5, 1]$ with equal probability, which implies it is uniformly distributed on $[0, 1]$ for any θ, and hence is privacy-preserving.

Our next result establishes that the reordered quantile signals are the maximals and completely characterize the set of privacy-preserving signals.

Theorem 1 (Characterization of Privacy-Preserving Signals)**.** *Suppose that* $\Omega \subseteq \mathbb{R}$.

(i) A signal is privacy-preserving if and only if it is Blackwell-dominated by some reordered quantile signal $\pi^\star_\Phi$.

(ii) Every reordered quantile signal is Blackwell-undominated among privacy-preserving signals.

Part (i) of Theorem 1 establishes that each privacy-preserving signal can be described by two components: First, a family of measure-preserving transformations that identify a reordered quantile signal, and second, a garbling which describes what information is not revealed. An immediate economic consequence of this result is that *every* privacy-preserving signal can be generated by adding noises to a reordered quantile signal. Thus, it is without loss of generality in decision problems to optimize only over reordered quantile signals instead of all privacy-preserving signals, as the decision-maker can always ignore additional information. Part (ii) of Theorem 1 establishes that every reordered quantile signal is Blackwell undominated. Thus, without imposing further structure on the decision problem, no further restriction of the set of privacy-preserving signals is without loss of generality.

Together, Theorem 1 characterizes the set of privacy-preserving signals, and identifies its Blackwell frontier as the reordered quantile signals. In §6, we develop the general version of this result, which does not impose any restrictions on the state space Ω.

[8] While in this example the randomization over the preimage of q is uniform, this is not necessarily the case for general reordered quantile signals.

Distributions of Posterior Means Our characterization of Theorem 1 can be further sharpened if one focuses on distributions of posterior means. Consider the case where $\Omega \subseteq \mathbb{R}$ and $\mathbb{E}[\omega \mid \theta]$ exists for all $\theta \in \Theta$. It is well-known that a CDF G is a distribution of posterior means $\mathbb{E}_\pi[\omega \mid s]$ under some signal π if and only if G is a mean-preserving contraction of the prior distribution $F(x) := \mathbb{P}[\omega \leq x]$ of ω (see e.g., Strassen 1965). However, not every mean-preserving contraction of F can be the distribution of posterior mean under a privacy-preserving signal. For example, if ω and θ are not independent, then F could never be the distribution of posterior means, since it can only be induced by fully revealing ω, which violates the privacy constraint.

Let $\overline{F}$ be defined as

$$\overline{F}(x) := \inf \left\{ q \in [0,1] : \mathbb{E}\left[F^{-1}(q \mid \theta) \right] \geq x \right\},$$

for all $x \in \mathbb{R}$. By definition, $\overline{F}$ is a CDF on $\mathbb{R}$. Note that $\overline{F}$ is exactly the distribution over posterior means $\mathbb{E}_{\pi^\star}[\omega \mid s]$ induced by the generalized quantile signal $\pi^\star$.

Theorem 2 (Distributions of Posterior Means)**.** *Suppose that $\Omega \subseteq \mathbb{R}$. Then a CDF G is the distribution of posterior means induced by some privacy-preserving signal if and only if G is a mean-preserving contraction of $\overline{F}$.*

Although privacy-preserving signals do not have a Blackwell-maximum in general according to Theorem 1, in settings where only the posterior means of ω are relevant, Theorem 2 implies that the generalized quantile signal is the most informative among all privacy-preserving signals. In other words, Theorem 2 means that the distributions of posterior means induced by reordered quantile signals can be further ranked under the mean-preserving contraction order, with the one induced by the generalized quantile signal being the most dispersed. Consequently, in settings where only posterior means of ω are relevant, the undominated privacy-preserving signals collapse to a singleton.

This observation reduces an optimization problem over the set of privacy-preserving signals to an optimization problem over mean preserving contractions of $\overline{F}$. Recent results in Kleiner, Moldovanu and Strack (2021) and Arieli, Babichenko, Smorodinsky and Yamashita (2023) characterize the extreme points of this set, which allows one to further restrict the set of signals one needs to consider.

In addition, Theorem 2 generalizes an elegant recent result of He et al. (2023), which characterizes the distributions of posterior means when ω can take only two values $\{0, 1\}$. Their proof relies on a very different methodology based on tools from mathematical tomography. Our methodology allows us to greatly simplify the proofs and extend the results beyond binary states.

Remark 1 (Conditionally Privacy-Preserving Signals)**.** In many economic applications, a signal is only required to be privacy-preserving conditional on certain information. For example, if some information y is already publicly available, then it would be natural to only restrict signal to not reveal *additional* information. This can be captured by considering the state space (ω, θ, y) and defining a signal to be *conditionally privacy-preserving*, if its realization is independent of θ, conditional on y. Mathematically, our Theorem 1 and Theorem 2 immediately extend to this case by simply applying them for each fixed value of y, defining the quantile signal as $q = F(\omega \mid \theta, y)$ and the generalized quantile signal analogously.

Remark 2 (Relaxing the Independence Criterion)**.** An immediate implication of Remark 1 is a relaxation of the definition of privacy-preserving signals. While we define privacy-preserving signals by the notion of independence, conditionally privacy-preserving signals relaxes the independence requirement by allowing for correlations with the component y. In particular, one may consider a joint distribution of (ω, θ, y) where y is independent of ω but correlated with θ.[9] Conditionally privacy-preserving signals in this environment can then be regarded as privacy-preserving signals with a less stringent requirement for independence, as correlation between s and θ would be allowed as long as it is through y.

4 Optimizing over Privacy-Preserving Signals

In this section, we apply Theorem 1 and Theorem 2 to characterize optimal signals for a decision-maker who takes an action after observing the signal realization. For the ease of exposition, we focus on the case where $\Omega \subset \mathbb{R}$ and assume that θ takes finitely many values: $\Theta = \{\theta_1, \ldots, \theta_J\}$. Moreover, we assume that the conditional expectation $\mathbb{E}[\omega \mid \theta]$ exists for all $\theta \in \Theta$.

[9] In fact, we can fully characterize these joint distributions, as they are equivalent to privacy-preserving signals for θ that are independent of ω.

Consider a (Bayesian) decision-making problem: The decision-maker chooses an action $a \in A$ to maximize expected payoff. The decision-maker's payoff is given by

$$u : \Omega \times \Theta \times A \to \mathbb{R},$$

where $u(\omega, \theta, a)$ denotes the agent's ex-post payoff when the state is (ω, θ) and the action is a. Before taking actions, the decision-maker observes a signal realization drawn from a privacy-preserving signal π.

Clearly, the optimal signal for the decision-maker who faces no privacy constraint is the one that fully reveals (ω, θ). However, this signal would not be privacy-preserving if θ is non-degenerate or if ω and θ are correlated. From Theorem 1 and Blackwell's theorem, it follows that there always exists an optimal signal that is a reordered quantile signal. Therefore, it is without loss to restrict attention to these signals. For any reordered quantile signal $\pi^\star_\Phi$ and for any signal realization $s \in [0,1]$, the state ω is determined by θ through $\omega = F^{-1}(\Phi_\theta(s) \mid \theta)$. Thus,

$$\mathbb{P}_{\pi^\star_\Phi}[\omega \in A, \theta = \theta_j \mid s] = \mathbf{1}\{F^{-1}(\Phi_{\theta_j}(s) \mid \theta_j) \in A\} \cdot \mathbb{P}[\theta = \theta_j],$$

for all $s \in [0,1]$, for all measurable $A \subseteq \Omega$, and for all $j \in \{1, \dots, J\}$.

As a result, the distribution of posteriors over $\Omega \times \Theta$ induced by $\pi^\star_\Phi$ can be summarized by the joint distribution of $(\tilde{\omega}_j)_{j=1}^J \in \Omega^J$ with

$$\tilde{\omega}_j = F^{-1}(\Phi_{\theta_j}(s) \mid \theta_j),$$

where s is uniformly distributed on $[0,1]$. Here, the i-th component of $(\tilde{\omega}_j)_{j=1}^J$ indicates the state revealed by the signal realization s conditional on $\theta = \theta_i$.

One obstacle to finding optimal privacy-preserving signals is that it involves an optimization over measure preserving transformations Φ. We next establish that this problem is equivalent to an optimal transport problem.

Let $\mathcal{M}$ be the set of joint distributions ρ on Ω^J such that the marginal of the j-th coordinate equals $F(\cdot \mid \theta_j)$. The next lemma shows that $\mathcal{M}$ characterizes the distribution of posteriors over $\Omega \times \Theta$ induced by all reordered quantile signals.

Lemma 1. *Let s be a the random variable uniformly drawn from $[0,1]$. For any $\rho \in \Delta(\Omega^J)$, ρ is the joint distribution of $(F^{-1}(\Phi_{\theta_j}(s) \mid \theta_j))_{j=1}^J$ for some family $\Phi = \{\Phi_{\theta_j}\}_{j=1}^J$ of measure-*

preserving transformations if and only if $\rho \in \mathcal{M}$.

Intuitively, we have established that the marginal distribution over posterior beliefs conditional on each characteristic θ is the same for each Blackwell undominated privacy-preserving signal. Thus, the only dimension on which a designer needs to optimize is the correlation structure across different components of the agent's belief. The above result establishes that indeed each correlation structure can be generated by some collection of measure preserving transformations. Thus, we might optimize over correlation structures instead of measure preserving transformations.

Optimal Privacy-Preserving Signal With Lemma 1, we can now characterize the optimal privacy-preserving signals for the decision-maker. To this end, let $V^\star$ be the optimal value of the decision-maker. That is,

$$V^\star = \sup_{\pi} \left\{ \mathbb{E}_\pi \left[\sup_{a \in A} \mathbb{E}_\pi [u(\omega, \theta, a) \mid s] \right] \right\},$$

where the first supremum is taken over all privacy-preserving signals.[10] Moreover, let $V : \Omega^J \to \mathbb{R}$ be defined as

$$V(\omega_1, \ldots, \omega_J) := \sup_{a \in A} \left(\sum_{j=1}^{J} u(\omega_j, \theta_j, a) \mathbb{P}[\theta = \theta_j] \right), \tag{4}$$

for all $(\omega_j)_{j=1}^J \in \Omega^J$. We then have the following characterization:

Proposition 2 (Optimal Privacy-Preserving Signal)**.** *The decision-maker's optimal value* $V^\star$ *among all privacy-preserving signals is given by*

$$V^\star = \sup_{\rho \in \mathcal{M}} \int_{\Omega^J} V(\omega_1, \ldots, \omega_J) \, \mathrm{d}\rho. \tag{5}$$

Moreover, any optimal privacy-preserving signal must be Blackwell-equivalent to a reordered quantile signal $\pi^\star_\Phi$ *such that the distribution of* $(F^{-1}(\Phi_{\theta_j}(s) \mid \theta_j))_{j=1}^J$ *is a solution of* (5)*, where* s *is a random variable uniformly drawn from* $[0, 1]$.

[10] From Theorem 1, the set of privacy-preserving signals, up to Blackwell-equivalent classes, is well-defined.

The optimization problem (5) is a multi-marginal optimal transport problem. The existence of solutions can be guaranteed if Ω is compact and if V is upper-semicontinuous. According to Sklar's theorem, the feasible set $\mathcal{M}$ of (5) can be represented by all the copulas on $[0,1]^J$. Therefore, one optimal privacy preserving signal must correspond to an extreme point of the set of copulas. There has been recent progress in mathematics towards a characterization of these extreme points (see, for instance, Ghosh and Bhandari 2017 and Perronea and Durante 2021).[11] In the special case where $J = 2$, (5) becomes the classical Kantorovich optimal transport problem, whose dual problem is given by

$$\begin{aligned} &\inf_{L,K} \left(\int L(\omega_1)\, \mathrm{d}F(\omega_1|\theta_2) + \int K(\omega_2)\, \mathrm{d}F(\omega_2|\theta_2) \right) \\ &\text{s.t. } L(\omega_1) + K(\omega_2) \geq V(\omega_1, \omega_2)\,,\ \forall (\omega_1, \omega_2) \in \Omega^2 \end{aligned} \tag{6}$$

where the infimum is taken over all bounded continuous functions on Ω. In particular, if Ω is compact and if V is upper-semicontinuous, then by Theorem 1.46 of Santambrogio (2015), the value of the dual problem (6) is $V^\star$.

Increasing Difference Payoffs While Proposition 2 characterizes the decision-maker's optimal privacy-preserving signals by (5) for any payoff function, we may in fact obtain a closed-form solution for a specific class of decision-making problems.

Proposition 3. *Suppose that A is a totally ordered set and that $u : \Omega \times \Theta \times A \to \mathbb{R}$ has increasing difference in (ω, a). Moreover, suppose that*

$$\operatorname*{argmax}_{a \in A} \sum_{j=1}^{J} u(\omega_j, \theta_j, a) \mathbb{P}[\theta = \theta_j]$$

is nonempty for all $(\omega_j)_{j=1}^J \in \Omega^J$. Then the generalized quantile signal $\pi^\star$ is optimal for the decision-maker. That is,

$$V^\star = \mathbb{E}_{\pi^\star} \left[\sup_{a \in A} \mathbb{E}_{\pi^\star}[u(\omega, \theta, a) \mid q] \right].$$

[11] Recall that a copula on $[0,1]^J$ is a joint CDF C such that the marginal distribution on each dimension is uniform. Ghosh and Bhandari (2017) show that a copula $C : [0,1]^J \to [0,1]$ is an extreme point only if it is singular with respect to the Lebesgue measure on $[0,1]^J$. Moreover, C is an extreme point if it assigns probability 1 to a set $\{(\omega_1, \ldots, \omega_n) : \omega_i = g(\omega_{-i})\}$ for some $i \in \{1, \ldots, n\}$ and for some measurable function $g : [0,1]^{J-1} \to [0,1]$.

According to Proposition 3, for any decision-maker who chooses a one-dimensional action a to maximize a payoff $u(\omega, \theta, a)$ that has increasing difference in (ω, a), the generalized quantile signal is the optimal signal. For example, if the decision-maker seeks to minimize a loss function $|\omega - a|^p$, for some $p \in (1, \infty)$, by choosing an action $a \in \mathbb{R}$ to match the state $\omega \in \mathbb{R}$, then the generalized quantile signal is optimal. The assumption of increasing difference preferences is natural in many applications. For example, if ω measures the probability of a borrower repaying a loan and a is the interest rate a bank requires from a borrower, then it would be natural to assume that the bank wants to charge a lower interest rate to those borrowers who are more likely to repay.

However, as our next example shows, for preferences that do not exhibit increasing difference, the generalized quantile signal may not be optimal.

Example 2. Let $A = \{0, 1\}$, $\Omega = \{0, 1, 2\}$ and suppose that $\Theta = \{\theta_1, \theta_2\}$, with equal probability. Suppose that ω follows $F_1 = \frac{1}{2}\delta_{\{0\}} + \frac{1}{2}\delta_{\{1\}}$ if $\theta = \theta_1$, and follows $F_2 = \frac{1}{2}\delta_{\{1\}} + \frac{1}{2}\delta_{\{2\}}$ if $\theta = \theta_2$. The possible reordered quantile signals are convex combinations of the following two signals: either pooling agents of characteristic $\theta = \theta_1, \omega = 0$ with agents of characteristic $\theta = \theta_2, \omega = 1$, or agents of characteristic $\theta = \theta_2, \omega = 2$. These two signals generate payoffs proportional to:

$$\left(\max_a u(0, \theta_1, a) + u(1, \theta_2, a)\right) + \left(\max_a u(1, \theta_1, a) + u(2, \theta_2, a)\right);$$
$$\left(\max_a u(1, \theta_1, a) + u(1, \theta_2, a)\right) + \left(\max_a u(0, \theta_1, a) + u(2, \theta_2, a)\right),$$

respectively. Note that either of these (privacy-preserving) signals can be optimal: For $u(1, \theta, 1) = 1, u(1, \theta, 0) = u(1, \theta, 2) = -2$, the later signal is optimal and if $u(\omega, \theta, a) = -(a - \omega)^2$ the former signal, which corresponds to the general quantile signal, is optimal.

Binary Actions An important special case is when the decision is only between two actions, $A = \{0, 1\}$ (e.g., a bank decides whether to extend a loan at an exogenously fixed interest rate). In this case, define $\Phi_\theta : [0, 1] \to [0, 1]$ to be any function such that for all $\theta \in \Theta$

$$\omega \mapsto u(\Phi_\theta(\omega), \theta, 1) - u(\Phi_\theta(\omega), \theta, 0) \text{ is non-decreasing.} \tag{7}$$

Then, $\Phi = \{\Phi_\theta\}_{\theta \in \Theta}$ is indeed the optimal family of measure preserving transformations.

Corollary 1. *Suppose that $A = \{0,1\}$. Then any reordered quantile signal $\pi^\star_\Phi$ such that Φ solves (7) is an optimal privacy-preserving signal.*

Corollary 1 follows immediately from Proposition 3. To see this, let $\tilde{\omega} = \Phi_\theta(\omega)$, the decision-maker's payoff then exhibits increasing difference $(\tilde{\omega}, a)$. Corollary 1 thus completely solves the design problem for binary actions, arbitrary payoffs and arbitrary privacy sets.

Characteristic-Independent Preferences Another natural assumption is that the payoff $u(\omega, \theta, a)$ is independent of the characteristic θ. For instance, in the aforementioned example of a bank extending a loan, if θ captures the race or gender of an applicant it is a natural assumption that the bank has no intrinsic preferences over race or gender. In this case the optimization problem reduces to

$$\sup_{\rho \in \mathcal{M}} \int_{\Omega^J} \sup_{a \in A} \left(\sum_{j=1}^{J} u(\omega_j, a) \mathbb{P}[\theta = \theta_j] \right) \mathrm{d}\rho \,.$$

As Example 2 shows, even if the utility does not depend on θ, the generalized quantile signal might not be optimal. However, our next result shows that if preferences are linear in the state in addition, then the generalized quantile signal is optimal.

Corollary 2. *If $u(\omega, \theta, a)$ is constant in θ and either (i) is affine in ω or (ii) the state is binary: $\omega \in \{0,1\}$, then the generalized quantile signal is optimal.*

This corollary follows directly from Theorem 2, which establishes that the generalized quantile signal is the signal that induces the most dispersion in the posterior means. The assumption that payoffs are linear in the state is often satisfied in applications. For instance, in the bank loan example, it corresponds to assuming that the banks preference depends only on the expected amount repaid by the borrower. Since the linearity assumption is always satisfied if the state is binary, the second part of the corollary follows.

5 Economic Applications

To illustrate the relevance of privacy-preserving signals and to demonstrate the implications of our main results, we discuss several economic examples.

5.1 Algorithmic Fairness

Our first application pertains to "algorithmic fairness". The literature on algorithmic fairness in computer science and legal studies aims to explore optimal algorithms for decision-making under constraints that capture some notion of fairness, by which the literature means that people of different protected characteristics, such as race, gender, or sexual orientation are treated equally.[12] Our characterization of privacy-preserving signals implies a generalization of results in this literature under one of the most commonly adopted notions of fairness.

Specifically, the literature on algorithmic fairness considers a decision problem, also referred to as a classification problem, where there is an underlying outcome x. A decision-maker, or an algorithm, observes covariates (ω, θ) that are correlated with the outcome x and has to take an action $a \in \{0, 1\}$. The decision-maker has payoff $u(x, a) \in \mathbb{R}$ when the outcome is x and when the action is a. While the decision-maker can contingent their actions on the covariates (ω, θ) in principle, the actions taken must satisfy a fairness constraint. A commonly adopted notion of fairness in computer science is (conditional) independence,[13] which requires the action to be independent of the protected characteristics θ (conditional on materially relevant characteristics that are part of ω), and is also interpreted as preventing *disparate impact* in legal studies (see, e.g., Yang and Dobbie 2020).[14]

For example, suppose that a bank, who faces many loan applicants with observable characteristics (ω, θ), needs to make loan decisions a. The relevant outcome is whether an applicant will default in the future, denoted by $x \in \{0, 1\}$. The Equal Credit Opportunity Act (15 U.S.C. 1691 et seq.)[15]

[12] To avoid confusion we follow the computer-science literature in calling decisions that do not discriminate based on certain characteristics "fair", even though "non-discriminatory" might be more descriptive.

[13] See, e.g., Darlington (1971); Calders et al. (2009); Dwork, Hardt, Pitassi, Reingold and Zemel (2012); Calders and Verwer (2010); Kamishima, Akaho and Sakuma (2011); Corbett-Davies et al. (2017); Kitagawa et al. (2021); Gillis et al. (2021).

[14] This notion is also referred to as demographic parity, statistical parity, or group fairness. Another two commonly adopted criteria are (i) *separation*, which requires balanced type-I and type-II errors (see, e.g., Hardt, Price and Srebro 2016), or more generally, independence between a and θ conditional on the true outcome ω; and (ii) *sufficiency*, which requires the action a to be a sufficient statistics for ω, so that the outcome ω is independent of θ conditional on a. It is well-known that none of any pairs of these three common fairness criteria can be satisfied at the same time, and hence the choice of a fairness criteria is necessary (see Barocas, Hardt and Narayanan (2019) and Carey and Wu (2023) for a comprehensive review of these criteria). With appropriate projections, our results can also be applied when the notion of separation, instead of independence, is adopted. See §7 for more details.

[15] See `https://www.justice.gov/crt/equal-credit-opportunity-act-3`.

> "prohibits creditors from discriminating against credit applicants on the basis of race, color, religion, national origin, sex, marital status, age [..]."

A concrete (although stringent) interpretation of this requirement taken in the algorithmic fairness literature is that the information about each individual's default probability the bank uses to make loan decisions must be independent of an individual's protected characteristics (potentially conditional on materially relevant information, such as income). This requirement avoids the problem that even when restricting the bank to not condition on race directly it might still do so indirectly through the use of covariates such as zip code. This is a well-known issue highlighted in the actuarial sciences and legal studies, for example Wiggins (2020) states:[16]

> "[...] race has become so highly correlated with other social statistics that actuarial science in general has developed a baked-in racial bias. Racial discrimination by proxy (e.g., zip code standing in for race) can be glimpsed in the disparate impact of data-driven decision-making in housing, healthcare, policing, sentencing, and more. Simply leaving out racial data in statistically aided decision-making distances institutions from claims of intentional discrimination, but a disparate, discriminatory impact lingers when other factors correlated with race power actuarial analyses."

Existing results in the fairness literature (see, e.g., Calders et al. 2009; Hardt et al. 2016; Corbett-Davies et al. 2017) solve the decision-maker's constraint optimization problem and characterize the optimal fair algorithm in a simple setting when the decision-maker's choice is binary, e.g., when the bank only decides whether to grant a loan, and when the payoff is given by $u(x,a) = a \cdot (1-x-c)$ for some $c \in (0,1)$. The optimal algorithm adopts different thresholds for different groups θ (conditional on materially relevant characteristics), and chooses action $a = 1$, e.g., grants the loan to an applicant, if and only if the conditional expectation $\mathbb{E}[x \mid \omega, \theta]$, e.g., expected default probability, is below their group-specific thresholds.

[16] Former Attorney Eric Holder also made a similar remark in the context of sentencing: "[...] basing sentencing decisions on static factors and immutable characteristics—like the defendant's education level, socioeconomic background, or neighborhood—they may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society." See https://www.justice.gov/opa/speech/attorney-general-eric-holder-speaks-national-association-criminal-defense-lawyers-57th.

However, in many applications, the payoff-relevant outcomes might be richer, the decision-maker may need to make more than a binary choice, and might care about protected characteristics θ directly. For example, a bank typically needs to decide—in addition to whether to grant the loan—how much to grant, what the interest rate should be, the amount of down payment, and the form of the collateral. What would optimal fair algorithms be for a general decision-making problem? With more than binary actions, the underlying decision-making problem may be highly complex. Using existing methods to characterize optimal fair algorithms in a systematic way would be challenging, as they rely on the specific payoff structure and the simplicity of the action space. In fact, we are not aware of any paper solving explicitly a model with more than two actions.

Nonetheless, Theorem 1 and Theorem 2 lead to a simple and complete characterization of optimal algorithms for arbitrary decision problems. For any decision problem, let $\tilde{u}(\omega, \theta, a) := \mathbb{E}[u(x, a) \mid \omega, \theta]$, the optimal fair algorithms can be characterized by solving the linear program given in (5) with the payoff function being $\tilde{u}$.[17] If $\tilde{u}$ is either affine in ω or exhibits increasing difference in (ω, a), then the optimal algorithms must be equivalent to the optimal decision rule under the generalized quantile signal, according to Corollary 2. In particular, if the underlying outcome is binary: $x \in \{0, 1\}$, as commonly encountered in the context of algorithm-assisted decision making, by redefining $\omega \in [0, 1]$ as the conditional expected outcome given all available covariates and letting $\tilde{u}(\omega, \theta, a) := \omega u(1, a) + (1 - \omega)u(0, a)$ for all $a \in A$, Corollary 2 then implies that the optimal algorithms can be found by simply computing the optimal decision rule under the generalized quantile signal.[18]

This provides a simple and detail-free method to finding the optimal fair algorithm for a wide class of decision problems, as the generalized quantile signal does not depend on the structure of the underlying decision problem. In particular, we may recover the optimal group-specific threshold policy discovered in the literature by simply computing the optimal decision rule under the generalized quantile signal in the special case with a binary action and a characteristic-independent payoff that is monotone and affine in ω. Note, however, that in general the optimal decision rule might involve randomization inherited from the randomness

[17] Although Proposition 2 is stated under the assumption that $\omega \in \mathbb{R}$, it can be extended to the case of multi-dimensional ω, as we discuss in §6. Therefore, the covariates ω are allowed to be high-dimensional.

[18] This result generalizes the characterization of algorithmically fair optimal decisions obtained in Theorem 2 of He et al. (2023), which assumes the outcome x to be binary, and the covariates ω to be *perfectly informative* about the outcome of interest x.

of the generalized quantile signal.

A Regulation Procedure In practice, our results provide a simple and detail-free approach for regulating algorithm, such as loan decisions, product offers, bail decisions, etc to ensure that they satisfy fairness constraints while preserving as much prediction power as possible. A regulator can ensure fairness without any knowledge of the underlying decision problem, through the following regulatory procedure:[19]

1. Identify materially relevant characteristics (e.g., credit history or employment status).
2. Allow decision-makers to use any algorithms and any covariates to predict the relevant outcome ω.
3. Require a "post-processing step" after generating the raw prediction and *orthogonalize* these predictions, by computing the quantiles of the predicted outcome for each group of protected characteristic, conditional on the materially relevant characteristics.
4. Allow decision-makers to reorder, possibly randomly, the quantiles, in a way that the reordered quantiles remain uniformly distributed.
5. Decision-makers take an action based on the reordered quantiles.

By Theorem 1, decisions made under this procedure are the most efficient among those that meet the same fairness criteria, *regardless of the decision-maker's objectives*. Moreover, if the underlying decision-making problem has a payoff that is independent of θ and affine in ω, then Theorem 2 further allows the regulator to eliminate Step 4, so that decision-makers would take actions based only on the quantiles of each protected group.

The implementation of this procedure is practical since the regulator can monitor the decision-maker by simply observing the empirical joint distribution of decisions, protected characteristics, and materially relevant characteristics. In essence, this procedure does not regulate the *inputs* or *algorithms* themselves, but rather the *outputs* of the algorithm. It requires the decision-maker to make final decisions based on the orthogonalized predictions while permitting them to use any algorithm and all available inputs.

[19] Kamiran et al. (2013) and Feldman, Friedler, Moeller, Scheidegger and Venkatasubramanian (2015) propose a similar procedure to "repair" unfair algorithms. See also, Calders and Verwer (2010) and Kamishima et al. (2011), for earlier work on repairing and regularizing unfair algorithms. This literature focuses on transforming any (potentially unfair) algorithm into a fair one. Our results imply that, not only does a similar procedure lead to a fair algorithm, it is in fact the *optimal* way to transform any algorithm into a fair one.

5.2 Price Discrimination

In addition to algorithmic fairness, our results can be applied to settings of price discrimination in the spirit of Bergemann et al. (2015). Consider a monopolist who uses consumer data to price-discriminate consumers. The monopolist sells a single product to a unit mass of consumers. Each customer demands a single unit and has value $\omega \in \Omega := [\underline{\omega}, \overline{\omega}] \subset \mathbb{R}_+$ for the product. With different combinations of consumer data, the monopolist is able to charge different prices to different groups of consumers and engage in third-degree price discrimination.

While consumer data enables the monopolist to engage in price discrimination, it is often required by law or regulations that consumers cannot be price-discriminated based upon their protected characteristics. For example, the Civil Rights Act of the state of California prohibits businesses from engaging in "unlawful discrimination [...] based on a person's sex, race, color, religion, ancestry, national origin, age, disability, medical condition, genetic information, marital status, sexual orientation, citizenship, primary language, or immigration status".[20] Given such legal constraints, it is natural to ask: What market segmentations allow the monopolist to price-discriminate, but are not based on protected characteristics?

Clearly, the market segmentation that fully segments consumers by their values allows the monopolist to extract all the surplus. This, however, would typically lead to price discrimination based on protected characteristics, in the sense that consumer of different characteristics, e.g. race, would face a different distribution of prices. Moreover, simply prohibiting the monopolist from using protected characteristics to segment consumers would not be privacy-preserving either, since the monopolist may have access to close proxies of these characteristics. For example, as noted by The White House (2015):

> "Big data naturally raises concerns among groups that have historically been victims of discrimination. Given hundreds of variables to choose from, it is easy to imagine that statistical models could be used to hide more explicit forms of discrimination by generating customer segments that are closely correlated with race, gender, ethnicity, or religion [...], even if the profit motive is different from,

[20] Similarly, a recent legislation (AB1287) specifically prohibits businesses from price-discriminating based on gender. Likewise, the Genetic Information Nondiscrimination Act prohibits health insurers from using genetic information to "determine if someone is eligible for insurance or to make coverage, underwriting or premium-setting decisions".

and in many cases fundamentally inconsistent with, the sort of prejudice that our antidiscrimination laws seek to prohibit."

Our results lead to a characterization of all market segmentations that prohibit the monopolist from price-discriminating consumers based on their protected characteristics, in the sense that consumers of different protected characteristics face the *same* distribution of prices. By Theorem 1, a market segmentation is non-discriminatory in the this sense if and only if it corresponds to a garbling of some reordered quantile signal $\pi^\star_\Phi$.

Seller-Optimal Segmentations A natural question is what non-discriminatory market segmentation maximizes the seller's profit. Proposition 2 shows that this question reduces to an multi-marginal optimal transport problem. To simplify expositions, we consider the case of two protected characteristics in this section $\Theta = \{\theta_1, \theta_2\}$. Let

$$u(\omega, \theta, p) := \mathbf{1}\{\omega \geq p\}\, p$$

be the seller's profit. The optimal pricing problem is then a decision problem with payoff u, and the price charged to each segment of consumers is given by the optimal price given the signal realization that corresponds to this segment.

For any pair of consumer values $(\omega_1, \omega_2) \in [\underline{\omega}, \overline{\omega}]^2$, let $V(\omega_1, \omega_2)$ be the maximal profit obtainable by selling either to only one or both types of consumers

$$V(\omega_1, \omega_2) = \max\left\{\min\{\omega_1, \omega_2\}, \omega_1 \mathbb{P}[\theta = \theta_1], \omega_2 \mathbb{P}[\theta = \theta_2]\right\}\,.$$

Proposition 2 implies that the profit-maximizing market segmentation, among all market segmentations that prohibit price discrimination based on protected characteristics, can be identified by finding the joint distribution ρ of ω_1, ω_2 with marginals equal to the conditional distribution of buyer values $F(\cdot \mid \theta_1), F(\cdot \mid \theta_2)$ that solves the optimal transport problem

$$\sup_\rho \int V(\omega_1, \omega_2)\, \mathrm{d}\rho\,.$$

Suppose that both types are equally likely $\mathbb{P}[\theta = \theta_1] = \mathbb{P}[\theta = \theta_2] = 1/2$ and and the minimal

$\underline{\omega}$ and maximal willingness to pay $\overline{\omega}$ satisfy

$$2\,\underline{\omega} \geq \overline{\omega}\,.$$

In that case we have that $V(\omega_1, \omega_2) = \min\{\omega_1, \omega_2\}$. This corresponds to the classical Monge optimal transport problem with transport cost $|\omega_1 - \omega_2|$ for which the assortative assignment is optimal.[21]

Proposition 4. *Suppose that* $2\,\underline{\omega} \geq \overline{\omega}$ *and* $\mathbb{P}[\theta = \theta_1] = \mathbb{P}[\theta = \theta_2] = 1/2$.

(i) The market segmentation corresponding to the generalized quantile signal maximizes the seller's revenue.

(ii) The outcome is efficient and every consumer purchases the good.

(iii) If, furthermore, consumers with characteristic θ_1 *have higher values than those with characteristic* θ_2 *in the sense of FOSD*[22] *then consumers with characteristic* θ_2 *retain zero surplus under the seller-optimal market segmentation while consumers with characteristic* θ_1 *retain positive surplus.*

The first part of the result follows as $V(\omega_1, \omega_2) = \min\{\omega_1, \omega_2\}$ is supermodular and thus the assortative matching (which corresponds to the generalized quantile signal) is optimal (see, e.g., Lorenz 1949 and Theorem 3.12 of Rachev and Rüschendorf 1998). Moreover, under the assumption of the proposition, consumers' values are not too dispersed,. Thus, it always optimal to sell to both consumer types in each segment and charge the lower valuation. As a result, if the valuations are ordered in FOSD then one of the types will always have a lower value and thus make zero surplus, while the other type with some strictly higher valuation will be left with positive surplus. Thus, relative to the case where no constraints on market segmentation are imposed (and thus both types make zero surplus), it is not the weaker, but the stronger type who benefits from these constraints. This observation may serve as a cautionary tale, as in practice the legislations imposing privacy constraints typically mention explicitly that they are meant to protect groups that plausibly have lower willingness to pay.

When the assumptions in Proposition 4 are violated, V is not super-modular and the generalized quantile signal may not be not optimal. In this case, non-trivial reorderings of

[21] See the Online Appendix for detailed derivations.

[22] $F(\omega \mid \theta_1) \leq F(\omega \mid \theta_2)$, with strict inequality for a positive measure of ω.

quantiles are involved in the seller-optimal segmentation. As an example, suppose that the buyer value ω takes three possible values, 1, 2, or 3. Suppose that $\theta = \theta_1$ indicates a male person and $\theta = \theta_2$ indicates a female person, and both characteristics equally likely. The conditional distribution of ω given $\theta = \theta_1$ equals $(1/2, 1/3, 1/6)$; while the conditional distribution of ω given $\theta = \theta_2$ equals $(1/6, 1/3, 1/2)$. One can show that the solution to the above optimal transport problem is given by the joint distribution $\rho^\star$, where $\rho^*(1,1) = \rho^*(3,3) = 1/6$, $\rho^*(2,2) = \rho^*(1,3) = 1/3$.[23] Note that $\rho^\star$ corresponds to the reordered quantile signal generated by the measure-preserving transformations:

$$\Phi_{\theta_1}(q) = q \qquad \Phi_{\theta_2}(q) = \begin{cases} q, & \text{if } q \in [0, 1/6] \cup (5/6, 1] \\ 1/2 + (q - 1/6), & \text{if } q \in (1/6, 1/2] \\ 1/6 + (q - 1/2), & \text{if } q \in (1/2, 5/6] \end{cases}.$$

This signal corresponds to the (optimal) market segmentation described by Figure 1. Under this market segmentation, all the value $\omega = 2$ male and female consumers are pooled together in a segment; $1/3$ of the male consumers with value $\omega = 1$ and all the female consumers with value $\omega = 1$ are pooled together in a segment; $1/3$ of the female consumers with value $\omega = 3$ and all the male consumers with value $\omega = 3$ are pooled together in a segment; and the remaining male consumers with values $\omega = 1$ are pooled together in a segment with the remaining female consumers with values $\omega = 3$. Note that in each of these segments, the fractions of male and female consumers are exactly one-half, which are the same as the population frequency. Therefore, the monopolist is not able to price-discriminate based on gender under this market segmentation. Intuitively, this signal obtains a high revenue as it completely reveals the values of consumers with value 2, female consumer with value 1, and male consumers with value 3, and charges them their valuation. It pools male consumers of value 1 and female consumers of value 3, which ensures that the mechanism treats male and female consumers equally, but excludes 2/3 of the value 1 male consumer from consumption to do so. In this example, no consumer is better off due to the fact that both groups of consumers have to be treated equally.

[23] See the Online Appendix for detailed arguments.

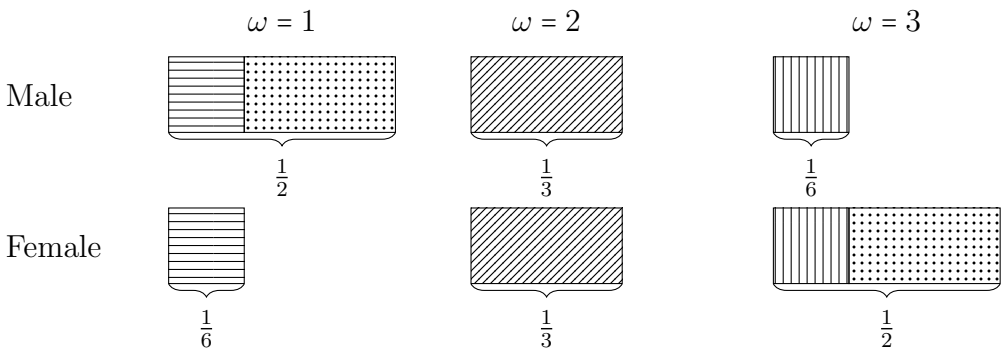

Figure 1: The profit-maximizing market segmentation. Segments of the customers which are shaded the same way are pooled together in the optimal privacy-preserving signal, and will hence face the same price. Columns signify the value of the consumer and rows their gender.

5.3 Private Private Information

Our results also lead to a generalization of an elegant recent result by He, Sandomirskiy and Tamuz (2023), henceforth HST. They ask, in a multi-agent setting, what private signals about a binary state have the property that they reveal no information about the signals received by other agents: $n \geq 2$ agents observe private signals about a binary state $\omega \in \{0,1\}$. An information structure is a random vector $(s_1, \ldots, s_n)$, whose distribution depends on the state, such that each agent i privately observes the realization of s_i. It is *private private* if the signals $s_1, \ldots, s_n$ are independent (unconditional on the state).

A main result of HST is a characterization of undominated private private information structures. A private private information structure is said to be (mean-)*undominated* if there does not exist any other private private information structure where every agent's distribution of posterior means is more dispersed (in the sense of mean-preserving spreads). HST characterize the undominated private private information structures in the setting where $\omega \in \{0,1\}$ and show that a private private information structure is undominated if and only if it corresponds to a *set of uniqueness* in the hypercube $[0,1]^n$. In the case when $n = 2$, as shown by Theorem 1 of HST, this is equivalent to saying that the distribution of posterior means of one agent equals the conjugate of the distribution of posterior means of the other agent.[24,25] HST also show how this result can be used to analyze algorithmic fairness similar

[24]This in turn generalizes Proposition 2 of Arieli, Babichenko, Sandomirskiy and Tamuz (2021)

[25]The conjugate of a CDF $G : [0,1] \to [0,1]$ is given by $\widehat{G}(x) := 1 - G^{-1}(1-x)$ for all $x \in [0,1]$.

to our application in §5.1.

We can encompass their notion of private private information by focusing on a given agent i and fixing the signals s_{-i} of other agents. An information structure is private private if and only if, for each agent i, their signal s_i is privacy-preserving with respect to the privacy sets generated by other agents' signals s_{-i}. The following corollary, which follows from Theorem 2, generalizes Theorem 1 of HST to more than two states and more than two agents, and characterizes the undominated private private information structures.[26]

Corollary 3. *A private private information structure $(s_1, \ldots, s_n)$ is undominated if and only if for every agent i, the distribution of their posterior mean $\mathbb{E}[\omega \mid s_i]$ is given by $\overline{F}_i$, where*

$$\overline{F}_i(x) := \inf \left\{ q \in [0,1] : \mathbb{E}\left[F^{-1}(q \mid s_{-i}) \right] \geq x \right\},$$

and $F(\cdot \mid s_{-i})$ is the conditional distribution of ω given s_{-i}.

When there are only two states $\omega \in \{0,1\}$ and two agents $n = 2$, $F(x \mid s_i) = 1 - \mathbb{E}[\omega \mid s_i]$ for all $x \in [0,1)$, and hence

$$F^{-1}(q \mid s_i) = \mathbf{1}\{\mathbb{E}[\omega \mid s_i] > 1 - q\}.$$

Since

$$\inf \left\{ q \in [0,1] : \mathbb{E}[F^{-1}(q \mid s_i)] \right\} = \inf \left\{ q \in [0,1] : \mathbb{E}[\mathbf{1}\{\mathbb{E}[\omega \mid s_i] > 1 - q\}] \right\}$$

is the conjugate of the distribution of $\mathbb{E}[\omega \mid s_i]$, Corollary 3 implies Theorem 1 of HST.[27]

[26] When $\omega \in \{0,1\}$, the notion of dominance defined in terms of distributions of posterior means is equivalent to dominance in terms of Blackwell's order. In general, the Blackwell order is coarser than the mean-preserving contraction order of distributions of posterior means. Theorem 5 of HST, which can be found in their appendix, shows that when the state space is finite, Blackwell-undominated private private information structures are equivalent to *partitions of uniqueness* on the hypercube $[0,1]^n$. This, however, as they note, is a complex mathematical object. Our Theorem 1 alone does not generalize their Theorem 5 in the appendix. Doing so requires a characterization of the maximal element of distributions over posteriors over Ω under the mean-preserving spread order. Although our Theorem 1 implies a characterization of maximal elements of the distributions of posteriors over $\Omega \times \Theta$, it does not lead to a characterization of the maximal elements of the distributions of posteriors over Ω. Generalizations and simplifications in this direction remain as open questions.

[27] Our Theorem 1 can be used to obtain a characterization of feasible distributions of posterior beliefs in addition to the undominated distributions of posterior means.

5.4 Bayesian Persuasion

Consider the Bayesian persuasion setting where a sender discloses information about (ω, θ) to a receiver who chooses an action a, and suppose that the sender is restricted to choose only signals that are privacy-preserving. For example, the sender might be the prosecutor as in Kamenica and Gentzkow (2011), trying to convince the judge that the defendant should not be released on bail, but is restricted to not using any information related to the race of the defendant even though such information might be predictive about the probability of reoffense. Let the sender's payoff be $u_S : \Omega \times \Theta \times A \to \mathbb{R}$ and the receiver's payoff be $u_R : \Omega \times \Theta \times A \to \mathbb{R}$.[28] Let $V_S^\star$ be the sender's value from choosing the optimal privacy-preserving signal. For simplicity, suppose again that $|\Theta| = J < \infty$ and write Θ as $\{\theta_1, \ldots, \theta_J\}$. Let $V_R : \Omega^J \times A \to \mathbb{R}$ be defined as

$$V_R(\omega_1, \ldots, \omega_J, a) := \sum_{j=1}^{J} u_R(\omega_j, \theta_j, a)\mathbb{P}[\theta = \theta_j],$$

for any $(\omega_j)_{j=1}^J \in \Omega^J$. Moreover, for any $\rho \in \Delta(\Omega^J)$, let

$$V_S(\rho) := \mathbb{E}_\rho\left[\sum_{j=1}^{J} u_S(\omega_j, \theta_j, a^\star(\rho))\mathbb{P}[\theta = \theta_j]\right],$$

where $a^\star(\rho)$ is the (sender-preferred) optimal action of the receiver that maximizes V_R when the posterior over $(\omega_j)_{j=1}^J$ is ρ. To ensure the existence of optimal signals, we assume that Ω is compact and that V_S is upper-semicontinuous. The next proposition characterizes the sender's value $V_S^\star$.

Proposition 5 (Value of Persuasion)**.** *Let $\overline{V}_S$ be the concave closure of V_S, the sender's value $V_S^\star$ is given by*

$$V_S^\star = \max_{\rho \in \mathcal{M}} \overline{V}_S(\rho),$$

where $\mathcal{M}$ is the set of joint distributions on Ω^J such that the marginal of the j-th coordinate equals $F(\cdot \mid \theta_j)$.

[28] Recall that Kamenica and Gentzkow (2011) characterize the sender's optimal value over all signals (including non-privacy-preserving ones) as the concave closure of the sender's indirect payoff as a function of posterior beliefs.

Proposition 5 states that the sender's value can be found by a two-step procedure: First, fix a joint distribution ρ and find the optimal garbling of it by computing $\overline{V}_S(\rho)$. Then, optimize across all reordered canonical signals.

Just as in standard persuasion problems, the characterization of Proposition 5 requires computing the concave closure of the function V_S, which is typically computationally demanding. Nonetheless, when payoffs are such that the sender's indirect utility is measurable with respect to the posterior mean, Theorem 2 provides a tractable way to characterize optimal signals. Specifically, suppose that the sender's indirect utility is a function only of the posterior mean $\mathbb{E}[\omega \mid s]$, which we denote by $U_S : \mathbb{R} \to \mathbb{R}$. Then the sender's payoff given a signal can be written as $\int_{\mathbb{R}} U_S(x)\,\mathrm{d}G$, where G is the CDF of posterior means induced by the signal. Theorem 2 implies the following characterization:

Proposition 6 (Value of Mean-Measurable Persuasion)**.** *Suppose that the sender's indirect utility is measurable with respect to posterior means and is denoted by $U_S : \mathbb{R} \to \mathbb{R}$. The sender's value $V_S^\star$ is given by*

$$V_S^\star = \sup_{G \preceq_{\mathrm{MPS}} \overline{F}} \int_{\mathbb{R}} U_S(x)\,\mathrm{d}G\,, \tag{8}$$

As a result, since the objective of (8) is affine and since the feasible set is convex, one of the solutions must be an extreme point of the feasible set. If U_S is upper-semicontinuous and if $\overline{F}$ is continuous, then by Theorem 2 of Kleiner et al. (2021), there must be a solution $G^\star$ that takes the form of

$$G^\star(x) = \begin{cases} \overline{F}(x), & \text{if } x \notin \bigcup_{i\in I}[\underline{x}_i, \overline{x}_i) \\ \overline{F}(\underline{x}_i), & \text{if } x \in [\underline{x}_i, \underline{y}_i) \\ x_i, & \text{if } x \in [\underline{y}_i, \overline{y}_i) \\ \overline{F}(\overline{x}_i), & \text{if } x \in [\overline{y}_i, \overline{x}_i) \end{cases},$$

for some collection of intervals $\{[\underline{x}_i, \overline{x}_i]\}_{i\in I}$ and $\{[\underline{y}_i, \overline{y}_i]\}_{i\in I}$ such that $[\underline{y}_i, \overline{y}_i] \subseteq [\underline{x}_i, \overline{x}_i]$, and for some $\{x_i\}_{i\in I} \subseteq \mathbb{R}$. Furthermore, $G^\star$ can be implemented by fulling revealing the realizations q of the generalized quantile signal whenever $q \notin \cup_{i\in I}[\underline{x}_i, \overline{x}_i)$, while pooling at most two signal realizations in each interval $[\underline{x}_i, \overline{x}_i)$ for all $i \in I$.

6 General Result and its Proof

In this section, we state the general version of our main result, which does not require Ω to be a subset of $\mathbb{R}$. To begin with, we introduce a family of signals that are analogous to the reordered quantile signals when Ω is not necessarily one-dimensional.

6.1 Canonical Signals and Reordered Canonical Signals

We first show that it is without loss to focus on the case where for each given value of θ, the state ω can be expressed as a random variable on a probability space where the only randomness is generated by drawing θ and an *independently* distributed uniformly distributed random variable.[29] The next lemma shows that this reduction is indeed without loss of generality. To this end, let λ denote the Lebesgue measure on $[0,1]$, and let ν_Θ be the marginal distribution of θ.

Lemma 2. *There exists a random variable $\omega^\star : [0,1] \times \Theta \to \Omega$ such that $(\omega, \theta(\omega))$ and $(\omega^\star(q,\theta),\theta)$ have the same distribution, where ω is distributed according to $\mathbb{P}$ and (q,θ) is distributed according to $\lambda \times \nu_\Theta$.*[30]

We henceforth refer to $[0,1]\times\Theta$ as the *canonical state space*, and to the random variable $\omega^\star : [0,1]\times\Theta \to \Omega$ as the *canonical map*. Note that when $\Omega \subseteq \mathbb{R}$, the canonical map can be chosen to be F^{-1}.

Reordered Canonical Signals By normalizing the state to the canonical state space, a privacy-preserving signal is naturally defined: The signal $\pi^\star$ that reveals the underlying noise q of the canonical state space conditional on $\omega^\star(q,\theta) = \omega$ and θ is, by definition, independent of θ.[31] By Proposition 1, it is thus privacy-preserving. We refer to the signal $\pi^\star$ as the *canonical signal*.

[29] Here we follow ideas from von Neumann (1932) and Rokhlin (1952) who show that for standard probability spaces it is without loss to assume that all randomness is generated by a uniform random variable.

[30] More precisely, there exists a random variable $\omega^\star : [0,1]\times\Theta \to \Omega$ defined on the probability space $([0,1]\times\Theta, \mathcal{B}\otimes\mathcal{G}, \lambda\times\nu_\Theta)$ such that $\mathbb{P}[(\omega,\theta(\omega)) \in A\times B] = \lambda\times\nu_\Theta(\{\omega^\star(q,\theta),\theta) \in A\times B\})$ for all $A\in\mathcal{F}$, $B\in\mathcal{G}$, where $\mathcal{B}$ is the Borel σ-algebra on $[0,1]$ and $\mathcal{G}$ is the σ-algebra of Θ.

[31] More formally, the conditional distribution $\pi^\star_{(\omega,\theta)}$ is defined as the transition probability implied by the joint distribution of $(\omega^\star(q,\theta),\theta,q)$ on the canonical state space. The transition probability exists due to the disintegration theorem (c.f., Çinlar (2010) pp.154)

We now define a class of signals that are analogous to the reordered quantile signals. For any family $\Phi = \{\Phi_\theta\}_{\theta\in\Theta}$ of measure-preserving transformations that preserve the Lebesgue measure and for any realization (ω, θ), draw the canonical signal q from $\pi^\star_{(\omega,\theta)}$. Then, for each realized generalized quantile q, further draw a signal $s \in \Phi_\theta^{-1}(q)$ randomly so that the distribution of s conditional on θ is uniform.[32] This defines a privacy-preserving signal, which is denoted by $\pi^\star_\Phi$, and referred to as the Φ-*reordered canonical signal*. Just like reordered quantile signals, reordered canonical signals are obtained by (randomly) reordering the canonical signal in a way that preserves the uniform measure.

6.2 Characterization of Privacy-Preserving Signals

We now present our main result, which generalizes Theorem 1 and characterizes what information can be revealed by a privacy-preserving signal.

Theorem 3 (Characterization of Privacy-Preserving Signals)**.**

(i) A signal is privacy-preserving if and only if it is Blackwell dominated by some reordered canonical signal $\pi^\star_\Phi$.

(ii) Every reordered canonical signal is Blackwell undominated among privacy-preserving signals.

When $\Omega \subseteq \mathbb{R}$, the canonical map $\omega^\star$ equals the generalized quantile function, and any Φ-reordered canonical signal is the Φ-reordered quantile signal. Thus, Theorem 3 generalizes Theorem 1.

6.3 Proof of Theorem 3

We now provide a proof of Theorem 3, which in turn implies Theorem 1. The proof consists of several lemmas. Proofs of these lemmas can be found in the Appendix.

To begin with, we first argue that any garbling of a privacy-preserving signal remains privacy-preserving, which, together with the fact that any reordered canonical signal is privacy-preserving, proves the sufficiency part of (i).

Lemma 3. *Any signal that is Blackwell dominated by a privacy-preserving signal is also privacy-preserving.*

[32] As noted in footnote 7, such signal is well-defined and perfectly reveals q.

Next, we prove the necessity part of (i). To this end, we first introduce a class of signals referred to as *conditionally revealing* signals. These are signals that fully reveal the state ω conditional on θ. Secondly, we show that every privacy-preserving signal is a garbling of a conditionally revealing privacy-preserving signal. Lastly, we argue that every conditionally revealing privacy-preserving signal is a reordered canonical signal.

Definition 3 (Conditionally Revealing Signals)**.** A privacy-preserving signal π is *conditionally revealing* if ω is fully revealed by s conditional on θ. That is, there exists a measurable function $\eta : S \times \Theta \to \Omega$ such that, for almost all $s \in S$, $\omega = \eta(s, \theta)$ with $\mathbb{P}_\pi[\cdot \mid s]$-probability 1.

Note that the canonical signal is conditionally revealing, with η being the canonical map $\omega^\star$. Conditionally revealing signals fully reveal ω if θ is known. Since a conditionally revealing privacy-preserving signal fully reveals ω conditional on θ, every privacy-preserving signal that has residual noise even if θ is known must be less informative than a conditionally revealing signal. The next proposition formalizes this intuition.

Lemma 4. *Every privacy-preserving signal is a garbling of some conditionally revealing privacy-preserving signal.*

Finally, we argue in Lemma 5 below that any conditionally revealing signal can be generated by reordering the canonical signal for each $\theta \in \Theta$.

Lemma 5. *Every conditionally revealing privacy-preserving signal is equivalent to a reordered canonical signal.*

Combining lemmas 3 through 5, we now prove Theorem 3.

Proof of Theorem 3.
Part (i): For sufficiency, as every reordered canonical signal is privacy-preserving, any garbling of the reordered canonical signal is also privacy-preserving, by Lemma 3. For necessity, consider any privacy-preserving signal π. Lemma 4 implies that π is Blackwell-less informative than a conditionally revealing privacy-preserving signal. From Lemma 5, this signal must be equivalent to a reordered canonical signal.

Part (ii): Suppose that π and π' are two Blackwell-nonequivalent reordered canonical signals, and suppose that π is Blackwell more informative than π'. By Lemma 5, both π and

π' are conditionally revealing privacy-preserving signals. However, since π is Blackwell more informative than π', π' is not conditionally revealing, a contradiction. Therefore, any two reordered canonical signals are either Blackwell-incomparable or Blackwell-equivalent.

For any reordered canonical signal π and any privacy-preserving signal $\hat{\pi}$. Part (i) implies that $\hat{\pi}$ must be a Blackwell-dominated by some reordered canonical signal π'. Since π and π' must be either Blackwell-equivalent or Blackwell-incomparable, π is not Blackwell-dominated by $\hat{\pi}$. □

7 Discussion

Relationship with Differential Privacy While we define privacy-preserving signals through an abstract collection of privacy sets, another notion of privacy is *differential privacy*. Specifically, suppose that Ω is a finite product set $\Omega_1 \times \cdots \times \Omega_n$, where each dimension Ω_i represents characteristics of a different agent. A signal π satisfies *ε-differential privacy* for $\varepsilon > 0$ if for every signal realization $s \in S$ and for any ω, ω' that differ only in the characteristic of a single agent i (i.e., $\omega_{-i} = \omega'_{-i}$),

$$\left| \log \frac{\mathbb{P}_\pi[\omega \mid s]}{\mathbb{P}_\pi[\omega' \mid s]} - \log \frac{\mathbb{P}[\omega]}{\mathbb{P}[\omega']} \right| \leq \varepsilon \,.$$

Intuitively, the log-likelihood induced by the signal cannot be influenced by more than ε by each individual agent. Our notion of privacy considers signals only depends on the characteristics of a single individual, and are restricted to not not reveal certain information. In contrast, differential privacy considers signals which depend on a whole population of agents, but who are only influenced to a limited extent by each individual agent. Mathematically, these notions are unrelated and aim to capture different aspects of privacy.

Separation as a Notion of Fairness In the literature on algorithmic fairness, there are other notions of fairness that do not require statistical independence, as discussed in §5. One of the most commonly used alternatives to statistical independence is called *separation*. Separation requires the decisions to be independent of protected characteristics *conditional on the true state*.

Our results can also be applied to this setting. To see this, suppose that the underlying outcome, z, is binary and takes values 0 or 1. Let ω be the expected probability of the

underlying state being $z = 1$, conditional on all the observable covariates (including protected characteristics θ). A signal would satisfy the requirement of separation if its realization is independent of θ conditional on z. Consider any conditionally privacy-preserving signals π for the extended state space (ω, θ, z). By definition, signal realizations s drawn from π would be independent of θ conditional on z. Moreover, a conditionally privacy-preserving signal is Blackwell-undominated if and only if it takes the form of (s, z), where s is drawn from some reordered canonical signal π^*_Φ conditional on z. Although the signal that reveals (s, z) may not be feasible, as the outcome z is typically unknown, one can project this signal by computing the conditional expectation of (s, z) given ω. This conditional expectation is thus, by construction, a garbling of ω, and is conditionally independent of θ given z. Furthermore, since taking the conditional expectation preserves the Blackwell order, this signal must remain Blackwell-undominated among all feasible signals.

8 Conclusion

We provide a characterization of signals which do not reveal certain information, and among others presented application to price discrimination, and algorithmic fairness. We believe the mathematical characterization of privacy preserving signals can be useful in other contexts. For instance, we conjecture that our results can be used to prove generalizations of Border's theorem (Border 1991; Hart and Reny 2015). Another interesting avenue for future research is to use the mathematical characterization presented in this paper to understand the consequences of different notion of privacy and fairness.

Appendix

Lemma A.1. *A signal is privacy-preserving with respect to $\mathcal{P} \subseteq \mathcal{F}$ if and only if it is privacy-preserving with respect to the σ-algebra generated by $\mathcal{P}$.*

Proof. Fix any nonempty collection $\mathcal{P} \subseteq \mathcal{F}$. Consider any signal π that is privacy-preserving with respect to the σ-algebra generated by $\mathcal{P}$, denoted by $\sigma(\mathcal{P})$. Since $\mathcal{P} \subseteq \sigma(\mathcal{P})$, π is privacy-preserving with respect to $\mathcal{P}$. Conversely, consider any signal π that is privacy-preserving with respect to $\mathcal{P}$. Let $\mathcal{P}^\pi \subseteq \mathcal{F}$ be the collection of events for which (1) holds for

all signal realizations s. Clearly, $\mathcal{P}^\pi$ is nonempty since π is privacy-preserving with respect to $\mathcal{P}$. Moreover, from the facts that $\mathbb{P}$ and $\mathbb{P}[\cdot \mid s]$ are probability measures for all $s \in S$, it follows that $\mathcal{P}^\pi$ is a λ-system. Therefore, by Dynkin's $\pi - \lambda$ theorem, since the π-system $\mathcal{P}$ is contained in the λ-system $\mathcal{P}^\pi$, the σ-algebra $\sigma(\mathcal{P})$ generated by $\mathcal{P}$ must also be contained in $\mathcal{P}^\pi$. Therefore, π is privacy-preserving with respect to $\sigma(\mathcal{P})$. □

Proof of Proposition 1. By Lemma A.1, it is without loss to assume that $\mathcal{P}$ is a σ-algebra. Let $\Theta := \Omega$, $\mathcal{G} := \mathcal{P}$, $(\Theta, \mathcal{G})$ is a measurable space. Moreover, let $\theta : (\Omega, \mathcal{F}) \to (\Theta, \mathcal{G})$ be the identity function, i.e., $\theta(\omega) := \omega$ for all $\omega \in \Omega$. Clearly, θ is measurable since for any $B \in \mathcal{G} = \mathcal{P}$, $\theta^{-1}(B) := \{\omega \in \Omega : \theta(\omega) \in B\} = B \in \mathcal{F}$. Furthermore, since for any $B \in \mathcal{P}$, $\theta^{-1}(B) = B$, it must be that $\sigma(\theta) = \mathcal{P}$. Lastly, for any signal π and for any $B \in \mathcal{P}$, since $(\Omega, \mathcal{F}, \mathbb{P})$ is a standard probability space, the conditional probability $\mathbb{P}_\pi[\theta \in B \mid s]$ is well-defined. Moreover,

$$\mathbb{P}_\pi[\theta \in B \mid s] = \mathbb{P}_\pi[\omega \in \theta^{-1}(B) \mid s] = \mathbb{P}_\pi[\omega \in B \mid s]\,.$$

Therefore, π is privacy-preserving if and only if for all $B \in \mathcal{G} = \mathcal{P}$,

$$\mathbb{P}_\pi[\theta \in B \mid s] = \mathbb{P}[\theta \in B]\,.$$ □

Proof of Theorem 2. For any privacy-preserving signals π, π', let G, G' be the distribution of posterior means induced by π, π', respectively. Then $G \preceq_{\text{MPS}} G'$ whenever π' Blackwell dominates π. Since $\overline{F}$ can be induced by the generalized quantile signal, Lemma 3 implies that every mean-preserving contraction of $\overline{F}$ can also be induced by a privacy-preserving signal. Thus, by Theorem 1, it suffices to show that, for any family $\Phi = \{\Phi_\theta\}_{\theta \in \Theta}$ of measure-preserving transformations, the distribution G of posterior means induced by $\pi^\star_\Phi$ is a mean-preserving contraction of $\overline{F}$. To see this, observe that the posterior mean after observing a signal realization s drawn from $\pi^\star_\Phi$ is given by

$$\mathbb{E}[F^{-1}(\Phi_\theta(s) \mid \theta)]\,. \tag{A.9}$$

Note, that the function $s \mapsto \mathbb{E}[F^{-1}(\Phi_\theta(s) \mid \theta)]$ is not necessarily monotone. The generalized quantile function G^{-1} of the posterior mean is given by the monotone rearrangement of the

above function. If we denote this rearrangement by $\psi : [0,1] \to [0,1]$, we have that

$$G^{-1}(s) = \mathbb{E}[F^{-1}(\Phi_\theta(\psi(s)) \mid \theta)]\,.$$

As $\Phi_\theta \circ \psi$ is a measure-preserving transformation, we have that for all $t \in [0,1]$,

$$\int_t^1 G^{-1}(s)\,\mathrm{d}s = \int_t^1 \mathbb{E}[F^{-1}(\Phi_\theta \circ \psi(s) \mid \theta)]\,\mathrm{d}s \leq \int_t^1 \mathbb{E}[F^{-1}(s \mid \theta)]\,\mathrm{d}s = \int_t^1 \overline{F}^{-1}(s)\,\mathrm{d}s\,.$$

The inequality in the above equation follows since for any value of $t \in [0,1]$ the measure-preserving transformation $\Phi_\theta \circ \psi$ that maximizes the above integral is the identity. The above equation shows that G^{-1} is majorized by $\overline{F}^{-1}$, which implies that G is majorized by $\overline{F}$ (see, for example, Shaked and Shanthikumar (2007), Section 3.A), which implies that G is a mean-preserving spread of $\overline{F}$. □

Proof of Lemma 2. By definition, θ has the same distribution in both cases. We thus only need show that for each fixed θ there exists a random variable $\omega^\star(q,\theta)$ such that if q is uniformly distributed $\omega^\star$ and ω have the same distribution conditional on θ. To this end, let $\phi : \Omega \to [0,1]$ be the Borel isomorphism. Since $(\Omega, \mathcal{F})$ is a standard Borel space, ϕ is well-defined and both ϕ and ϕ^{-1} are measurable. For any $\theta \in \Theta$, let $\nu(C \mid \theta) := \mathbb{P}[\{\omega \in \Omega : \phi(\omega) \in C\} \mid \theta]$, for all Borel set $C \subseteq [0,1]$. Then, for $F^{-1}(q \mid \theta)$ defined as

$$F^{-1}(q \mid \theta) := \inf\{x \in [0,1] : \nu([0,x] \mid \theta) \geq q\}\,,$$

for all $q \in [0,1]$ and for all $\theta \in \Theta$, both $q \mapsto F^{-1}(q \mid \theta)$ and $\theta \mapsto F^{-1}(q \mid \theta)$ are measurable. Moreover,

$$\lambda(\{q \in [0,1] : F^{-1}(q \mid \theta) \in C\}) = \nu(C \mid \theta)\,,$$

for all Borel-measurable set $C \subseteq [0,1]$ and for all $\theta \in \Theta$. Now let $\omega^\star(q,\theta) : [0,1] \times \Theta \to \Omega$ be defined as $\omega^\star(q,\theta) := \phi^{-1}(F^{-1}(q \mid \theta))$. Then $\omega^\star$ is measurable. Moreover, for any $\theta \in \Theta$ and for any measurable $A \in \mathcal{F}$,

$$\begin{aligned}\lambda(\{q \in [0,1] : \omega^\star(q,\theta) \in A\}) &= \lambda(\{q \in [0,1] : \phi^{-1}(F^{-1}(q \mid \theta)) \in A\}) = \lambda(\{q \in [0,1] : F^{-1}(q \mid \theta) \in \phi(A)\}) \\ &= \nu(\phi(A) \mid \theta) = \mathbb{P}[\phi^{-1}(\phi(A)) \mid \theta] = \mathbb{P}[A \mid \theta]\,,\end{aligned}$$

as desired. □

Proof of Lemma 3. Suppose that π is Blackwell-more informative than π'. Fix any privacy set $P \in \mathcal{P}$ and consider the decision problem where $A = [0,1]$ and $u(a,\omega) = -(\mathbf{1}\{\omega \in P\} - a)^2$. Taking the first order condition yields that the unique optimum in the optimization problem

$$\max_{a \in [0,1]} \mathbb{E}_\pi \left[-(\mathbf{1}\{\omega \in P\} - a)^2 \mid s \right] \tag{A.10}$$

is given by

$$a^\star(s \mid \pi) = \mathbb{E}_\pi \left[\mathbf{1}\{\omega \in P\} \mid s \right] = \mathbb{P}_\pi \left[\omega \in P \mid s \right] = \mathbb{P}\left[\omega \in P \right].$$

where the last equality in the above equation follows as π is privacy-preserving. The optimal action thus does not depend on the signal realization and consequently a decision-maker observing the signal π' can guarantee themselves the same expected payoff by using the constant action $\mathbb{P}[\omega \in P]$. As the expected payoff under the Blackwell dominated signal π' is weakly lower, it follows that $a = \mathbb{P}[\omega \in P]$ must be an optimal action under the signal π' for all signal realizations. Since the optimal action of the decision problem (A.10) is unique, it follows that

$$a^\star(s \mid \pi') = \mathbb{E}_{\pi'} \left[\mathbf{1}\{\omega \in P\} \mid s \right] = \mathbb{P}_{\pi'} \left[\omega \in P \mid s \right] = \mathbb{P}\left[\omega \in P \right]$$

which implies that π' is privacy-preserving. □

Proof of Lemma 4. Consider any privacy-preserving signal π, let $\gamma \in \Delta\Delta(\Omega \times \Theta)$ be the distribution over posteriors on $\Omega \times \Theta$ induced by π. Since π is privacy-preserving, for γ-almost all $\mu \in \Delta(\Omega \times \Theta)$, the marginal of μ on Θ must be ν_Θ. Since Ω is standard-Borel, there exists a transition probability $\tilde{\mu} : \Theta \to \Delta(\Omega)$ such that

$$\mu(A \times B) = \int_B \tilde{\mu}(A \mid \theta) \, \mathrm{d}\nu_\Theta,$$

for all measurable $A \subseteq \Omega$ and for all measurable $B \subseteq \Theta$.

Furthermore, since $(\Omega, \mathcal{F})$ is a standard Borel space, there exists a Borel isomorphism $\phi : \Omega \to [0,1]$ such that both ϕ and $\phi^{-1} : [0,1] \to \Omega$ are measurable. For any $\mu \in \Delta(\Omega \times \Theta)$ and for any $\theta \in \Theta$, let $F_\mu(C \mid \theta) := \tilde{\mu}(\{\omega \in \Omega : \phi(\omega) \in C\} \mid \theta)$ for all Borel measurable $C \subseteq [0,1]$.

Let

$$F_\mu^{-1}(q \mid \theta) := \inf\{x \in [0,1] : F_\mu([0,x] \mid \theta) \geq q\}\,,$$

for all $q \in [0,1]$, for all $\theta \in \Theta$, and for all $\mu \in \Delta(\Omega \times \Theta)$. Note that the functions $q \mapsto F_\mu^{-1}(q \mid \theta)$, $\theta \mapsto F_\mu^{-1}(q \mid \theta)$, and $\mu \mapsto F_\mu^{-1}(q \mid \theta)$ are measurable.

Now let $\tilde{\omega}^\mu(q,\theta) := \phi(F_\mu^{-1}(q \mid \theta))$ for all $q \in [0,1]$, for all $\theta \in \Theta$, and for all $\mu \in \Delta(\Omega \times \Theta)$. Then for all $\theta \in \Theta$ and for all measurable $A \in \mathcal{F}$,

$$\begin{aligned}\lambda(\{q \in [0,1] : \tilde{\omega}^\mu(q,\theta) \in A\}) =& \lambda(\{q \in [0,1] : \phi(F_\mu^{-1}(q \mid \theta)) \in A\}) = \lambda(\{q \in [0,1] : F_\mu^{-1}(q \mid \theta) \in \phi^{-1}(A)\}) \\ =& F_\mu(\phi^{-1}(A) \mid \theta)) = \tilde{\mu}(\phi(\phi^{-1}(A)) \mid \theta) = \tilde{\mu}(A \mid \theta)\,.\end{aligned}$$

That is, the conditional distribution of $\tilde{\omega}^\mu(q,\theta)$ given θ, equals $\tilde{\mu}(\cdot \mid \theta)$ for all $\theta \in \Theta$, whenever $q \in [0,1]$ follows the uniform distribution.

For any $q \in [0,1]$, define $\nu^{q,\mu} \in \Delta(\Omega \times \Theta)$ by

$$\nu^{q,\mu}(A \times B) := \int_B \mathbf{1}\{\tilde{\omega}^\mu(q,\theta) \in A\}\, \mathrm{d}\nu_\Theta\,.$$

Note that the functions $q \mapsto \nu^{q,\mu}$ and $\mu \mapsto \nu^{q,\mu}$ are measurable since $q \mapsto \tilde{\omega}^\mu(q,\theta)$ and $\mu \mapsto \tilde{\omega}^\mu(q,\theta)$ are measurable. Also note that the marginal of $\nu^{q,\mu}$ equals ν_Θ for all $q \in [0,1]$. Furthermore, for any measurable $A \subseteq \Omega$ and for any measurable $B \subseteq \Theta$,

$$\int_0^1 \nu^{q,\mu}(A \times B)\, \mathrm{d}q = \int_0^1 \int_B \mathbf{1}\{\tilde{\omega}^\mu(q,\theta) \in A\}\, \mathrm{d}\nu_\Theta\, \mathrm{d}q = \int_B \tilde{\mu}(A \mid \theta)\, \mathrm{d}\nu_\Theta = \mu(A \times B)\,.$$

Now define $\gamma^\star \in \Delta\Delta(\Omega \times \Theta)$ as

$$\gamma^\star(E) := \int_\Gamma \lambda(\{q \in [0,1] : \nu^{q,\mu} \in E\})\, \mathrm{d}\gamma(\mu)\,.$$

Theorem 2 of Strassen (1965) then implies that $\gamma^\star$ is a mean-preserving spread of γ. Furthermore, note that by construction, for $\gamma^\star$-almost all $\tilde{\nu}$, $\tilde{\nu} = \nu^{q,\mu}$ for some $q \in [0,1]$ and for some $\mu \in \Delta\Delta(\Omega \times \Theta)$. In particular, for $\gamma^\star$-almost all $\tilde{\nu}$, the marginal of $\tilde{\nu}$ over Θ is ν_Θ, and $\tilde{\nu}(\{(\omega,\theta) \in \Omega \times \Theta : w = \eta(\tilde{\nu},\theta)\}) = 1$, for some measurable function $\eta : \Delta\Delta(\Omega \times \Theta) \times \Theta \to \Omega$.

As $\gamma^\star$ is a mean-preserving spread of the Dirac measure on $\mathbb{P}$, from Blackwell's theorem, there exists a signal $\pi^\star$ that induces $\gamma^\star$ as its distribution of posteriors over $\Omega \times \Theta$. Since $\gamma^\star$ is

a mean-preserving spread of γ, $\pi^\star$ is Blackwell-more informative than π. Since the marginal over Θ of $\gamma^\star$-almost every posterior is ν_Θ, the signal $\pi^\star$ can be chosen so that every posterior induced by $\pi^\star$ has marginal ν_Θ over Θ, and hence, $\pi^\star$ is privacy-preserving. Likewise, since $\gamma^\star$-almost every posterior $\tilde{\nu}$ assigns probability 1 to the event $\{\omega = \eta(\tilde{\nu}, \theta)\}$, $\pi^\star$ can be chosen so that for any s, the posterior assigns probability 1 to the event $\{\omega = \tilde{\eta}(s, \theta)\}$, for some $\tilde{\eta} : S \times \Theta \to \Omega$, and hence is conditionally revealing. This completes the proof. □

Lemma A.2. *For any signal π, there exists a Blackwell-equivalent signal $([0,1], \hat{\pi})$ such that for any measurable $C \subseteq [0,1]$,*

$$\int_{\Omega\times\Theta} \pi_{(\omega,\theta)}(C)\,\mathrm{d}\mathbb{P} = \lambda(C)\,.$$

Proof. Let $\gamma \in \Delta\Delta(\Omega \times \Theta)$ be the distribution of posteriors over $\Omega \times \Theta$ associated with π. Define a joint distribution $\Pi \in \Delta(\Omega \times \Theta \times \Delta(\Omega \times \Theta))$ as

$$\Pi(A \times B \times C) := \int_C \mu(A \times B)\,\mathrm{d}\gamma(\mu)\,,$$

for all measurable $A \subseteq \Omega$, $B \subseteq \Theta$, and $C \subseteq \Delta(\Omega\times\Theta)$. Then, since $\Delta(\Omega\times\Theta)$ is standard-Borel, there exists a transition probability $\tilde{\pi} : \Omega \times \Theta \to \Delta(\Omega \times \Theta)$ such that

$$\Pi(A \times B \times C) = \int_{A\times B} \tilde{\pi}_{(\omega,\theta)}(C)\,\mathrm{d}\mathbb{P}\,,$$

for all measurable $A \subseteq \Omega$, $B \subseteq \Theta$, and $C \subseteq \Delta(\Omega \times \Theta)$. Then π is Blackwell equivalent to $(\Delta(\Omega \times \Theta), \tilde{\pi})$. Moreover, since $\Delta(\Omega \times \Theta)$ is standard-Borel, there exists a signal $([0,1], \hat{\pi})$ that is equivalent to $(\Delta(\Omega \times \Theta), \tilde{\pi})$ and such that

$$\int_{\Omega\times\Theta} \hat{\pi}_{(\omega,\theta)}(C)\,\mathrm{d}\mathbb{P} = \lambda(C)\,,$$

for all measurable $C \subseteq [0,1]$. This completes the proof. □

Proof of Lemma 5. Consider any conditionally revealing privacy-preserving signal π. By Lemma A.2, it is without loss to assume that $S = [0,1]$ and that the marginal distribution of s is uniform. Let $\eta : [0,1] \to \Theta$ be the measurable function such that $\omega = \eta(s, \theta)$ with $\mathbb{P}_\pi[\cdot \mid s]$ probability 1 for all $s \in S$. Note that this implies, for any measurable $A \subseteq \Omega$, and for

ν_Θ-almost all θ,

$$\lambda(\{s \in [0,1] : \eta(s,\theta) \in A\}) = \mathbb{P}[\omega \in A \mid \theta] = \lambda(\{s \in [0,1] : \omega^\star(s,\theta) \in A\}).$$

That is, for ν_Θ-almost all θ, the random variables $\eta(\cdot,\theta) : [0,1] \to \Omega$ and $\omega^\star(\cdot,\theta) : [0,1] \to \Omega$ have the same distribution. It then remains to show that, for any two random variables $\eta_1 : [0,1] \to \Omega$ and $\eta_2 : [0,1] \to \Omega$ that have the same distribution, there exists a measure-preserving transformation Φ such that $\eta_1(s) = \eta_2(\Phi(s))$ for all $s \in [0,1]$.

To this end, note that since $(\Omega, \mathcal{F})$ is a standard Borel space, there exists a Borel isomorphism $\phi : [0,1] \to \Omega$ such that both ϕ and $\phi^{-1} : \Omega \to [0,1]$ are measurable. Consider any two integrable functions $\eta_1, \eta_2 : [0,1] \to \Omega$ such that

$$\lambda(\{s \in [0,1] : \eta_1(s) \in A\}) = \lambda(\{s \in [0,1] : \eta_2(s) \in A\},$$

for all measurable $A \in \mathcal{F}$. Then, for any Borel measurable set $C \subseteq [0,1]$,

$$\lambda(\{s \in [0,1] : \phi \circ \eta_1(s) \in C\}) = \lambda(\{s \in [0,1] : \phi \circ \eta_2(s) \in C\}).$$

That is, $\phi \circ \eta_1(s)$ and $\phi \circ \eta_2(s)$ have the same distribution whenever $s \in [0,1]$ follows the uniform distribution. Since both $\phi \circ \eta_1$ and $\phi \circ \eta_2$ are in $L^1([0,1])$, by Proposition 3 of Ryff (1970), there exists a measure-preserving transformation $\Phi : [0,1] \to [0,1]$ such that $\phi \circ \eta_1(s) = \phi \circ \eta_2(\Phi(s))$ for all $s \in [0,1]$. Since ϕ is an isomorphism, it follows that $\eta_1(s) = \eta_2(\Phi(s))$ for all $s \in [0,1]$. □

References

Aguirre, Inaki, Simon Cowan, and John Vickers (2010) "Monopoly Price Discrimination and Demand Curvature," *American Economic Review*, 100 (4), 1601–1615.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner (2016) "Machine Bias," https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Arieli, Itai, Yakov Babichenko, Fedor Sandomirskiy, and Omer Tamuz (2021) "Feasible Joint Posterior Beliefs," *Journal of Political Economy*, 129 (9), 2546–2594.

Arieli, Itai, Yakov Babichenko, Rann Smorodinsky, and Takuro Yamashita (2023) "Optimal persuasion via bi-pooling," *Theoretical Economics*, 18 (1), 15–36.

Arnold, David, Will Dobbie, and Peter Hull (2022) "Measuring Racial Discrimination in Bail Decisions," *American Economic Review*, 112 (9), 2992–3038.

Aumann, Robert J. and Michael B. Maschler (1995) *Repeated Games with Incomplete Information*: MIT Press.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan (2019) *Fairness and Machine Learning: Limitations and Oppurtunities*: MIT Press.

Bergemann, Dirk, Benjamin Brooks, and Stephen Morris (2015) "The Limits of Price Discrimination," *American Economic Review*, 105 (3), 921–957.

Blackwell, David (1953) "Equivalent Comparisons of Experiments," *Annals of Mathematical Statistics*, 24 (2), 265–272.

Border, Kim C. (1991) "Implementation of Reduced Form Mechanisms: A Geometric Approach," *Econometrica*, 59 (4), 1175–1187.

Calders, Toon, Faisal Kamiran, and Mykola Pechenizkiy (2009) "Building Classifiers with Indenpendency Constraints," in *IEEE International Conference on Data Mining Workshops*, 13–18, 10.1109/ICDMW.2009.83.

Calders, Toon and Sicco Verwer (2010) "Three Naive Bayes Approaches for Discrimination-Free Classification," *Data Mining and Knowledge Discovery*, 21, 277–292.

Carey, Alycia N. and Xintao Wu (2023) "The Statistical Fairness Field Guide: Perspectives from Social and Formal Sciences," *AI and Ethics*, 3, 1–23.

Çinlar, Erhan (2010) *Probability and Stochastics*: Springer.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq (2017) "Algorithmic Decision Making and the Cost of Fairness," in *23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797–806, 10.1145/3097983.3098095.

Cowan, Simon (2016) "Welfare-Increasing Third-Degree Price Discrimination," *RAND Journal of Economics*, 47 (2), 326–340.

Darlington, Richard B. (1971) "Another Look at Cultural Fairness," *Journal of Educational Measurement*, 8 (2), 71–82.

Doval, Laura and Alex Smolin (2023) "Persuasion and Welfare,"Technical report, CEPR Discussion Papers.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel (2012) "Fairness through Awareness," *ACM ITCS Proceedings*, 214–226.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith (2006) "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, 265–284, Springer.

Feldman, Michael, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian (2015) "Certifying and Removing Disparate Impact," in *21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268, 10.1145/2783258.2783311.

Fuster, Andreas, Paul S. Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther (forthcoming) "Predictably Unequal? The Effects of Machine Learning on Credit Markets," *Journal of Finance*.

Gentzkow, Matthew and Emir Kamenica (2016) "A Rothschild-Stiglitz Approach to Bayesian Persuasion," *American Economic Review: Papers and Proceedings*, 106, 597–601.

Ghosh, Partha Pratim and Subir Kumar Bhandari (2017) "Characterizations of Extreme Copulas," arXiv preprint arXiv:1709.02472.

Gillis, Talia, Bryce McLaughlin, and Jann Spiess (2021) "On the Fairness of Machine-Assisted Human Decisions," Working Paper.

Haghpanah, Nima and Ron Siegel (2022) "The Limits of Multi-Product Price Discrimination," *American Economic Review: Insight*, 4 (4), 443–458.

——— (Forthcoming) "Pareto Improving Segmentation of Multi-Product Markets," *Journal of Political Economy*.

Hardt, Moritz, Eric Price, and Nathan Srebro (2016) "Equality of Opportunity in Supervised Learning," arXiv preprint arXiv:1610.02413.

Hart, Sergiu and Philip J. Reny (2015) "Implementation of Reduced Form Mechanisms: A Simple Approach and a New Characterization," *Economic Theory Bulletin*, 3, 1–8.

He, Kevin, Fedor Sandomirskiy, and Omer Tamuz (2023) "Private Private Information," arXiv preprint arXiv:2112.14356.

Kamenica, Emir and Matthew Gentzkow (2011) "Bayesian Persuasion," *American Economic Review*, 101 (6), 2560–2615.

Kamiran, Faisal, Indré Žilobaitė, and Toon Calders (2013) "Quantifying Explaniable Discriminaition and Removing Illegal Discrimination in Automated Decision Making," *Knowledge and Information Systems*, 35 (3), 613–644.

Kamishima, Toshihiro, Shotaro Akaho, and Jun Sakuma (2011) "Fairness-aware Learning through Regularization Approach," in *IEEE International Conference on Data Mining Workshops*, 643–650, 10.1109/ICDMW.2011.83.

Kitagawa, Toru, Shosei Sakaguchi, and Aleksey Tetenov (2021) "Constrained Classification and Policy Learning," Working Paper.

Kleiner, Andreas, Benny Moldovanu, and Philipp Strack (2021) "Extreme Points and Majorization: Economic Applications," *Econometrica*, 89 (4), 1557–1593.

Liang, Annie, Jay Lu, and Xiaosheng Mu (2023) "Algorithm Design: A Fairness-Accuracy Frontier," Working Paper.

Lorenz, George G. (1949) "A Problem of Plane Measure," *American Journal of Mathematics*, 71 (2), 417–426.

von Neumann, John (1932) "Einige sätze über messbare abbildungen," *Ann. of Math.(2)*, 33 (3), 574–586.

Perronea, Elisa and Fabrizio Durante (2021) "Extreme Points of Polytopes of Discrete Copulas," *Atlantis Studies in Uncertainty Modelling*, 3, 596–601.

Puccetti, Giovanni and Ruodu Wang (2015) "Extremal Dependence Concepts," *Statistical Science*, 30 (4), 485–571.

Rachev, Svetlozar T. and Ludger Rüschendorf (1998) *Mass Transportation Problems*: Springer.

Rokhlin, VA (1952) *On the Fundamental Ideas of Measure Theory:(Matematiceskiui Sbornik (ns) 25 (67) 107-150 (1949))* (71): American Mathematical Society.

Ryff, John V. (1970) "Measure Preserving Transformations and Rearrangements," *Journal of Mathematical Analysis and Applications*, 31, 449–458.

Santambrogio, Filippo (2015) *Optimal Transport for Applied Mathematics*: Springer.

Schmutte, Ian M. and Nathan Yoder (2022) "Information Design for Differential Privacy," Working Paper.

Shaked, Moshe and J. George Shanthikumar (2007) *Stochastic Orders*: Springer.

Strassen, Volker (1965) "The Existence of Probability Measures with Given Marginals," *Annuals of Mathematical Statistics*, 36, 423–439.

Tchen, Andre H. (1980) "Inequalities for Distributions with Given Marginals," *Annals of Probability*, 8 (4), 814–827.

The White House (2015) "Big Data and Differential Pricing."

Varian, Hal R. (1985) "Price Discrimination and Social Welfare," *American Economic Review*, 75 (4), 870–875.

Wiggins, Benjamin (2020) *Calculating Race: Racial Riscrimination in Risk Assessment*: Oxford University Press.

Yang, Crystal S. and Will Dobbie (2020) "Equal Protection under Algorithms: A New Statistical and Legal Framework," *Michigan Law Review*, 119 (2), 291–396.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi (2015) "Learning Fair Classifiers," *arXiv: Machine Learning*.

Online Appendix

Proof of Lemma 1. Consider any family $\Phi = \{\Phi_{\theta_j}\}_{j=1}^J$ of measure-preserving transformations. Since $F^{-1}(\cdot \mid \theta_j)$ is the generalized quantile function, the distribution of $F^{-1}(\Phi_{\theta_j}(s) \mid \theta_j)$ is $F(\cdot \mid \theta_j)$ for all $j \in \{1, \dots, J\}$. Therefore, the joint distribution ρ of $(F^{-1}(\Phi_{\theta_j}(s) \mid \theta_j))_{j=1}^J$ is in $\mathcal{M}$.

Conversely, consider any $\rho \in \mathcal{M}$. Since Ω^J is standard Borel, there exists measurable functions $\{\eta_j\}_{j=1}^J$ such that the joint distribution of $(\eta_j(s))_{j=1}^J$ is ρ, where s is the uniform random variable on $[0,1]$. Since for all $j \in \{1, \dots, J\}$, $\eta_j(s)$ and $F^{-1}(s \mid \theta_j)$ have the same distribution, there exists a measure-preserving transformation $\Phi_{\theta_j} : [0,1] \to [0,1]$ such that $\eta_j(s) = F^{-1}(\Phi_{\theta_j}(s) \mid \theta_j)$ for all $s \in [0,1]$ and for all $j \in \{1, \dots, J\}$ by Proposition 3 of Ryff (1970), as desired. □

Proof of Proposition 2. By Theorem 3 and Blackwell's theorem, any privacy-preserving signal yields a (weakly) lower payoff to the decision-maker than some reordered canonical signal. Together with Lemma 1, it then follows that

$$V^\star = \sup_{\rho \in \mathcal{M}} \int_{\Omega^J} V(\omega_1, \dots, \omega_J) \, \mathrm{d}\rho \, .$$

Moreover, by Theorem 3, any privacy-preserving that yields $V^\star$ must be a Φ-reordered canonical signal $\pi^\star_\Phi$, for some family $\Phi = \{\Phi_{\theta_j}\}_{j=1}^J$ of measure-preserving transformations. Thus, by Lemma 1, the joint distribution of $(\omega^\star(\Phi_{\theta_j}(s), \theta_j))_{j=1}^J$ must be a solution of (5). □

Proof of Proposition 3. Let $\widehat{V} : \Omega^J \times A \to \mathbb{R}$ be defined as

$$\widehat{V}(\omega_1, \dots, \omega_J, a) := \sum_{j=1}^J u(\omega_j, \theta_j, a) \mathbb{P}[\theta = \theta_j].$$

We first show that $\widehat{V}$ has increasing difference in $(\omega_1, \dots, \omega_J)$ and a, and is supermodular in $(\omega_1, \dots, \omega_J)$. Indeed, for any $a, a' \in A$ and $\boldsymbol{\omega} = (\omega_j)_{j=1}^J, \boldsymbol{\omega}' = (\omega'_j)_{j=1}^J \in \Omega^J$ such that $a \geq a'$ and

$\omega_j \geq \omega'_j$ for all j,

$$\begin{aligned}\widehat{V}(\boldsymbol{\omega}, a) - \widehat{V}(\boldsymbol{\omega}, a') =& \sum_{j=1}^{J} [u(\omega_j, \theta_j, a) - u(\omega'_j, \theta_j, a')]\mathbb{P}[\theta = \theta_j] \\ \geq& \sum_{j=1}^{J} [u(\omega'_j, \theta_j, a) - u(\omega'_j, \theta_j, a')]\mathbb{P}[\theta = \theta_j] \\ =& \widehat{V}(\boldsymbol{\omega}', a) - \widehat{V}(\boldsymbol{\omega}', a'),\end{aligned}$$

where the inequality follows from the increasing difference property of u. Furthermore, for any $\boldsymbol{\omega} = (\omega_j)_{j=1}^J, \boldsymbol{\omega}' = (\omega'_j)_{j=1}^J \in \Omega^J$ and for all $a \in A$,

$$\begin{aligned}\widehat{V}(\boldsymbol{\omega} \vee \boldsymbol{\omega}', a) + \widehat{V}(\boldsymbol{\omega} \wedge \boldsymbol{\omega}', a) =& \sum_{j=1}^{J} [u(\max\{\omega_j, \omega'_j\}, \theta_j, a) + u(\min\{\omega_j, \omega'_j\}, \theta_j, a)]\mathbb{P}[\theta = \theta_j] \\ =& \sum_{j=1}^{J} [u(\omega_j, \theta_j, a) + u(\omega'_j, \theta_j, a)]\mathbb{P}[\theta = \theta_j] = \widehat{V}(\boldsymbol{\omega}, a) + \widehat{V}(\boldsymbol{\omega}', a),\end{aligned}$$

We next show that $V : \Omega^J \to \mathbb{R}$ defined in (4) is supermodular. Since $\operatorname{argmax}_{a \in A} \widehat{V}(\boldsymbol{\omega}, a)$ is nonempty for all $\boldsymbol{\omega} \in \Omega^J$, for any $a^\star(\boldsymbol{\omega}) \in \operatorname{argmax}_{a \in A} \widehat{V}(\boldsymbol{\omega}, a)$ and for any $\boldsymbol{\omega} = (\omega_j)_{j=1}^J \in \Omega^J$, $V(\boldsymbol{\omega}) = \widehat{V}(\boldsymbol{\omega}, a^\star(\boldsymbol{\omega}))$. Therefore, it suffices to show that

$$\widehat{V}(\boldsymbol{\omega} \vee \boldsymbol{\omega}', a^\star(\boldsymbol{\omega} \vee \boldsymbol{\omega}')) + \widehat{V}(\boldsymbol{\omega} \wedge \boldsymbol{\omega}', a^\star(\boldsymbol{\omega} \wedge \boldsymbol{\omega}')) \geq \widehat{V}(\boldsymbol{\omega}, a^\star(\boldsymbol{\omega})) + \widehat{V}(\boldsymbol{\omega}', a^\star(\boldsymbol{\omega}')),$$

for all $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega^J$. To see this, consider any $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega^J$. Since A is totally ordered, it is

without loss to assume that $a^\star(\boldsymbol{\omega}) \geq a^\star(\boldsymbol{\omega}')$. As a result,

$$
\begin{aligned}
&\widehat{V}(\boldsymbol{\omega} \vee \boldsymbol{\omega}', a^\star(\boldsymbol{\omega} \vee \boldsymbol{\omega}')) + \widehat{V}(\boldsymbol{\omega} \wedge \boldsymbol{\omega}', a^\star(\boldsymbol{\omega} \wedge \boldsymbol{\omega}'))) \\
=&\widehat{V}(\boldsymbol{\omega} \vee \boldsymbol{\omega}', a^\star(\boldsymbol{\omega})) + \widehat{V}(\boldsymbol{\omega} \wedge \boldsymbol{\omega}', a^\star(\boldsymbol{\omega})) \\
&+ [\widehat{V}(\boldsymbol{\omega} \vee \boldsymbol{\omega}', a^\star(\boldsymbol{\omega} \vee \boldsymbol{\omega}')) - \widehat{V}(\boldsymbol{\omega} \vee \boldsymbol{\omega}', a^\star(\boldsymbol{\omega}))] + [\widehat{V}(\boldsymbol{\omega} \wedge \boldsymbol{\omega}', a^\star(\boldsymbol{\omega} \wedge \boldsymbol{\omega}')) - \widehat{V}(\boldsymbol{\omega} \wedge \boldsymbol{\omega}', a^\star(\boldsymbol{\omega}))] \\
\geq&\widehat{V}(\boldsymbol{\omega}, a^\star(\boldsymbol{\omega})) + \widehat{V}(\boldsymbol{\omega}', a^\star(\boldsymbol{\omega})) \\
&+ [\widehat{V}(\boldsymbol{\omega} \vee \boldsymbol{\omega}', a^\star(\boldsymbol{\omega} \vee \boldsymbol{\omega}')) - \widehat{V}(\boldsymbol{\omega} \vee \boldsymbol{\omega}', a^\star(\boldsymbol{\omega}))] + [\widehat{V}(\boldsymbol{\omega} \wedge \boldsymbol{\omega}', a^\star(\boldsymbol{\omega} \wedge \boldsymbol{\omega}')) - \widehat{V}(\boldsymbol{\omega} \wedge \boldsymbol{\omega}', a^\star(\boldsymbol{\omega}))] \\
=&\widehat{V}(\boldsymbol{\omega}, a^\star(\boldsymbol{\omega})) + \widehat{V}(\boldsymbol{\omega}', a^\star(\boldsymbol{\omega}')) + \widehat{V}(\boldsymbol{\omega}', a^\star(\boldsymbol{\omega})) - \widehat{V}(\boldsymbol{\omega}', a^\star(\boldsymbol{\omega}')) \\
&+ [\widehat{V}(\boldsymbol{\omega} \vee \boldsymbol{\omega}', a^\star(\boldsymbol{\omega} \vee \boldsymbol{\omega}')) - \widehat{V}(\boldsymbol{\omega} \vee \boldsymbol{\omega}', a^\star(\boldsymbol{\omega}))] + [\widehat{V}(\boldsymbol{\omega} \wedge \boldsymbol{\omega}', a^\star(\boldsymbol{\omega} \wedge \boldsymbol{\omega}')) - \widehat{V}(\boldsymbol{\omega} \wedge \boldsymbol{\omega}', a^\star(\boldsymbol{\omega}))] \\
\geq&\widehat{V}(\boldsymbol{\omega}, a^\star(\boldsymbol{\omega})) + \widehat{V}(\boldsymbol{\omega}', a^\star(\boldsymbol{\omega}')) + \widehat{V}(\boldsymbol{\omega} \wedge \boldsymbol{\omega}', a^\star(\boldsymbol{\omega})) - \widehat{V}(\boldsymbol{\omega} \wedge \boldsymbol{\omega}', a^\star(\boldsymbol{\omega}')) \\
&+ [\widehat{V}(\boldsymbol{\omega} \vee \boldsymbol{\omega}', a^\star(\boldsymbol{\omega} \vee \boldsymbol{\omega}')) - \widehat{V}(\boldsymbol{\omega} \vee \boldsymbol{\omega}', a^\star(\boldsymbol{\omega}))] + [\widehat{V}(\boldsymbol{\omega} \wedge \boldsymbol{\omega}', a^\star(\boldsymbol{\omega} \wedge \boldsymbol{\omega}')) - \widehat{V}(\boldsymbol{\omega} \wedge \boldsymbol{\omega}', a^\star(\boldsymbol{\omega}))] \\
=&\widehat{V}(\boldsymbol{\omega}, a^\star(\boldsymbol{\omega})) + \widehat{V}(\boldsymbol{\omega}', a^\star(\boldsymbol{\omega}')) \\
&+ [\widehat{V}(\boldsymbol{\omega} \vee \boldsymbol{\omega}', a^\star(\boldsymbol{\omega} \vee \boldsymbol{\omega}')) - \widehat{V}(\boldsymbol{\omega} \vee \boldsymbol{\omega}', a^\star(\boldsymbol{\omega}))] + [\widehat{V}(\boldsymbol{\omega} \wedge \boldsymbol{\omega}', a^\star(\boldsymbol{\omega} \wedge \boldsymbol{\omega}')) - \widehat{V}(\boldsymbol{\omega} \wedge \boldsymbol{\omega}', a^\star(\boldsymbol{\omega}'))] \\
\geq&\widehat{V}(\boldsymbol{\omega}, a^\star(\boldsymbol{\omega})) + \widehat{V}(\boldsymbol{\omega}', a^\star(\boldsymbol{\omega}'))\,,
\end{aligned}
$$

where the first inequality follows from supermodularity of $\widehat{V}$, the second inequality follows from the increasing difference property of $\widehat{V}$ and from $a^\star(\boldsymbol{\omega}) \geq a^\star(\boldsymbol{\omega}')$, and the third inequality follows from optimality of $a^\star$.

Finally, note that by Lemma 1, (5) is equivalent to choosing a family $\{\Phi_j\}_{j=1}^J$ of measure-preserving transformations to maximize

$$\int_0^1 V(F^{-1}(\Phi_1(q) \,|\, \theta_1), \ldots, F^{-1}(\Phi_J(q) \,|\, \theta_J))\, \mathrm{d}q.$$

Since V is supermodular, corollary 3 of Tchen (1980) (see also, Theorem 2.1 of Puccetti and Wang 2015) implies that

$$\int_0^1 V(F^{-1}(\Phi_1(q) \,|\, \theta_1), \ldots, F^{-1}(\Phi_J(q) \,|\, \theta_J))\, \mathrm{d}q \leq \int_0^1 V(F^{-1}(q \,|\, \theta_1), \ldots F^{-1}(q \,|\, \theta_J))\, \mathrm{d}q$$

for any family $\{\Phi_j\}_{j=1}^J$ of measure-preserving transformations. Together with Lemma 1, $V^\star$ is attained by the generalized quantile signal, as desired. □

Proof of Proposition 5. For any $\rho \in \mathcal{M}$, Lemma 1 implies that there exists a family $\Phi = \{\Phi_\theta\}_{\theta\in\Theta}$ of measure-preserving transformations such that the joint distribution of $(\omega^\star(\Phi_{\theta_j}(s), \theta_j))_{j=1}^J$ is ρ, where s follows the uniform distribution on $[0,1]$. Consider the problem where the sender is restricted to choose garblings of the Φ-reordered canonical signal. Standard arguments (Aumann and Maschler 1995; Kamenica and Gentzkow 2011) implies that the sender's value in this restricted problem is $\overline{V}_S(\rho)$. By Theorem 3, since every privacy-preserving signal is a garbling of some reordered canonical signal, the sender's value $V_S^\star$ in the original problem must be given by

$$\max_{\rho\in\mathcal{M}} \overline{V}_S(\rho)\,.$$

□

Transforming the Optimal Transport Problem In Section 5.2, we claim that the optional transport problem (5) with $V(\omega_1,\omega_2) = \min\{\omega_1,\omega_2\}$ is equivalent to the standard problem with an absolute transport cost. To see this, note that $-\min\{\omega_1,\omega_2\} = \max\{x_1,x_2\} - (x_1+x_2) = 0.5\max\{x_1-x_2, x_2-x_1\} - 0.5(x_1+x_2) = 0.5|x_1-x_2| - 0.5(x_1+x_2)$. Consequently, as the marginal distribution of ω_1,ω_2 is fixed we have that

$$\operatorname*{argmax}_\rho \int V(\omega_1,\omega_2)\,\mathrm{d}\rho = \operatorname*{argmin}_\rho \int |\omega_1-\omega_2|\,\mathrm{d}\rho\,.$$

Verifying the Optimality of ρ^* In Section 5.2, we claim that the joint distribution ρ^* is optimal in our example. To see this, recall that a joint distribution $\rho \in \mathcal{M}$ is a solution of the associated optimal transport problem if and only if there exists Lagrange multipliers $L, K : \Omega \to \mathbb{R}$ that satisfy the complementary slackness condition: $L(\omega_1) + K(\omega_2) \geq V(\omega_1,\omega_2)$, for all $(\omega_1,\omega_2) \in \Omega^2$, with equality on the support of ρ. It can then be verified that the complementary slackness condition is satisfied under the Lagrange multipliers $(L(\omega))_{\omega\in\Omega} = (1,2,5/2)$ and $(K(\omega))_{\omega\in\Omega} = (0,0,1/2)$, and hence $\rho^\star$ is indeed a solution.