DISCUSSION PAPER SERIES

DP18104

PERSUASION AND WELFARE

Laura Doval and Alex Smolin

INDUSTRIAL ORGANIZATION AND ORGANIZATIONAL ECONOMICS

ISSN 0265-8003

PERSUASION AND WELFARE

Laura Doval and Alex Smolin

Discussion Paper DP18104
Published 22 April 2023
Submitted 22 April 2023

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

PERSUASION AND WELFARE

Abstract

Information policies such as scores, ratings, and recommendations are increasingly shaping society's choices in high-stakes domains. We provide a framework to study the welfare implications of information policies on a population of heterogeneous agents. We define and characterize the Bayes welfare set, consisting of the population's utility profiles that are feasible under some information policy. The Pareto frontier of this set can be recovered by a series of standard Bayesian persuasion problems, in which a utilitarian planner takes the role of the information designer. We provide necessary and sufficient conditions under which an information policy exists that Pareto dominates the no-information policy. We extend our results to the case in which information policies are restricted in the data they can use and show that ``blinding" algorithms to sensitive inputs is welfare decreasing. We illustrate our results with applications to privacy, recommender systems, and credit ratings.

JEL Classification: N/A

Keywords: Bayesian persuasion

Laura Doval - laura.doval@columbia.edu
Columbia University Graduate School of Business and CEPR

Alex Smolin - alexey.v.smolin@gmail.com
Toulouse School of Economics and CEPR

Persuasion and Welfare*

Laura Doval† Alex Smolin‡

April 21, 2023

Abstract

Information policies such as scores, ratings, and recommendations are increasingly shaping society's choices in high-stakes domains. We provide a framework to study the welfare implications of information policies on a population of heterogeneous individuals. We define and characterize the *Bayes welfare set*, consisting of the population's utility profiles that are feasible under some information policy. The Pareto frontier of this set can be recovered by a series of *standard Bayesian persuasion* problems, in which a utilitarian planner takes the role of the information designer. We provide necessary and sufficient conditions under which an information policy exists that Pareto dominates the no-information policy. We extend our results to the case in which information policies are restricted in the data they can use and show that "blinding" algorithms to sensitive inputs is welfare decreasing. We illustrate our results with applications to privacy, recommender systems, and credit ratings.

Keywords: Bayesian persuasion, information design, welfare economics, algorithms, information policies.

*For valuable suggestions and comments, we would like to thank Odilon Câmara, Emir Kamenica, Antonio Penta, Jean Tirole as well as seminar participants at Toulouse School of Economics, Columbia, Bonn Winter Theory Workshop 2021, Warwick Theory Workshop 2022, Stony Brook 2022, ESSET 2022, and EEA-ESEM 2022. Smolin acknowledges funding from the French National Research Agency (ANR) under the Investments for the Future (Investissements d'Avenir) program (grant ANR-17-EURE-0010).

†Columbia Business School and CEPR. E-mail: laura.doval@columbia.edu.

‡Toulouse School of Economics and CEPR. E-mail: alexey.v.smolin@gmail.com.

1 Introduction

Information has increasingly become a tool for shaping society's choices in high-stakes domains. Consider, for instance, the role of algorithms in making recommendations for bail (Angwin et al., 2016), recruiting (Raghavan et al., 2020; Li et al., 2020), health (Obermeyer et al., 2019), education (Kučak et al., 2018), and lending (Jagtiani and Lemieux, 2019), among others. The role of information as a policy instrument is not confined to the *big data* economy. Indeed, information policies in the form of scores and ratings have been in place since long before algorithmic recommendations to determine school placement, promotions, and who receives credit. As society becomes reliant on information to guide decisions in policy-relevant domains, understanding the welfare implications of information-based policies becomes a first-order concern.

In this paper, we provide a framework to study the welfare impact of information policies in a population of heterogeneous agents. Formally, we study the following model. There is a unit mass population with characteristics in a finite type space, distributed according to a given prior distribution. We model an information policy as an information structure, which associates to each type a distribution over signals and hence, via Bayes' rule, a distribution over posterior beliefs (Kamenica and Gentzkow, 2011). Our primitive is an ex post welfare function that represents for each (posterior) type distribution the welfare of individuals of a given type. Given the welfare function, each information structure induces a *Bayes welfare profile*, which describes for each type in the population their expected payoff under the information structure. We define the Bayes welfare set to be the set of all such profiles.

We characterize the Bayes welfare set and study its properties. The Bayes welfare set allows us to reduce society's choice of an information policy to the choice of a Bayes welfare profile. Instead of imposing properties that the information policy must satisfy, society's preferences over the welfare distribution in the population determine the properties of the chosen information policy. Our perspective thus complements that of the algorithmic fairness literature which remains agnostic about the population's payoffs, focusing instead on statistical properties of information policies, such as accuracy, parity, or fairness. It also complements that of the literature in Bayesian persuasion (Rayo and Segal, 2010; Kamenica and Gentzkow, 2011), which characterizes the (maximum) average welfare consistent with some information structure, but not necessarily the Bayes welfare profiles that give rise to the average welfare.

Theorem 1 characterizes the Bayes welfare set via the convex-hull of a vector-valued function. In doing so, we extend the geometric characterizations of Aumann and Maschler (1995) and Kamenica and Gentzkow (2011) of the feasible set of ex ante payoffs to the characterization of the Bayes welfare set. Whereas a Bayes welfare

profile depends on the distribution over posteriors conditional on each type, we show it can be alternatively expressed as the unconditional expectation over posteriors of a *truth-adjusted* payoff function, where the adjustment is proportional to the posterior likelihood ratio of each type. Evaluated at a given type, the truth-adjusted welfare function allows us to characterize the welfare individuals with a given type may obtain under some information structure. In turn, interpreting the truth-adjusted payoff function as a vector-valued function allows us to capture the across-type restrictions imposed by Bayes' rule and precisely characterize the Bayes welfare set.

Theorem 2 characterizes the Pareto frontier of the Bayes welfare set. Points in the Pareto frontier are natural candidates for being the outcome of efficient bargaining over information structures or a social planner's choice. Theorem 1 implies the Bayes welfare set is convex so that it can be alternatively characterized via its supporting hyperplanes. Theorem 2 shows the points in the Pareto frontier of the Bayes welfare set can be recovered by a series of *standard* Bayesian persuasion problems, in which a utilitarian planner takes the role of an information designer. We use Theorem 2 throughout the paper to characterize optimal information structures in specific applications. Leveraging Theorem 2, Proposition 2 provides a necessary and sufficient condition under which an information structure exists that Pareto dominates the no-information policy.

Theorem 3 enriches our characterization in the case in which the welfare function is equal to the expectation of a one-dimensional random variable, the support of which we call the *reputation vector*. This special case constitutes a natural benchmark and is commonly used in the literature on career concerns (Holmström, 1999), social image (Bénabou and Tirole, 2006, Tirole, 2021), and policy prediction problems (Mullainathan, 2018). Theorem 3 shows a welfare profile belongs to the Bayes welfare set if and only if it can be represented as the product between the reputation vector and a *completely positive matrix* that satisfies a version of Bayes plausibility.[1] In addition, we show that the Bayesian persuasion problems that characterize the *relative* boundary of the Bayes welfare set correspond to instances of the problem in Rayo and Segal (2010). It follows that the information structures that induce payoffs in the relative boundary of the Bayes welfare set can be characterized using the graph-theoretic approach in Rayo and Segal (2010). We leverage their approach in Proposition 4, where we show that the information structures that maximize the welfare of a given type in the population correspond to a noisy version of the priority mechanisms studied in the matching literature (e.g., Celebi and Flynn, 2022).

Whereas the analysis so far presumes that *all* information policies are allowed, Section 6 extends our framework to accommodate limits on how much information can

[1] A matrix $\mathrm{C} \in \mathbb{R}^{N \times N}$ is completely positive if non-negative vectors $\mathrm{c}_1, \ldots, \mathrm{c}_K \in \mathbb{R}^N_+$ exist such that $\mathrm{C} = \sum_{i=1}^{K} \mathrm{c}_i \mathrm{c}_i^T$ (Berman, 1988).

be disclosed about the individuals in the population. We achieve this by distinguishing between the individuals' types (the variable on which we condition payoffs) and the data source (the variable we provide information about). Theorem 4 shows the characterization in Theorem 1 extends verbatim to this setting. Furthermore, Proposition 5 establishes the analogue of Blackwell's theorem in our setting (Blackwell, 1953). Consequently, as the data source becomes more informative, the associated Bayes welfare set increases in the set inclusion order. In other words, "blinding" the information policy to the individuals' payoff-relevant characteristics may be welfare decreasing.

Finally, we note that by interpreting our welfare function as an individual's type-dependent payoff function, the Bayes welfare set is also the object of interest in more standard information design applications. For instance, the types may represent the private information of an informed principal who can commit to an information structure only *after* observing her type, as in Perez-Richet (2014) and Koessler and Skreta (Forthcoming). Similarly, in the study of mechanism design with limited commitment, Doval and Skreta (2022) describe the principal's mechanism as an information structure that must satisfy an informed agent's incentive constraints. Similar constraints appear in the studies of information design without commitment, as in Fréchette et al. (2022), Lipnowski and Ravid (2020), Salamanca (2021), and in the analysis of tests subject to participation constraints in Rosar (2017). Thus, the Bayes welfare set can be viewed as a unifying concept that underlies the incentive constraints the equilibrium information structure must satisfy. As we show in our first working paper version, Doval and Smolin (2021), our tools also open the door to the study of new problems in this literature.

Related Literature: Our work contributes to the literature on information design reviewed in the introduction. Starting from the work of Kamenica and Gentzkow (2011) and Rayo and Segal (2010), a series of papers investigate the limits imposed by common knowledge of Bayesian rationality (Aumann, 1987). Whereas the Bayesian persuasion literature studies the (maximum) *average* welfare that can be achieved under some information policy, we characterize instead the welfare profiles that are consistent with some information structure.

Whereas ours is the first characterization of the Bayes welfare set, a small literature studies certain Bayes welfare profiles within applications. Assuming the information designer is an informed principal, Perez-Richet (2014) refers to the payoff profile induced by an information structure as an *interim* payoff and studies the informed principal's preferred Bayes welfare profile. Recently, Galperti et al. (Forthcoming) study the Bayes welfare profile that gives rise to the sender's maximum average payoff and relate it to the Lagrange multiplier in the Bayes' plausibility constraint. Restricting attention to the case in which the welfare function is linear in beliefs, Saeedi and

Shourideh (2020) characterize a subset of the Bayes welfare set that satisfies certain incentive compatibility constraints, thus obtaining a different characterization. Finally, as the analysis below makes clear, a Bayes welfare profile depends on the posterior distribution induced by an information structure *conditional* on each type. Whereas Levy et al. (2021) and Arieli et al. (2022) characterize the set of conditional distributions over posteriors consistent with the prior, we follow a complementary approach that allows us to carry only the (unconditional) posterior distribution induced by an information structure (see Claim 1).

We also contribute to the economics literature that studies algorithmic fairness. Mullainathan (2018), Kleinberg et al. (2018), and Rambachan et al. (2020) argue for letting the social planner's objective determine the properties of algorithms. The analysis in Section 6 relates to Liang et al. (2022), who study decision-making algorithms. Starting from a fixed joint distribution over groups, covariates, and states, Liang et al. (2022) model an algorithm as taking actions directly as a function of covariates and study the group error profiles in the fairness-accuracy frontier as the algorithm varies. Instead, we model an algorithm as an information structure that sends non-binding action recommendations to an unmodeled receiver, whose actions determine the population's welfare and characterize the set of all Bayes welfare profiles as we vary the algorithm. We expand on the connection between the two papers in Section 6 (see Remark 4).

Finally, our work contributes indirectly to the literature on higher-order beliefs. Indeed, when the welfare function is linear in beliefs as in Section 5, the welfare profile can be seen as a profile of second-order expectations. Starting with Samet (1998), a body of work uses Markov matrices to represent such higher-order beliefs and expectations of higher-order beliefs for a *given* information structure (see, e.g., Cripps et al., 2008; Golub and Morris, 2017). Instead, our result in Theorem 3 identifies the set of matrices that correspond to *some* information structure.

In lieu of an organizational paragraph, we summarize below the notation used throughout the paper:

Notation: For ease of presentation, we sometimes find it convenient to denote a function from a set Θ to $\mathbb{R}$ as a vector in $\mathbb{R}^N$, where N is the cardinality of Θ. In this case, we reserve the italic notation x for the function $x : \Theta \mapsto \mathbb{R}$ and the upright notation x for the vector in $\mathbb{R}^N$. Any vector $\mathrm{x} \in \mathbb{R}^N$ is taken to be a column vector; we denote its i^{th} component by x_i or $x(\theta_i)$ interchangeably. If $\mathrm{x} \in \mathbb{R}^N$ is a column vector, x^T denotes its transpose. If x, y are two vectors, $\mathrm{x} * \mathrm{y}$ denotes their Hadamard (element-wise) product and x/y denotes their Hadamard division. We denote by $\mathrm{e} \in \mathbb{R}^N$ the vector with $\mathrm{e}_1 = \cdots = \mathrm{e}_N = 1$. When we want to emphasize that x is a random variable, we write it as $\tilde{x}$.

2 Model

A unit mass population has characteristics in a finite type space, $\theta \in \Theta \equiv \{\theta_1, \ldots, \theta_N\}$. Letting $\Delta(\Theta)$ denote the set of probability distributions over Θ, we denote by $\mu_0 \in \Delta(\Theta)$ the frequency of characteristics in the population. We denote by $\Delta(\Delta(\Theta))$ the set of posterior distributions, and by $\Delta_{\mu_0}(\Delta(\Theta))$ the set of posterior distributions with mean equal to the prior μ_0.

Welfare function: An individual's welfare depends on her type θ and an (unmodeled) outside observer's belief about her type. We represent this by an ex post welfare function $w : \Delta(\Theta) \times \Theta \mapsto \mathbb{R}$ that represents for each belief μ and each type θ, the welfare of individuals of type θ under belief μ, $w(\mu, \theta)$. We assume throughout that w is bounded.

A welfare profile is a vector $\mathrm{w} \in \mathbb{R}^N$, where w_i describes the welfare level of individuals with type θ_i. Any welfare profile w induces an ex ante welfare of $\mu_0^T \mathrm{w} = \sum_{\theta \in \Theta} \mu_0(\theta_i)\mathrm{w}_i$. When we average a profile w using weights different than the prior μ_0, we refer to *average* welfare instead. We are interested in characterizing those welfare profiles that are induced by some information policy, to which we turn next.

Information policies: We model information policies as information structures. Formally, an information structure $\Pi = (\pi, S)$ consists of a countable set of labels S, and a mapping π, which associates to each type θ a distribution over signals $\pi(\cdot|\theta) \in \Delta(S)$. Given an information structure Π and a signal realization $s \in S$, the corresponding posterior belief $\mu_s \in \Delta(\Theta)$ obtained by Bayes' rule is defined by

$$\mu_s(\theta) = \frac{\mu_0(\theta)\pi(s|\theta)}{\sum_{\theta' \in \Theta} \mu_0(\theta')\pi(s|\theta')}.$$

Thus, an information structure can be seen as inducing a distribution over posterior beliefs $\{\mu_s : s \in S\}$. In what follows, two such distributions are of interest: the distribution over posterior beliefs *conditional* on an individual's type – as induced by $\pi(\cdot|\theta)$– and the *unconditional* distribution over posterior beliefs – as induced by the prior μ_0 and the signal distribution. When taking expectations using these distributions, we use the notations $\Pi|\theta$ and Π to denote the conditional and unconditional distributions, respectively.

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

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

That is, for each type θ, $w_\Pi(\theta)$ describes the (expected) welfare of type-θ individuals under information structure Π. Note that in computing the welfare of type-θ individuals, their type θ enters twice: directly through the welfare function, $w(\cdot, \theta)$, and indirectly through the signal distribution, $\pi(\cdot|\theta)$.

We now present our two main objects of study:

Definition 1 (Bayes welfare profile). *A welfare profile* $\mathrm{w} \in \mathbb{R}^N$ *is a* Bayes welfare profile *if an information structure,* Π*, exists such that for all types* θ*,* $\mathrm{w}(\theta) = w_\Pi(\theta)$*.*

Definition 2 (Bayes welfare set). *The Bayes welfare set is the set of all Bayes welfare profiles; that is,*

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

The Bayes welfare set W represents the utility possibility set in an economy where the allocations are given by information structures. As such, it describes the welfare effects that different information structures have for individuals with different characteristics in applications such as grading schemes in the case of schooling (Ostrovsky and Schwarz, 2010), disclosure about job performance (Mukherjee, 2008), affirmative action in the case of college admissions or the job market, rating systems in the case of platforms (Saeedi and Shourideh, 2020), and market segmentations (Bergemann et al., 2015).

Throughout, we illustrate our results using the following examples:

Example 1 (Privacy). Consumers concerned about their privacy wish to maximize an outside observer's uncertainty about their types. We formalize this as follows. There are two types of consumer, $\Theta = \{\theta_A, \theta_B\}$. Letting μ_0 denote the frequency of consumers of type θ_B (i.e., $\mu_0 \equiv \mu_0(\theta_B)$), we assume that $\mu_0 = 0.3$. That is, consumers with θ_B constitute the minority of the population. When the outside observer believes the consumer's type is θ_B with probability $\mu \in [0, 1]$, a consumer of any type experiences a welfare loss of $w(\mu) = -(\mu - 1/2)^2$. The consumer's welfare loss is minimal when the market is maximally confused about the consumer's type (i.e., $\mu = 1/2$). Instead, the consumer's welfare loss is maximal when the market has precise information about her type (i.e., $\mu \in \{0, 1\}$). When choosing what information to disclose about the consumer, a regulator is in fact choosing a welfare profile in W. $\lozenge$

Example 2 (Online marketplace). An online marketplace makes algorithmic recommendations to a buyer about whether to purchase a product from a seller. There are two seller types, low (θ_1) and high (θ_2), and they are equally likely, so that $\mu_0(\theta_1) = \mu_0(\theta_2) = 1/2$. Consumers prefer to buy from high-quality sellers. Thus, a seller's profit depends on the likelihood μ that the consumer attaches to the seller being of

high quality. In particular, we assume a seller's profits as a function of the consumers' beliefs are as follows:

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

When designing its recommendation algorithm, the online marketplace is in fact choosing a profit profile in W. ◊

Example 3 (Credit ratings). A credit agency makes lending decisions based on a consumer's perceived repayment probability. A consumer of type θ repays loans with probability $\rho(\theta)$. The credit agency approves loans with probability proportional to the expected value of ρ. A regulator wishes to maximize the probability that consumers of type θ_i receive a loan by choosing the information on which the credit agency can condition its approval decision. When choosing its information policy, the regulator is choosing a profile of expected loan-approval probabilities in W. ◊

We conclude this section by commenting on the model's assumptions (Remark 1) and the connection to Bayesian persuasion (Remark 2). Both can be skipped without loss of continuity:

Remark 1 (Implicit assumptions). *In the spirit of the mechanism design and information design literatures, our model allows for arbitrary information structures and an abstract welfare function. Our formulation implicitly assumes the following. First, until Section 6, we assume* any *information can be provided about an individual's payoff-relevant characteristics. Second, because the welfare function depends only on the first-order beliefs about an individual's type, our model only accounts for strategic interactions that follow the realization of a public signal (see, e.g., Laclau and Renou, 2017).*

Remark 2 (Connection to Kamenica and Gentzkow, 2011). *Our model encompasses the Bayesian persuasion model of Kamenica and Gentzkow (2011). In their model, a sender, who knows the state of the world $\theta\in\Theta$, provides information to a receiver, who then chooses an action $a\in A$. Denote by $v(a,\theta)$ the sender's payoffs when the state is θ and the receiver takes action a. If the receiver takes action $a(\mu)$ when her posterior belief is μ, the sender's ex post payoff is given by $v(a(\mu),\theta)\equiv w(\mu,\theta)$.*[2]

[2]Two remarks are in order. First, our ex post welfare function, w, differs from the sender's *indirect* utility function in Kamenica and Gentzkow (2011), usually denoted by $\hat{v}$. The indirect utility function is the expectation under μ of the ex post welfare function, $w(\mu,\cdot)$. That is, $\hat{v}(\mu)=\sum_{\theta\in\Theta}\mu(\theta)v(a(\mu),\theta)=\sum_{\theta\in\Theta}\mu(\theta)w(\mu,\theta)$. Second, our ex post welfare function presumes a *fixed* selection out of the receiver's best-response correspondence. The working paper Doval and Smolin (2021) characterizes the Bayes welfare set over all information structures *and* selection rules, showing the characterization in the current paper extends verbatim.

Under this interpretation, the set W *represents the profile of* interim *payoffs the sender can achieve for a given information structure. The set* W *is the relevant object of study in problems where either the sender does not have commitment, as in Lipnowski and Ravid (2020), or the sender can commit to the information structure but only chooses the information structure after observing the realization of the state* θ*, as in Perez-Richet (2014) and Koessler and Skreta (Forthcoming), or moral hazard is present, as in Saeedi and Shourideh (2020). In each of these cases, equilibrium considerations imply incentive constraints that may be written in terms of the sender's profile of interim payoffs.*

3 Characterization

Section 3 presents our first characterization of the Bayes welfare set via the convex hull of the graph of a vector-valued function, in the spirit of the belief-based approach of Kamenica and Gentzkow (2011).

Truth-drifting: An apparent obstacle in following the belief approach in Kamenica and Gentzkow (2011) is that the elements of W are expressed in terms of expectations *conditional* on a given type $\theta \in \Theta$, rather than unconditional expectations. Indeed, as shown in Francetich and Kreps (2014), conditional expectations do not satisfy the martingale property; rather, they *drift toward the truth*. More precisely, for any type θ and for any information structure, the expectation of the posterior probability of θ conditional on $\tilde{\theta} = \theta$ is higher than the prior probability of θ. That is,[3]

$$\mathbb{E}_{\Pi|\theta}\left[\frac{\tilde{\mu}(\theta)}{\mu_0(\theta)}\right] = \sum_{s\in S} \pi(s|\theta)\frac{\mu_s(\theta)}{\mu_0(\theta)} \geq 1. \tag{TD}$$

Instead of pursuing a characterization of *conditional* distributions of posteriors, we recover the belief approach in Kamenica and Gentzkow (2011) by studying a suitably modified welfare function.

Truth-adjusted welfare: We show any element $\mathrm{w} \in \mathrm{W}$ can be expressed as the unconditional expectation of an adjusted version of the welfare function. Indeed, define the *truth-adjusted* welfare function $\hat{w} : \Delta(\Theta) \times \Theta \mapsto \mathbb{R}$ to be

$$\hat{w}(\mu, \theta) \equiv \frac{\mu(\theta)}{\mu_0(\theta)} w(\mu, \theta). \tag{AW}$$

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

[3] Claim 2 provides a more general version of this result based on Theorem 3.

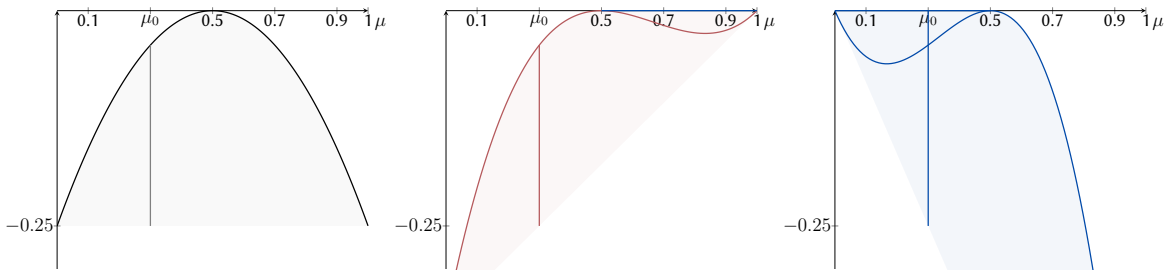

Figure (a) Welfare (w)

Figure (b) Adjusted welfare ($\hat{w}(\mu,\theta_A)$)

Figure (c) Adjusted welfare ($\hat{w}(\mu,\theta_B)$)

Figure 1: Truth-adjusted welfare function in Example 1

The truth-adjusted welfare function combines the preferences of individuals of type θ for a particular belief – $w(\mu,\theta)$ – and the *resource constraint* – Bayes' plausibility – in our economy, where information structures take the role of allocations. Indeed, the likelihood-ratio adjustment $\mu(\theta)/\mu_0(\theta)$ captures the constraint that comes from Bayes' plausibility. Intuitively, type-θ individuals have an endowment equal to $\mu_0(\theta)$ that can be spread over different beliefs $\mu(\theta)$, and the Bayes' plausibility constraint ensures this spread is done in a way that respects the budget.[4]

Example 1 (continued). We illustrate the truth-adjusted welfare function in the context of Example 1. Figure 1 depicts the ex post welfare function w (Figure 1a) and the truth-adjusted welfare function for θ_A (Figure 1b) and θ_B (Figure 1c). Whereas ex post welfare is type-independent, the truth-adjusted welfare function is type-dependent. This natural consequence of the likelihood-ratio adjustment reflects that consumers of different types benefit differently from various induced beliefs and therefore, from the same information structure. ♢

Claim 1 justifies our interest in the truth-adjusted welfare function $\hat{w}$:[5]

Claim 1 (From conditional to unconditional expectations). *For any information structure* Π *and any type* $\theta \in \Theta$*, the following holds:*

$$w_\Pi(\theta) = \mathbb{E}_\Pi\left[\hat{w}(\tilde{\mu},\theta)\right]. \tag{4}$$

The proof of Claim 1 and of other results can be found in the appendix.

[4] Formally, for any information structure, Π, and type $\theta \in \Theta$, $\mathbb{E}_\Pi[\tilde{\mu}(\theta)/\mu_0(\theta)] = 1$.

[5] In solving their respective design problems, Rosar (2017); Quigley and Walther (2019); Doval and Skreta (2022) similarly observe that the distribution over posterior beliefs conditional on an individual's type can be written in terms of the modified unconditional distribution. However, these papers do not provide the characterization result contained in Theorem 1.

Claim 1 implies the expectation of w under Π *conditional* on type θ can be expressed as the *unconditional* expectation of $\hat{w}$ under Π. Because the definition of $\hat{w}$ does not depend on Π, the distribution over posteriors induced by Π is enough to determine $\mathbb{E}_{\Pi}[\hat{w}(\cdot, \theta)]$. It follows that the expected value of $\hat{w}(\cdot, \theta)$ under posterior distributions with mean equal to μ_0 characterize the range of welfare values that individuals of type θ may obtain (Aumann and Maschler, 1995; Kamenica and Gentzkow, 2011). Indeed, let $w_*(\theta), w^*(\theta)$ denote the minimum and maximum welfare individuals of type θ can obtain under some information structure. That is, $w_*(\theta) = \inf\{w(\theta) : \mathrm{w} \in \mathrm{W}\}$ and $w^*(\theta) = \sup\{w(\theta) : \mathrm{w} \in \mathrm{W}\}$. We have the following:

Proposition 1 (Individually feasible welfare bounds). *For any type θ, the following holds*[6]*:*

$$w_*(\theta) = \operatorname{vex} \hat{w}(\mu_0, \theta),\ w^*(\theta) = \operatorname{cav} \hat{w}(\mu_0, \theta).$$

Proposition 1 follows from the main result in Kamenica and Gentzkow (2011). The vertical solid lines in Figures 1b and 1c illustrate the individually feasible welfare values for θ_A and θ_B in Example 1. An implication of Proposition 1 is that any Bayes welfare profile w satisfies that for all types θ, $\mathrm{w}(\theta) \in [\operatorname{vex} \hat{w}(\mu_0, \theta), \operatorname{cav} \hat{w}(\mu_0, \theta)]$.

Whereas Proposition 1 characterizes what is *individually* feasible for each type in the population, it does not deliver the characterization of the Bayes welfare set. The reason is that it ignores the across-type restrictions imposed by Bayes' rule. For instance, the truth-adjusted welfare function (AW) highlights that only types on the support of belief μ get to enjoy the payoff of inducing said belief. Similarly, inspection of Figures 1b and 1c shows θ_A's preferred information structure is no disclosure, whereas θ_B would prefer *some* disclosure to no disclosure. In other words, the profile $(w^*(\theta_A), w^*(\theta_B))$ is not *jointly* feasible.

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

Theorem 1 (Belief-based characterization). *The Bayes welfare set W satisfies the following:*

$$\mathrm{W} = \left\{\mathrm{w} \in \mathbb{R}^N : (\mu_0, \mathrm{w}) \in \operatorname{co}(\operatorname{graph} \hat{\mathrm{w}})\right\}. \tag{5}$$

Theorem 1 provides a geometric characterization of the set W: it is the section at the prior of the convex hull of the graph of the truth-adjusted welfare function $\hat{\mathrm{w}}$. Relying

[6] Recall that for a real-valued function f, $\operatorname{cav} f$ denotes the smallest concave function that dominates f, whereas $\operatorname{vex} f$ denotes the highest convex function that is dominated by f (see Hiriart-Urruty and Lemaréchal, 2004).

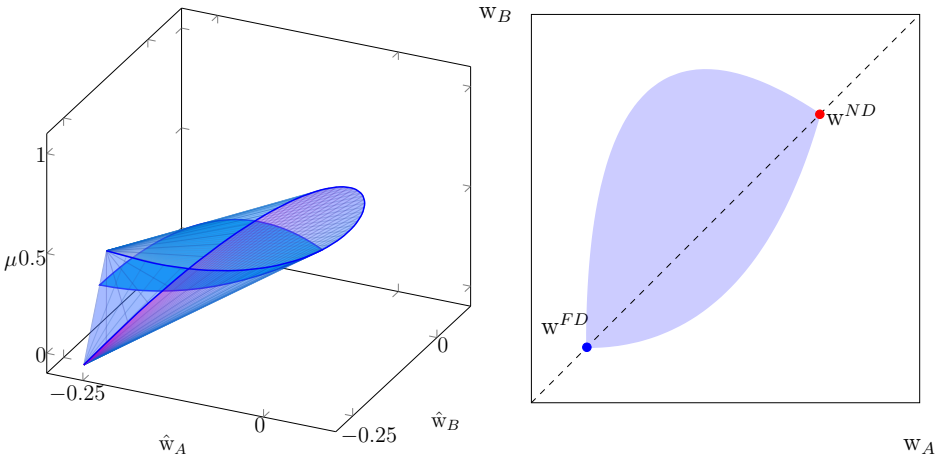

Figure (a) The convex hull of the graph of $\hat{w}$ Figure (b) The Bayes welfare set

Figure 2: Constructing the Bayes welfare set in Example 1

on the result in Kamenica and Gentzkow (2011) that any Bayes' plausible distribution over posteriors is the outcome of some information structure,[7] Theorem 1 characterizes a more primitive object, the set of welfare profiles that can be generated by some information structure. Indeed, whereas the main result in Kamenica and Gentzkow (2011) would allow us to characterize the (maximal) ex ante welfare a population with ex post welfare function w can obtain, Theorem 1 characterizes the welfare profiles whose average leads to that welfare.

Calculating the Bayes welfare set: Relying on Theorem 1, Figure 2 illustrates the construction of the Bayes welfare set in Example 1. Figure 2a depicts the convex hull of the graph of $\hat{\mathrm{w}}$. Applying Theorem 1, the resulting Bayes welfare set is the section of this convex hull at $\mu_0 = 0.3$. The blue color in Figure 2a depicts this section, which is represented in Figure 2b.

Figure 2b illustrates which welfare profiles are jointly feasible under *some* information structure in Example 1. For instance, fully revealing or concealing consumers' types is always possible, so that the full and no-disclosure profiles, w^{FD} and w^{ND}, are feasible. All Bayes welfare profiles Pareto dominate the full-disclosure profile, whereas no Bayes welfare profile dominates the no-disclosure one. As discussed above, simultaneously giving all consumer types their maximum welfare $w^*(\theta)$ is not possible, because θ_B's welfare is maximized by a policy that sometimes reveals a consumer is of type θ_A.

[7] See also Aumann and Maschler (1995) and Rayo and Segal (2010).

Cardinality: Theorem 1 has an immediate implication for the cardinality of the information structures that generate points in W:

Corollary 1. *Let* $\mathrm{w} \in \mathrm{W}$*. Then, an information structure* Π *with at most* $2N$ *signals exists such that* $\mathrm{w}_i = w_\Pi(\theta_i)$ *for all* $i \in \{1, \ldots, N\}$*.*

Corollary 1 stands in contrast to the result in Bayesian persuasion that finding an information structure that delivers the sender's maximal payoff and employs at most N posteriors is always possible. There are two reasons behind this difference. First, Theorem 1 characterizes *all* welfare profiles consistent with some information structure so that without further assumptions on the truth-adjusted welfare function, we rely on Carathédory's theorem to obtain Corollary 1. Instead, Kamenica and Gentzkow (2011) characterize the sender's maximal payoff and assume the indirect utility function is upper-semicontinuous. For that reason, Kamenica and Gentzkow (2011) can rely on Fénchel-Bunt's theorem instead of Carathéodory to obtain an upper bound of N instead of $N + 1$. Second, we are interested not just in the payoff of one "sender," but in the payoff of N, one for each type.

In Example 1, the upper bound in Corollary 1 is loose. Because the truth-adjusted welfare function is continuous, we can rely on Fénchel Bunt's theorem to reduce the upper bound in Corollary 1 by 1. Furthermore, the Bayes welfare profiles in the boundary of W are induced by information structures with at most two signals. We explain the reason for this further reduction in Section 4, where we characterize the boundary of the Bayes welfare set and, in particular, its Pareto frontier.

4 The Pareto frontier of the Bayes welfare set

We characterize in this section the Pareto frontier of W. Points in the Pareto frontier are natural candidates for being the outcome of efficient bargaining over information structures or a social planner's choice. Indeed, these points correspond to the solution of a utilitarian planner as we vary the weights the planner assigns to different types. Theorem 2 shows these points can be recovered as solutions to *standard* Bayesian persuasion problems, where an information designer takes the role of the utilitarian planner. Armed with this characterization, Proposition 2 provides a necessary and sufficient condition for the no-disclosure profile to be part of the Pareto frontier. In other words, it provides a necessary and sufficient condition for information disclosure to lead to a Pareto improvement relative to no disclosure.

We define the Pareto frontier of W to be the set of weak Pareto efficient Bayes welfare profiles. Formally,

$$\mathrm{W}_P = \{\mathrm{w} \in \mathrm{W} : (\nexists \mathrm{w}' \in \mathrm{W})\mathrm{w}' > \mathrm{w}\}. \tag{P}$$

Theorem 1 implies that for any $\mathrm{w} \in \mathrm{W}_P$, a direction $\lambda \in \mathbb{R}^N_+ \setminus \{0\}$ exists such that

$$\begin{aligned}\lambda^T \mathrm{w} = \max\left\{\lambda^T \mathrm{w}' : \mathrm{w}' \in \mathrm{W}\right\} &= \max\left\{\lambda^T \mathbb{E}_\tau\left[\hat{\mathrm{w}}(\tilde{\mu})\right] : \tau \in \Delta_{\mu_0}\left(\Delta(\Theta)\right)\right\} \\ &= \max\left\{\mathbb{E}_\tau\left[\lambda^T \hat{\mathrm{w}}(\tilde{\mu})\right] : \tau \in \Delta_{\mu_0}\left(\Delta(\Theta)\right)\right\}. \qquad (6)\end{aligned}$$

To obtain Equation 6, we invoke Theorem 1 twice: First, an immediate implication of Theorem 1 is that W is convex, and hence, the supporting hyperplane theorem applies. Second, Theorem 1 implies we can exchange the maximization over welfare profiles in W for a maximization over posterior distributions.[8] Note we can interchangeably talk about Pareto efficient Bayes welfare profiles and Pareto efficient information structures, and we do this in what follows.

Once we note we can restrict attention to directions $\lambda \in \Delta(\Theta)$, Equation 6 has two economic interpretations. First, consider the problem of a social planner who assigns weight $\lambda(\theta)$ to type θ and wishes to maximize the weighted sum of utilities of each type. Under this interpretation, Equation 6 states that w is a solution to the social planner's problem. Second, we can interpret $\lambda^T \mathrm{w}$ as the expectation with respect to θ of the welfare profile w under the measure λ. In this case, Equation 6 implies w is the vector of *interim* payoffs of a sender with ex post payoff function w and prior λ. For instance, when $\lambda = \mu_0$, so that the sender's prior coincides with that of the outside observer, the above problem coincides with that of Kamenica and Gentzkow (2011). Instead, whenever the direction λ is any element of $\Delta(\Theta)$, the above problem coincides with that considered by Alonso and Camara (2016).[9]

Moreover, Equation 6 has an important practical implication: any Pareto efficient Bayes welfare profile is induced by the solution to a *supporting Bayesian persuasion problem*. A supporting Bayesian persuasion problem is an instance of the model in Kamenica and Gentzkow (2011) in which the sender's indirect utility function equals

$$\hat{v}_\lambda(\mu) = \lambda^T \hat{\mathrm{w}}(\mu) = \sum_{\theta \in \Theta} \mu(\theta) \frac{\lambda(\theta)}{\mu_0(\theta)} w(\mu, \theta). \qquad (7)$$

This indirect utility is the product of the truth-adjusted welfare function and the Pareto weights, which capture the rate of substitution between the truth-adjusted welfare of different types. Theorem 2 summarizes the above discussion:

[8] Indeed, for any $\mathrm{w}' \in \mathrm{W}$, a Bayes' plausible distribution over posteriors τ exists such that $\mathrm{w}'_i = \mathbb{E}_\tau[\hat{w}(\mu, \theta_i)]$ for all $i \in \{1, \ldots, N\}$, and vice versa.

[9] Thus, one can always interpret the heterogeneous priors model in Alonso and Camara (2016) as a model in which the sender and the receiver share the same prior, but the sender assigns weights different than those under the prior μ_0 to each of his possible types.

Theorem 2 (Pareto frontier). *The welfare profile* $\mathrm{w} \in \mathbb{R}^N$ *is in the Pareto frontier of* W *if and only if a direction* $\lambda \in \Delta(\Theta)$ *exists such that* w *is the profile induced by an information structure that solves the supporting Bayesian persuasion problem with indirect utility function* $\hat{v}_\lambda$.

Note that because W is convex, the characterization in Theorem 2 extends to all points on the boundary of W once we allow for any direction $\lambda \in \mathbb{R}^N \setminus \{0\}$.

Theorem 2 recovers the Pareto frontier of W by connecting the solution of the utilitarian planner with weights λ to the Bayesian persuasion problem of the sender with indirect utility function $\hat{v}_\lambda$. Whereas Theorem 1 characterizes the Bayes welfare set via the convex hull of the graph of a vector-valued function, $\hat{\mathrm{w}}$, the results in Kamenica and Gentzkow (2011) imply each of the supporting Bayesian persuasion problems can be solved by concavifying a real-valued function, $\hat{v}_\lambda$. This observation provides us with a tractable way of characterizing the Pareto efficient information structures (see, e.g., the analysis in Section 5). However, Pareto efficient information structures can result from maximizing non-utilitarian objective functions on W, such as Rawls' criterion or those that obtain from efficient bargaining solutions, such as Nash Bargaining. These objective functions cannot be ex ante reduced to a given set of Pareto weights λ so that knowledge of the whole (Pareto frontier of) W is important. For this reason, we see the characterization in Theorem 2 as complementary to that in Theorem 1.

Theorem 2 has yet another practical implication: when a (Pareto efficient) w is an extreme point of W, an information structure exists that employs at most N signals and generates w. We record this result below:

Corollary 2 (Extreme points of W). *Let* w *be an extreme point of* W. *Then, an information structure* Π *with at most* N *signals exists such that* $\mathrm{w}_i = w_\Pi(\theta_i)$ *for all* $i \in \{1, \ldots, N\}$.

In Example 1, all points in W_P are extreme (Figure 2b). Corollary 2 then implies information structures that induce at most two posteriors are enough to characterize the Pareto frontier W in Example 1.

We use Example 2 to illustrate properties of the Bayes welfare profiles in the Pareto frontier and why the bound in Corollary 2 applies only to *extreme* points on the Pareto frontier of W. In doing so, we highlight the difference between the value of the program in Equation 6 and the Bayes welfare profiles consistent with that value:

Example 2 (continued). Figure 3 illustrates the convex hull of the graph of $\hat{\mathrm{w}}$ (Figure 3a) and the Bayes welfare set (Figure 3b) for the online marketplace example. We note the following features of the Pareto frontier of W in this example, depicted in black in Figure 3b. First, as the flat segment at the top of W shows, the profits of low-quality sellers can be increased without decreasing those of high-quality sellers.

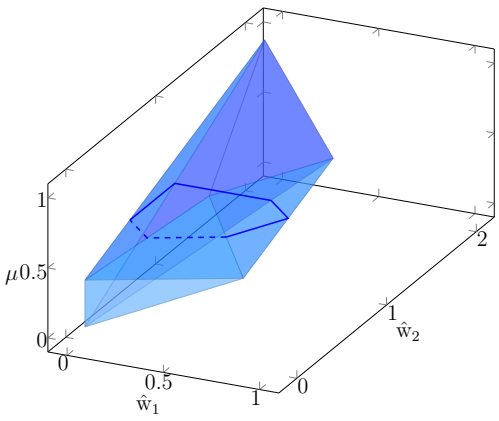

Figure (a) The convex hull of the graph of $\hat{w}$

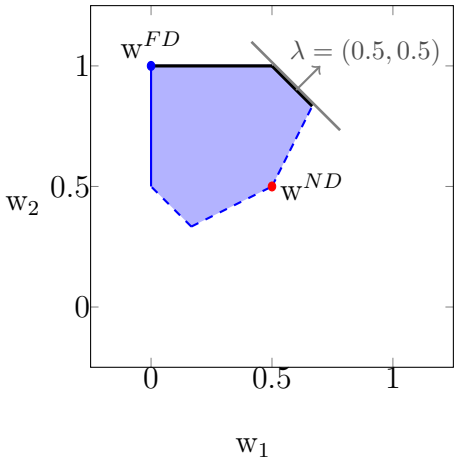

Figure (b) The Bayes welfare set W

Figure 3: Constructing the Bayes welfare set in Example 2; w^{FD} and w^{ND} denote the full- and no-information profit profiles, respectively

Second, the Pareto frontier lies entirely above the 45-degree line. In other words, information structures that are fair in the sense of equalizing profits across high- and low-quality sellers are inefficient. In this particular example, the only fair information structure is no disclosure.

Third, contrary to Example 1, not every point on the boundary of W is an extreme point. Consider, for instance, the points on the boundary in the direction $\lambda = (0.5, 0.5)$ in Figure 3b. It is possible to show that the points in the interior of this line are generated by information structures with three signals.[10] Note all the points in that line lead to the same value in the Bayesian persuasion problem with indirect utility function $\hat{v}_\lambda$, namely, $\text{cav } \hat{v}_\lambda(\mu_0)$. However, they correspond to different welfare profiles with different implications regarding how high- and low-quality sellers share the payoff $\text{cav } \hat{v}_\lambda(\mu_0)$. This highlights a benefit of the perspective we develop in this paper: insofar as one cares about the cross-sectional implications of different information structures, studying only the ex ante welfare induced by a given information structure may not be sufficient. ◊

[10] For instance, the point $w = (6/10, 9/10)$ can only be generated by an information structure that employs at least three signals. One such information structure is given by

$$\pi : \begin{pmatrix} 1/5 & 2/5 & 2/5 \\ 0 & 1/5 & 4/5 \end{pmatrix}.$$

4.1 When is (no) information disclosure efficient?

A natural question to ask is when providing some information is the efficient thing to do; in other words, when does an information structure exist that is a Pareto improvement relative to no disclosure? When such an information structure exists, we say all types *benefit from disclosure*. Examples 1 and 2 offer an interesting contrast in this respect: because the no-disclosure profile w^{ND} is not in the Pareto frontier in Example 2, all seller types benefit from disclosure. Instead, the no-disclosure profile w^{ND} is in the Pareto frontier in Example 1. Indeed, in Example 1 information disclosure necessarily hurts at least one of the types (in this case, θ_A).

Proposition 2 provides a necessary and sufficient condition for all types to benefit from disclosure. To introduce this condition, recalling a definition from Kamenica and Gentzkow (2011) is useful: a sender with indirect utility function $\hat{v}$ *benefits from persuasion* if the concavification of $\hat{v}$ at the prior exceeds the value of $\hat{v}$ at the prior; that is, $\hat{v}(\mu_0) < \operatorname{cav} \hat{v}(\mu_0)$. We have the following:

Proposition 2 (Efficiency of disclosure). *All types benefit from disclosure if and only if, for all directions $\lambda \in \Delta(\Theta)$, a sender with indirect utility $\hat{v}_\lambda$ benefits from persuasion.*

In other words, Proposition 2 states that the no-disclosure profile w^{ND} is in the Pareto frontier of W if and only if a direction $\lambda^{ND} \in \Delta(\Theta)$ exists such that a sender with indirect utility $\hat{v}_{\lambda^{ND}}$ does not benefit from persuasion. Alternatively, a social planner with Pareto weights λ^{ND} would find no disclosure to be the optimal information structure. Recall Example 1: in that case, when $\lambda = \mu_0$, the indirect utility function $\hat{v}_{\mu_0}(\mu)$ equals the strictly concave function $-(\mu - 1/2)^2$, so that no disclosure is optimal. Corollary 7 in Section A.2 shows that this observation reflects a more general result: if for all $\theta \in \Theta$ the ex post welfare function takes the form $a(\theta)w(\mu) + b(\theta)$, where $a(\theta) > 0$ and w is concave, w^{ND} is in the Pareto frontier of W.

Whereas Proposition 2 provides a condition in terms of the indirect utility function $\hat{v}_\lambda$, Corollary 3 provides conditions on the truth-adjusted welfare function $\hat{w}$ under which all types do or do not benefit from disclosure[11]:

Corollary 3 (Disclosure benefits). *The following hold:*

1. *If, for all $\theta \in \Theta$, $\hat{w}(\cdot, \theta)$ is strictly convex in a neighborhood of the prior, all types benefit from disclosure.*
2. *Instead, if a type θ exists such that $\hat{w}(\cdot, \theta)$ is concave in μ, no disclosure is Pareto efficient.*

As in Kamenica and Gentzkow (2011), concavity and convexity properties of a pay-

[11] Section A.2 provides weaker conditions under which all types benefit from disclosure.

off function determine whether information disclosure is beneficial. In contrast to Kamenica and Gentzkow (2011), the concavity and convexity properties of the truth-adjusted welfare function are what determine whether any given type can benefit from disclosure. This result can be clearly seen in Example 1, where the truth-adjusted welfare of θ_A is concave around the prior and that of θ_B is strictly convex around the prior, even though each type's welfare function is strictly concave.[12]

To understand the role of the convexity (concavity) properties of the truth-adjusted welfare function in determining the benefits from persuasion, note that information disclosure affects the welfare of a given type θ through two channels: directly through its impact on the welfare function as in Kamenica and Gentzkow (2011) and indirectly through the truth-drift adjustment. The second channel is most easily seen in the case of binary types. In that case, the property in Equation TD ensures that the posterior belief drifts along a straight line toward the true type θ. When the welfare function of individuals of type θ is convex *and* increasing in $\mu(\theta)$, both effects are positive, ensuring that individuals of type θ benefit from disclosure. Corollary 4 below summarizes this discussion:

Corollary 4 (Binary types). *Let $N = 2$. If for all $\theta \in \Theta$, $w(\mu, \theta)$ is strictly convex in μ and increasing in $\mu(\theta)$, full disclosure is uniquely Pareto efficient and both types benefit from persuasion. Instead, if for some $\theta \in \Theta$, $w(\mu, \theta)$ is concave in μ and decreasing in $\mu(\theta)$, no disclosure is Pareto efficient.*

5 Expected Reputation

We now specialize our results to the case in which the welfare function equals the expectation of some one-dimensional variable of interest, such as an individual's productivity, quality, or trade value. This is a standard way to model reputation, image, or career concerns in economics (see, e.g., Holmström, 1999, Bénabou and Tirole, 2006). It also captures the class of prediction policy problems in Rambachan et al. (2020), where a decision maker – our outside observer – selects a treatment (e.g., hiring, bail, loan-approval) on the basis of the prediction of an outcome of interest (e.g., productivity, recidivism, creditworthiness).

[12] For an even starker example, consider $\Theta = \{\theta_1, \theta_2\}$ and $w(\mu, \theta_1) = -\mu^2 - 2\mu + 3$, $w(\mu, \theta_2) = -\mu^2 + 4\mu$. Each welfare function is concave; however, for any prior distribution, the truth-adjusted welfare functions are globally convex. As a result, the unique Pareto efficient information structure is full disclosure.

Formally, we assume a *reputation vector* $\rho \in \mathbb{R}^N$ exists such that for all $\theta_i, \theta_j \in \Theta$,

$$w(\mu, \theta_i) = w(\mu, \theta_j) = \mathbb{E}_\mu[\rho(\theta)] = \sum_{k=1}^{N} \mu(\theta_k)\rho(\theta_k) = \mu^T \rho. \tag{8}$$

We refer to w as the individual's *reputation*. Thus, a Bayes welfare profile is a profile of *expected* reputations. Without loss of generality, we label ρ in increasing order, that is, $\rho_1 \leq \cdots \leq \rho_N$, so that types are labeled in increasing order of their values under ρ.

The analysis in this section allows us to focus on the *redistributive* role of information. Indeed, all information structures lead to the same ex ante welfare. That is, for any information structure Π and the corresponding Bayes welfare profile w, the ex ante welfare is:

$$\mu_0^T \mathrm{w} = \sum_{\theta \in \Theta} \mu_0(\theta) \mathbb{E}_{\Pi|\theta} \left[\tilde{\mu}^T \rho\right] = \mathbb{E}_\Pi \left[\tilde{\mu}^T \rho\right] = \mu_0^T \rho. \tag{9}$$

However, as the results in this section illustrate, different information structures lead to different welfare profiles, so that the chosen information structure determines how the different types in the population share the ex ante welfare.

When w is as in Equation 8, we can provide an alternative characterization of the set W. From Section 3, it follows that $\mathrm{w} \in \mathrm{W}$ if and only if we can find a Bayes' plausible posterior distribution $\tau \in \Delta_{\mu_0}(\Delta(\Theta))$ such that

$$\mathrm{w} = \mathbb{E}_\tau \left[\hat{\mathrm{w}}(\tilde{\mu})\right] = \mathbb{E}_\tau \left[\frac{\tilde{\mu}}{\mu_0}(\tilde{\mu}^T \rho)\right] = \mathrm{D}_0 \mathbb{E}_\tau \left[\tilde{\mu}\tilde{\mu}^T\right] \rho, \tag{10}$$

where D_0 denotes a diagonal matrix with (i,i)-th element equal to $1/\mu_0(\theta_i)$.

Equation 10 shows a Bayes welfare profile can be represented as the product of three terms: the reputation vector ρ, the prior-normalizing matrix D_0, and the matrix $\mathbb{E}_\tau[\mu\mu^T]$. Furthermore, the matrix $\mathbb{E}_\tau\left[\mu\mu^T\right]$ satisfies the following two properties. First, it is a *completely positive matrix* (Berman, 1988): an $N \times N$ matrix C is completely positive if it can be written as $\sum_{m=1}^{M} \mathrm{x_m x_m^T}$ for some finite collection of non-negative vectors $\mathrm{x_m} \in \mathbb{R}_+^N$.[13] Second, the rows of the matrix $\mathbb{E}_\tau\left[\mu\mu^T\right]$ add up to the prior: $\mathbb{E}_\tau\left[\mu\mu^T\right] \mathrm{e} = \mathbb{E}_\tau\left[\mu(\mu^T \mathrm{e})\right] = \mathbb{E}_\tau\left[\mu\right] = \mu_0$. Theorem 3 shows these two properties are not only necessary but also sufficient and thus fully characterize the Bayes welfare set:[14]

[13] Completely positive matrices have been studied extensively as they play an important role in optimization theory, machine learning, and other applications (Berman and Shaked-Monderer, 2003). A completely positive matrix is symmetric and positive-semidefinite, with positive elements; for $N < 5$, the converse is also true.

[14] An analogous characterization appears in concurrent work by Sayin and Başar (2021).

Theorem 3. *Given the reputation vector* ρ, $\mathrm{w} \in \mathrm{W}$ *if and only if a completely positive matrix* $\mathrm{C} \in \mathbb{R}^{N \times N}$ *exists such that* $\mathrm{Ce} = \mu_0$ *and*

$$\mathrm{w} = \mathrm{D}_0 \mathrm{C} \rho.$$

Putting together the properties in Theorem 3, we obtain that any Bayes welfare profile w is the product of the reputation vector ρ with a matrix P, where $\mathrm{P} \equiv \mathrm{D}_0\mathrm{C}$ is the transition matrix of a time-reversible Markov chain with invariant distribution μ_0. That is, (i) $\mu_0^T \mathrm{P} = \mu_0^T$, (ii) $\mathrm{Pe} = \mathrm{e}$, and (iii) P satisfies the *detailed balance conditions*: for all $i, j \in N$, $\mu_{0i}\mathrm{P}_{ij} = \mu_{0j}\mathrm{P}_{ji}$. The first property captures the pure redistribution of welfare highlighted in Equation 9. The second property implies any Bayes welfare profile can be viewed as a *garbled* version of the full information profile ρ. The third property delineates the limits of how payoffs can be redistributed by linking how much of $\rho(\theta_i)$ can be attributed to θ_j, and vice versa. Indeed, because P is the transition matrix of a time-reversible Markov chain, we obtain that there is *mean reversion* in the redistribution of payoffs across types. To see this, note that if $\mathrm{w} = \mathrm{P}\rho \in \mathrm{W}$, then also $\mathrm{Pw} \in \mathrm{W}$.[15] Because μ_0 is the invariant distribution of P, we have that $\mathrm{P}^k \mathrm{w} \to_{k \to \infty} (\mu_0^T \mathrm{w}) * \mathrm{e} = (\mu_0^T \rho) * \mathrm{e} = \mathrm{w}^{ND}$, where w^{ND} is the no-disclosure profile.

Remark 3 (Connections to the literature). *Reversible Markov chains are prominent in the study of higher-order beliefs and expectations of higher-order beliefs (Samet, 1998; Cripps et al., 2008; Golub and Morris, 2017). Indeed, note that when the welfare function is linear and type independent, a Bayes welfare profile is a vector of second-order expectations: for any type* θ *and any* $\mathrm{w} \in \mathrm{W}$, $\mathrm{w}(\theta)$ *is the expectation under some information structure of the random variable* $\mu^T \rho$ *of an individual that knows* θ*. Whereas that literature takes the information structure as given and shows (sequences of) higher-order expectations can be obtained by iteratively applying the transition matrix of a reversible Markov chain, Theorem 3 identifies* which *transition matrices are consistent with* some *information structure and shows complete positivity is the key property they must satisfy.*

Theorem 3 also relates to the literature on majorization (Hardy et al., 1952): if all types are equally likely, P *is doubly stochastic and* ρ *majorizes* w*. However, not any profile majorized by* ρ *is a Bayes welfare profile, because not all doubly stochastic matrices are symmetric, and hence, some do not satisfy the detailed balance conditions.*

We now show how Theorem 3 delivers a more general version of the truth-drifting property discussed in Section 3. This property has been obtained in different forms in the literature that studies the feasible evolution of beliefs (e.g., Francetich and Kreps, 2014, Hart and Rinott, 2020). Truth-drifting states that whereas an information structure can occasionally "deceive" the outside observer about an individual's true type,

[15] $\mathrm{P}^2\mathrm{e} = \mathrm{Pe} = \mathrm{e}$ and $\mathrm{P}^2 = \mathrm{D}_0\mathrm{C}'$, where $\mathrm{C}' \equiv \mathrm{C}\mathrm{D}_0\mathrm{C}$ is completely positive because C is symmetric.

it cannot systematically do so. This property underlies the limits of using information as a tool to distribute welfare in the population.

Formally, consider any event X that is correlated with types according to the conditional probability function $\beta \in [0,1]^N$, $\beta_i \equiv \Pr(X \mid \theta_i)$, so that the prior probability of the event is $\Pr(X) = \mu_0^T \beta$.[16] If all $\beta_i \in \{0,1\}$, the event effectively indicates a subset of types. More generally, the event may involve extraneous uncertainty, and the types may be only imperfectly informative about it. We show that if the event is true, the average posterior probability that the observer attaches to this event must be at least as large as the prior probability:

Claim 2 (Truth drifting). *For any event X and information structure Π,*

$$\mathbb{E}_\Pi \left[\Pr(X \mid s) \mid X \right] \geq \Pr(X).$$

Francetich and Kreps (2014) obtain this result, relying on the properties of Kullback-Leibler divergence. Hart and Rinott (2020) obtain a version of Claim 2 in the special case of $X \subseteq \Theta$, relying on the monotone-likelihood ratio property. Instead, our proof of Claim 2, presented in the appendix, Section A.3, builds on the fact that an underlying matrix C is completely positive and thus necessarily positive semi-definite.

Boundary information structures: Recall that Theorem 2 characterizes the boundary of the Bayes welfare set by means of supporting Bayesian persuasion problems. In the reputation model, Equation 9 implies the Bayes welfare set lies within a hyperplane with an orthogonal vector μ_0. Thus, instead of studying the boundary of the Bayes welfare set, we focus on its *relative* boundary, which consists of all Bayes welfare profiles not in the relative interior of the Bayes welfare set.[17] In a slight abuse of terminology, we refer to the profiles in the relative boundary of W and the information structures that induce them as *boundary* profiles and information structures, respectively.

As we show next, the supporting Bayesian persuasion problems in the reputation model take a well-known structure. Indeed, fix a direction λ not collinear with μ_0 and consider the induced supporting Bayesian persuasion problem:

$$\begin{aligned} \max_{\tau \in \Delta(\Delta(\Theta))} \mathbb{E}_\tau \left[\lambda^T \hat{w}(\mu) \right] &= \max_{\tau \in \Delta_{\mu_0}(\Delta(\Theta))} \mathbb{E}_\tau \left[\left(\frac{\lambda}{\mu_0}^T \mu \right) \left(\rho^T \mu \right) \right] \qquad (RS_\lambda) \\ &= \max_{\tau \in \Delta_{\mu_0}(\Delta(\Theta))} \mathbb{E}_\tau \left[\mathbb{E}_\mu \left[\frac{\lambda(\theta)}{\mu_0(\theta)} \right] \mathbb{E}_\mu \left[\rho(\theta) \right] \right], \end{aligned}$$

[16] For concreteness, X can be seen as a subset of $\Theta \times [0,1]$ equipped with a probability measure that agrees with μ_0 on Θ (Green and Stokey, 1978; Gentzkow and Kamenica, 2017).

[17] Recall that the relative interior of a set X is the interior of X within its affine hull, which is the set of all affine combinations of elements in X.

where the first equality uses the form of w and the definition of $\hat{w}$. Equation RS_λ shows that if an information structure Π delivers a profile w in the relative boundary of W, the information structure solves an instance of the information design problem in Rayo and Segal (2010). To be precise, Rayo and Segal (2010) consider the following problem. A sender owns a prospect, and his objective is that the receiver accepts it. When the sender's type is θ and the receiver accepts the prospect, the sender and the receiver obtain a payoff $\gamma(\theta) \equiv \lambda(\theta)/\mu_0(\theta)$ and $\rho(\theta) \in [0,1]$, respectively. Instead, if the receiver rejects the prospect, the sender obtains a payoff of 0, whereas the receiver obtains a payoff u distributed uniformly over $[0,1]$ independently of θ. The sender chooses an information structure, Π, without observing the realization of u. Thus, when Π induces a belief μ, the sender expects the receiver to accept the project with probability, $\rho^T\mu$. It follows that the last term in Equation RS_λ represents the sender's expected payoff when τ is the posterior distribution induced by information structure Π.

Proposition 3. (Boundary profiles) *A welfare profile* w *is in the relative boundary of* W *if and only if a direction* $\lambda \in \mathbb{R}^N \setminus \{0\}$ *not collinear with* μ_0 *exists such that* w *is induced by an information structure that solves Equation* RS_λ.

Proposition 3 allows us to rely on the approach of Rayo and Segal (2010) to characterize the *shape* of the information structures that achieve the boundary Bayes welfare profiles. This approach relies on a graphical representation of an information structure, in which the prospect values $\{(\gamma(\theta), \rho(\theta)) : \theta \in \Theta\}$ are the nodes (see Remark 5). The results in Rayo and Segal (2010) have immediate implications for the information structures that induce the boundary profiles of W:

Corollary 5. *In the reputation model, the following hold:*

1. *An information structure that induces a boundary Bayes welfare profile in the direction* λ *separates types* θ_i *and* θ_j *whenever their ranking under the vector* λ/μ_0 *and the reputation vector* ρ *is the same;*

2. *The full and no-disclosure Bayes welfare profiles are in the relative boundary of* W.

The first part of Corollary 5 highlights a natural feature of optimal information provision in the reputation model in terms of the alignment of preferences of the social planner, captured by λ/μ_0, and of the outside observer, captured by ρ: the planner should separate any two types as long as the planner and the outside observer are in agreement about the types' relative ranking. The second part of Corollary 5 shows that the full and no-disclosure Bayes welfare profiles are on the boundary of the Bayes welfare set. Whereas this property holds in our illustrative examples, despite the welfare function not being linear, this property is not a general one, as we illustrate in

Example 4 in the appendix.

Individual reputation bounds: We can further build on the graphical approach of Rayo and Segal (2010) to characterize the information structures that deliver maximal (or minimal) welfare to any given type. This exercise provides a rough way to bound the Bayes welfare set W and also suggests how information may be employed to boost (or dilute) the reputation of particular types in the population. In the context of our credit agency example, Example 3, the information structure that maximizes the welfare of individuals of type θ_i maximizes the probability that individuals of type θ_i obtain credit.

Formally, given a target type θ_i, we want to solve the following problem:

$$\max_{\mathrm{w}\in\mathrm{W}} \mathrm{w}_i. \tag{i-MAX}$$

Proposition 4 below shows a particular class of information structures solves the problem i-MAX.

Definition 3 (Noisy priority). *A θ_i-noisy-priority policy with threshold k is an information structure (π, S) such that $S = \Theta$, and the likelihood function π satisfies:*

1. *If $j \neq i$, $\pi(s = \theta_j \mid \theta_j) = 1$,*
2. *If $j < k$, $\pi(s = \theta_j \mid \theta_i) = 0$, and*
3. *If $j \geq k$, $\pi(s = \theta_j \mid \theta_i) > 0$.*

In other words, a θ_i-noisy-priority policy pairwise pools the target type θ_i with all types with indices above some threshold and separates all other types. A noisy-priority policy has an implementation akin to the priority mechanisms in the matching literature (Celebi and Flynn, 2022), and hence its name. Indeed, a noisy-priority policy with threshold $k \geq i$ can be implemented by first assigning a perfectly revealing score to each type equal to their index, and then prioritizing the target type by increasing this type's score by a random number.

Proposition 4. *The Bayes welfare profile that solves i-MAX is induced by a θ_i-noisy-priority policy with threshold $k \geq i$.*

The proof is in Section A.3 in the appendix. One part of Proposition 4 is straightforward: if one wishes to increase the reputation of θ_i, then θ_i should be separated from all types with lower indices. What might be less obvious is that whenever θ_i is pooled with some other type, θ_i should be pooled with it pairwise. In a sense, pooling several types together redistributes the reputation from higher-quality types to lower-quality types. Pairwise pooling then allows the target type to obtain maximal reputation gains from any other type without sharing the gains with others. Finally,

randomized pooling with many types ensures no signal is overly "muddled," which in turn ensures an overall high reputation for θ_i.

By simply reversing signs, Proposition 4 can be used to characterize the information structure that minimizes the expected reputation of individuals of type θ_i: this information structure should pairwise pool the target type with types whose indices are below some threshold. Such adversarial pairwise pooling inflicts maximal reputation losses and can be viewed as a *noisy-degrading policy*.

We conclude this section by illustrating the Bayes welfare set in the context of Example 3 and highlight the importance of population heterogeneity as captured by the number of types.

Example 3 (continued). Figure 4 illustrates the results of this section in the context of Example 3. Recall that in that case $\rho(\theta)$ denotes the probability that an individual of type θ repays the loan, and hence, $w(\mu, \theta)$ is the expected repayment probability under belief μ. Like in Rayo and Segal (2010), we assume the credit agency has a uniform outside option. Assuming individuals wish to maximize the probability the lending agency approves the loan justifies that $w(\cdot)$ is their ex post welfare.

Figure 4a depicts the individually feasible welfare profiles (dashed square) and the Bayes welfare set (blue line) in the case of $N = 2$. Proposition 1 implies any payoff between $\rho_1 = 0$ and $\mu_0^T \rho$ is feasible for θ_1, whereas any payoff between $\mu_0^T \rho$ and $\rho_2 = 1$ is feasible for θ_2. As Figure 4a illustrates, the Cartesian product $[\rho_1, \mu_0^T \rho] \times [\mu_0^T \rho, \rho_2]$ is a rather lax bound in this example. In particular, the latter set ignores that all Bayes welfare profiles satisfy $\mu_0^T \mathrm{w} = \mu_0^T \rho = 0.5$ (Equation 9). In the case of $N = 2$, adding this restriction is enough to pin down the Bayes welfare set. In particular, because as noted in Theorem 3, all Bayes welfare profiles can be obtained by "garbling" the full-disclosure Bayes welfare profile, ρ, and in the case of binary types, this garbling turns out to span a linear segment. Finally, note the structure of the Bayes welfare set implies a social planner with Pareto weights λ finds it optimal to provide no or full information, depending on whether the planner weighs the welfare of type-θ_1 individuals more than that of type-θ_2 individuals (that is, $\lambda_1 \lessgtr \lambda_2$).

Figure 4b depicts the Bayes welfare set W in the case of $N = 3$. In contrast to the binary-type case, the boundary of the Bayes welfare set is non-linear and features a continuum of extreme points. This example illustrates that the constraints imposed by Bayes' plausibility are richer than the simple garbling constraint that characterizes the set when $N = 2$. As we show in Section A.3, four classes of information structures span the boundary of the Bayes welfare set: the noisy-priority and noisy-degrading policies for θ_1 and the noisy-priority and noisy-degrading policies for θ_3. The θ_1-noisy-degrading and θ_3-noisy-priority policy policies span the linear segments in Figure 4b. The first minimizes the loan-approval probability of individuals of θ_1 by separating

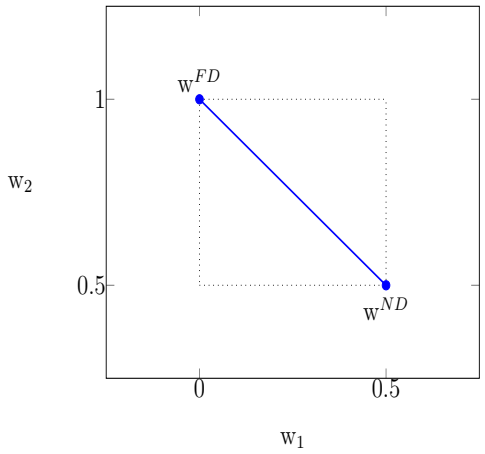

Figure (a) $N = 2$, $\rho = (0, 1)$, $\mu_0 = (1/2, 1/2)$

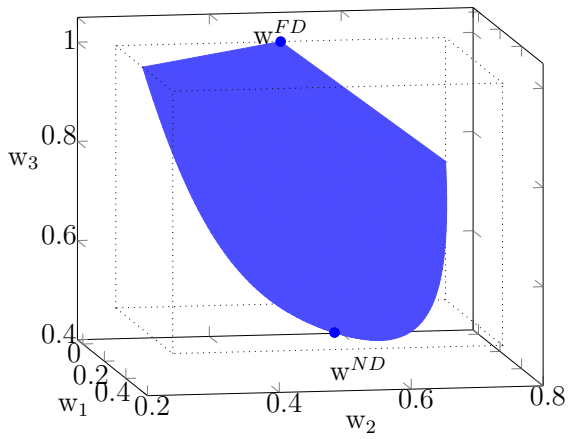

Figure (b) $N = 3$, $\rho = (0, 0.5, 1)$, $\mu_0 = (1/3, 1/3, 1/3)$

Figure 4: Expected reputation in Example 3. The blue color marks Bayes welfare set $\mathbb{W}$. The dashed segments outline the Cartesian product of individual welfare sets.

them from θ_2 and θ_3 individuals, whereas the second maximizes the loan-approval probability of individuals of θ_3 by separating them from θ_1 and θ_2 agents. Instead, the θ_1-noisy-priority policy and θ_3-noisy degrading policies span the nonlinear segment of the boundary. Notably, for almost all boundary points, type-θ_2 individuals are pooled with individuals of some other type. Intuitively, individuals of θ_2 exert *pooling* externalities on individuals of types θ_1 and θ_3.[18] For instance, individuals of θ_2 enable boosting the loan-approval probability of individuals of θ_1 in the case of the θ_1-noisy-priority policy, but exert a negative externality on individuals of θ_3 in the θ_3-noisy-degrading policy. ◊

6 Data Limits

The analysis in the previous sections assumes that the information structure can arbitrarily condition on an individual's payoff-relevant characteristics. However, this assumption does not necessarily hold in many applications of interest. For instance, regulation may prevent the disclosure of protected characteristics, such as gender or race. Thus, when θ encompasses such characteristics, considering information structures that respect these restrictions is natural.

In this section, we extend our analysis by removing this assumption. Formally, we con-

[18] We follow the terminology in (Galperti et al., Forthcoming), who highlight that individuals with certain types are valuable precisely because of the possibility of pooling them with other types.

sider the following extension of the model in Section 2. Together with the individuals' characteristics, we are given a data source that is potentially informative about these characteristics. The data source has realizations in a finite set $D \equiv \{d_1, \ldots, d_M\}$. We describe the joint distribution over payoff-relevant characteristics and data via the prior distribution on Θ, μ_0, and a system of conditional probabilities $\{\nu_0(\cdot|\theta) : \theta \in \Theta\}$, describing the distribution of the data source d conditional on the characteristics θ. We let $\eta_0 \in \Delta(D)$ denote the induced marginal distribution on D.[19] The model in Section 2 corresponds to the case in which $\Theta = D$ and $\nu_0(d|\theta) = \mathbb{1}[d = \theta]$.

We assume information can be provided to the outside observer only about data-source realizations, but not an individual's type. Formally, an information structure $\Pi = (\pi, S)$ consists of a countable set of labels S and a mapping π, which associates to each data-source realization, d, a distribution over signals $\pi(\cdot|d) \in \Delta(S)$. Given an information structure Π and a signal realization $s \in S$, updated beliefs about θ depend only on the updated belief about the realization of d. Indeed,

$$\mu_s(\theta) = \sum_{d \in D} \frac{\mu_0(\theta)\nu_0(d|\theta)}{\eta_0(d)} \eta_s(d), \tag{11}$$

where η_s is the marginal on D of the updated joint belief on $\Theta \times D$. It follows that we can define the welfare function as depending on beliefs about d rather than about θ. That is, we can define the function $w_\dagger : \Delta(D) \times \Theta \mapsto \mathbb{R}$ as follows:

$$w_\dagger(\eta, \theta) = w(\mu(\eta), \theta),$$

where the function $\mu(\eta)$ is determined by Equation 11.

Given an information structure (π, S), the welfare of an individual of type θ is

$$w_\Pi(\theta) \equiv \mathbb{E}_{\Pi|\theta}[w_\dagger(\tilde{\eta}, \theta)] = \sum_{s \in S} \sum_{d \in D} \nu_0(d|\theta)\pi(s|d) w_\dagger(\eta_s, \theta), \tag{12}$$

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

We now show the analysis in the previous sections extends verbatim. Indeed, by the same arguments as in Section 2, the welfare of an individual of type θ under information structure $\Pi = (\pi, S)$ can be written as:

$$w_\Pi(\theta) = \mathbb{E}_\Pi[\hat{w}_\dagger(\tilde{\eta}, \theta)] = \sum_{s \in S} \Pr_\Pi(s) \hat{w}_\dagger(\eta_s, \theta), \tag{13}$$

where the truth-adjusted welfare function $\hat{w}_\dagger$ now takes the form:

$$\hat{w}_\dagger(\eta, \theta) = \sum_{d \in D} \nu_0(d|\theta) \frac{\eta(d)}{\eta_0(d)} w_\dagger(\eta, \theta). \tag{14}$$

[19] Although we should index η_0 by $\left(\mu_0, (\nu_0(\cdot|\theta))_{\theta \in \Theta}\right)$, we omit this dependence to simplify notation.

By separating the variable on which welfare is conditioned on – the payoff-relevant characteristics θ– from the variable about which information is provided – the data source d– Equation 14 allows us to provide further insight into the truth-adjusted welfare function in the model in Section 2. Indeed, note the likelihood correction is based on the variable d, highlighting that it corresponds to the variable about which information is provided. Similar to before, we can interpret the likelihood-ratio adjustment as describing that each data-source realization d has a budget $\eta_0(d)$ to be distributed across different (data) posteriors η. Unlike the analysis before, individuals of type θ only own a fraction $\nu_0(d|\theta)$ of this ratio.

Equation 13 implies Theorem 1 immediately extends to this setting:

Theorem 4. *The Bayes welfare set* W *satisfies the following:*

$$\mathrm{W} = \left\{ \mathrm{w} \in \mathbb{R}^N : (\eta_0, \mathrm{w}) \in \mathrm{co}\left(\mathrm{graph}\, \hat{\mathrm{w}}_\dagger\right) \right\}. \tag{15}$$

In what follows, we explore how the Bayes welfare set changes as we change the informativeness of the data source. Intuitively, we would expect that the Bayes welfare set shrinks as data becomes *less precise*. The extreme cases in which data provides no or full information about types provide a simple illustration. When data provides no information about types, no scope exists to provide different welfare to individuals with different characteristics, so that the Bayes welfare set effectively collapses to a point. Instead, when the data source perfectly reveals the type – as is the case in the analysis of the previous sections– the scope for payoff redistribution across individuals with different types is the largest. Proposition 5 below shows that the above intuition holds when the notion of less precise coincides with the notion of *garbling* as in Blackwell (1953).

Formally, given the distribution of payoff-relevant characteristics $\mu_0 \in \Delta(\Theta)$, we wish to understand the effect of different data sources, as described by data-source realizations D' and conditional probability systems $\{\nu_0'(\cdot|\theta) \in \Delta(D') : \theta \in \Theta\}$. Following Blackwell (1953), we say (D', ν_0') is a *garbling* of (D, ν_0) if a stochastic matrix $G : D \mapsto \Delta(D')$ exists such that for every data-type pair (d', θ),

$$\nu_0'(d'|\theta) = \sum_{d \in D} G(d'|d)\nu_0(d|\theta).$$

Let $\mathrm{W}(\mu_0, w, D, \nu_0)$ denote the Bayes welfare set for prior type distribution θ and welfare function w, as we vary the (informativeness of the) data source (D, ν_0). We then have the following:

Proposition 5 (Data Comparison). $\mathrm{W}(\mu_0, w, D', \nu_0') \subseteq \mathrm{W}(\mu_0, w, D, \nu_0)$ *for all welfare functions* w *and type distributions* μ_0 *if and only if* (D', ν_0') *is a garbling of* (D, ν_0).

Proposition 5 builds on Blackwell's theorem. One direction is straightforward: if (D', ν_0') is a garbling of (D, ν_0), any distribution of signals conditional on payoff-relevant characteristics induced by some information structure under data source (D', ν_0') is feasible under data source (D, ν_0). Consequently, any welfare profile that can be induced by some information structure under (D', ν_0') can be induced under (D, ν_0).

The other direction of Proposition 5 is based on the insight that for a given welfare function, the Bayes welfare set can be viewed as the set of interim payoffs of a decision maker under different garblings of the data source. In that case, the value of the data source equals the maximal average welfare in W, weighted by the prior distribution. If one data source is not a garbling of another, one can construct a welfare function and a prior distribution that would make this data source strictly more valuable, which would violate the presumed set inclusion.

Remark 4 (When to blind an algorithm)**.** *Proposition 5 stands in contrast with the recommendation in the algorithmic fairness literature to "blind" algorithms to sensitive inputs such as race or gender. Indeed, having taken into account the impact of the outside observer's incentives in the population's welfare, allowing the information structure to condition on the individuals' payoff-relevant characteristics leads to the largest Bayes welfare set, thereby increasing the value of any social welfare function that is used to choose what information structure to implement.*

There may be other reasons outside our model that could justify blinding the information structure to the individuals' types. For instance, Liang et al. (2022) show that when an agent different from the social planner selects a decision-making algorithm, the social planner may prefer to restrict the inputs into the agent's algorithm. Even though in our model algorithms are information structures and not mechanisms as in Liang et al. (2022), a similar result would hold in our setting.

We conclude this section with an example that illustrates the following two points. First, whereas Proposition 5 shows that less precise data sources limit the ability to generate and distribute welfare via information, the example shows that this effect is not uniform across individuals of different types. Second, the Bayes welfare set may collapse to the no-disclosure Bayes welfare profile for data sources that are strictly more informative than no information in the Blackwell order.

Example 2 (continued; Noisy Data)**.** Suppose the online marketplace only has access to a noisy estimate of the seller's type, perhaps from past consumer reviews. We model this as a data source that reveals a seller's type with a fixed precision $\sigma \in [1/2, 1]$: $D = \{d_1, d_2\}$ and $\nu_0(d_i|\theta_i) = \sigma$. When $\sigma = 1$, the data source is perfectly informative about a seller's type; when $\sigma = 1/2$, the data source is pure noise. More generally, if $\sigma < \sigma'$, the data source that corresponds to σ is a garbling of the data source that

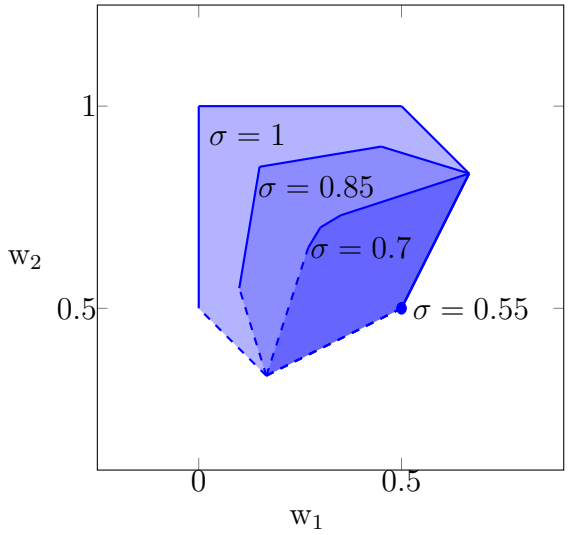

Figure 5: Noisy data in the online marketplace. Bayes welfare sets for different values of $\sigma \in \{0.55, 0.7, 0.85, 1\}$.

corresponds to σ'.

Figure 5 illustrates the Bayes welfare set W for different values of precision σ. Three features are worth noting. First, in line with Proposition 5, Bayes welfare sets resulting from data sources with lower precision are subsets of those with higher precision. When $\sigma = 1$, the Bayes welfare set naturally coincides with the one in Figure 3b. Second, at high values of σ, lower data precision has asymmetric effects across types: it decreases the maximal payoff of high-quality sellers without affecting their minimal payoff, yet it increases the minimal payoff of low-quality sellers without affecting their maximal payoff. Indeed, for sufficiently low values of σ, the unique Pareto efficient information structure is the one that maximizes the payoff of low-quality sellers. That is, in this example, lower data precision benefits low-quality sellers. Finally, whereas it is immediate that the Bayes welfare set coincides with the no-information profile w^{ND} when $\sigma = 1/2$, the Bayes welfare set actually collapses to this point at $\sigma = 2/3$: Once $\sigma < 2/3$, generating Bayes' plausible posterior distributions with support outside the interval $[1/3, 2/3]$ is not possible, and on this interval, w is constant. This feature highlights that an incrementally more informative data source may have a discontinuous impact on welfare redistribution possibilities. ◊

7 Conclusion

We provide a framework to study the potentially disparate impact of information policies in a population of heterogeneous individuals. As information policies increasingly shape society's choices in high-stakes domains, the Bayes welfare set describes the limits of what society can achieve under some information structure and the welfare trade-offs implied by the choice between different information structures. In the spirit of mechanism design and information design, our characterization of the Bayes welfare set provides a unifying tool to evaluate the welfare implications of different information policies across a wide array of objective functions.

We see several avenues worth exploring in future work. First, motivated by recent policies, Tirole (2021) studies the use of information in the form of a social score to incentivize good behavior in the population. Whereas Tirole (2021) studies this question in the context of a parametric family of information structures, our initial explorations show the Bayes welfare set allows us to extend his results by allowing *any* information structure. More generally, information has been suggested as a substitute for monetary incentives, and the Bayes welfare set describes what can be achieved with information alone.

Second, whereas we characterize individuals' welfare as a function of their type, thinking of applications in which we care instead about the welfare of *groups* is natural. For instance, an individual's type could encompass their gender and their ability, and the social planner is concerned with the welfare different genders may obtain. Our working paper version, Doval and Smolin (2021), extends our characterization of the Bayes welfare set to this setting but leaves the shape of efficient information structures as an open question.

References

Alonso, R. and O. Camara (2016): "Bayesian Persuasion with Heterogeneous Priors," *Journal of Economic Theory*, 165, 672–706.

Angwin, J., J. Larson, S. Mattu, and L. Kirchner (2016): "Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased Against Blacks." *ProPublica*, 23, 77–91.

Arieli, I., Y. Babichenko, and F. Sandomirskiy (2022): "Persuasion as Transportation," in *Proceedings of the 23rd ACM Conference on Economics and Computation, New York, NY, USA: Association for Computing Machinery, EC*, vol. 22, 468.

Aumann, R. (1987): "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica*, 55, 1–18.

Aumann, R. and M. Maschler (1995): *Repeated Games with Incomplete Information*, MIT Press.

Bénabou, R. and J. Tirole (2006): "Incentives and Prosocial Behavior," *American Economic Review*, 96, 1652–1678.

Bergemann, D., B. Brooks, and S. Morris (2015): "The Limits of Price Discrimination," *American Economic Review*, 105, 921–57.

Berman, A. (1988): "Complete Positivity," *Linear Algebra and its Applications*, 107, 57 – 63.

Berman, A. and N. Shaked-Monderer (2003): *Completely Positive Matrices*, World Scientific.

Blackwell, D. (1953): "Equivalent Comparisons of Experiments," *The Annals of Mathematical Statistics*, 265–272.

Celebi, O. and J. Flynn (2022): "Adaptive Priority Mechanisms," .

Cripps, M. W., J. C. Ely, G. J. Mailath, and L. Samuelson (2008): "Common Learning," *Econometrica*, 76, 909–933.

Doval, L. and V. Skreta (2022): "Mechanism Design with Limited Commitment," *Econometrica*, 90, 1463–1500.

Doval, L. and A. Smolin (2021): "Information Payoffs: An Interim Perspective," *arXiv preprint arXiv:2109.03061*.

Francetich, A. and D. Kreps (2014): "Bayesian Inference Does Not Lead You Astray… On Average," *Economics Letters*, 125, 444–446.

Fréchette, G. R., A. Lizzeri, and J. Perego (2022): "Rules and Commitment in Communication: An Experimental Analysis," *Econometrica*, 90, 2283–2318.

Galperti, S., A. Levkun, and J. Perego (Forthcoming): "The Value of Data Records," *Review of Economic Studies*.

Gentzkow, M. and E. Kamenica (2017): "Bayesian Persuasion with Multiple Senders and Rich Signal Spaces," *Games and Economic Behavior*, 104, 411–429.

Golub, B. and S. Morris (2017): "Higher-Order Expectations," *Available at SSRN 2979089*.

Green, J. and N. Stokey (1978): "Two Representations of Information Structures and their Comparisons," IMSSS, Stanford University.

Hardy, G. H., J. E. Littlewood, G. Pólya, G. Pólya, D. Littlewood, et al. (1952): *Inequalities*, Cambridge University Press.

Hart, S. and Y. Rinott (2020): "Posterior Probabilities: Dominance and Optimism," *Economics Letters*, 194, 109352.

Hiriart-Urruty, J.-B. and C. Lemaréchal (2004): *Fundamentals of convex analysis*, Springer Science & Business Media.

Holmström, B. (1999): "Managerial Incentive Problems - A Dynamic Perspective," *Review of Economic Studies*.

Jagtiani, J. and C. Lemieux (2019): "The Roles of Alternative Data and Machine Learning in Fintech Lending: Evidence from the LendingClub Consumer Platform," *Financial Management*, 48, 1009–1029.

Kamenica, E. and M. Gentzkow (2011): "Bayesian Persuasion," *American Economic Review*, 101, 2590–2615.

Kleinberg, J., J. Ludwig, S. Mullainathan, and A. Rambachan (2018): "Algorithmic Fairness," in *AEA Papers and Proceedings*, vol. 108, 22–27.

Koessler, F. and V. Skreta (Forthcoming): "Informed Information Design," *Journal of Political Economy*.

Kučak, D., V. Juričić, and G. Đambić (2018): "Machine Learning in Education– a Survey of Current Research Trends," *Annals of DAAAM & Proceedings*, 29.

Laclau, M. and L. Renou (2017): "Public Persuasion," .

Levy, G., I. Moreno de Barreda, and R. Razin (2021): "Feasible Joint Distributions of Posteriors: A Graphical Approach," *Working Paper*.

Li, D., L. R. Raymond, and P. Bergman (2020): "Hiring as Exploration," *National Bureau of Economic Research*.

Liang, A., J. Lu, and X. Mu (2022): "Algorithmic Design: Fairness versus Accuracy," in *Proceedings of the 23rd ACM Conference on Economics and Computation*, 58–59.

Lipnowski, E. and D. Ravid (2020): "Cheap Talk with Transparent Motives," *Econometrica*, 88, 1631–1660.

Mukherjee, A. (2008): "Sustaining Implicit Contracts When Agents Have Career Concerns: the Role of Information Disclosure," *The RAND Journal of Economics*, 39, 469–490.

Mullainathan, S. (2018): "Algorithmic fairness and the social welfare function," in *Proceedings of the 2018 ACM Conference on Economics and Computation*, 1–1.

Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan (2019): "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science*, 366, 447–453.

Ostrovsky, M. and M. Schwarz (2010): "Information Disclosure and Unraveling in Matching Markets," *American Economic Journal: Microeconomics*, 2, 34–63.

Perez-Richet, E. (2014): "Interim Bayesian Persuasion: First Steps," *American Economic Review*, 104, 469–74.

Quigley, D. and A. Walther (2019): "Contradiction-Proof Information Design," *Working Paper*.

Raghavan, M., S. Barocas, J. Kleinberg, and K. Levy (2020): "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469–481.

Rambachan, A., J. Kleinberg, J. Ludwig, and S. Mullainathan (2020): "An Economic Perspective on Algorithmic Fairness," in *AEA Papers and Proceedings*, vol. 110, 91–95.

Rayo, L. and I. Segal (2010): "Optimal Information Disclosure," *Journal of Political Economy*, 118, 949–987.

Rosar, F. (2017): "Test Design Under Voluntary Participation," *Games and Economic Behavior*, 104, 632–655.

Saeedi, M. and A. Shourideh (2020): "Optimal Rating Design," *arXiv preprint arXiv:2008.09529*.

Salamanca, A. (2021): "The Value of Mediated Communication," *Journal of Economic Theory*, 192, 105191.

Samet, D. (1998): "Iterated Expectations and Common Priors," *Games and Economic Behavior*, 24, 131–141.

Sayin, M. O. and T. Başar (2021): "Bayesian Persuasion with State-Dependent Quadratic Cost Measures," *IEEE Transactions on Automatic Control*, 67, 1241–1252.

Tirole, J. (2021): "Digital Dystopia," *American Economic Review*, 111, 2007–48.

A Omitted results and proofs from the main text

A.1 Omitted proofs from Section 3

Proof of Claim 1. For an information structure Π, let supp(Π) denote the support of the posterior belief distribution induced by Π. Then, for a given type θ, their welfare under information structure Π can be written as follows:

$$\begin{aligned} w_\Pi(\theta) = \mathbb{E}_{\Pi|\theta}\left[w(\tilde{\mu},\theta)|\theta\right] &= \sum_{\mu\in\text{supp}(\Pi)}\sum_{s\in S:\mu_s=\mu} \pi(s|\theta)w(\mu,\theta) \\ &= \sum_{\mu\in\text{supp}(\Pi)}\sum_{s\in S:\mu_s=\mu} \Pr_\Pi(s)\frac{1}{\mu_0(\theta)}\frac{\mu_0(\theta)\pi(s|\theta)}{\Pr_\Pi(s)}w(\mu,\theta) \\ &= \sum_{\mu\in\text{supp}(\Pi)}\sum_{s\in S:\mu_s=\mu} \Pr_\Pi(s)\frac{\mu(\theta)}{\mu_0(\theta)}w(\mu,\theta) \\ &= \sum_{\mu\in\text{supp}(\Pi)}\sum_{s\in S:\mu_s=\mu} \Pr_\Pi(s)\hat{w}(\mu,\theta) = \mathbb{E}_\Pi\left[\hat{w}(\tilde{\mu},\theta)\right]. \end{aligned}$$

□

Proof of Theorem 1. By definition, the point $(\mu_0, \mathrm{w}) \in \mathrm{co}\,(\mathrm{graph}\,\hat{\mathrm{w}})$ if and only if a Bayes' plausible posterior distribution exists such that $\mathbb{E}[\hat{\mathrm{w}}\,(\mu)] = \mathrm{w}$. At the same time, an information structure can induce a distribution over beliefs if and only if $\mathbb{E}[\mu] = \mu_0$. By Equation 4, the result follows. □

A.2 Omitted proofs from Section 4

Proof of Proposition 2. If all types (strictly) benefit from disclosure, providing no information is not Pareto efficient, and by Theorem 2 persuasion is beneficial in the problem of any direction $\lambda \in \mathbb{R}^N_+$, in which case, $\mathrm{cav}\hat{v}_\lambda(\mu_0) > \hat{v}_\lambda(\mu_0)$, as shown by Kamenica and Gentzkow (2011). For the opposite direction, if the strict Pareto improvement is not possible, providing no information is Pareto efficient and by Theorem 2, a direction $\lambda \in \mathbb{R}^N_+$ must exist such that providing no information is optimal in the corresponding supporting Bayesian persuasion problem, in which case, $\mathrm{cav}\hat{v}_\lambda(\mu_0) = \hat{v}_\lambda(\mu_0)$. □

Proof of Corollary 4. The result follows from Proposition 2. For the first part, the condition ensures $\mu(\theta)w(\mu,\theta)$ is strictly convex for both types, so the indirect utility function $\hat{v}_\lambda$ is strictly convex for any $\lambda \in \Delta(\Theta)$ and full disclosure is the unique solution to the supporting Bayesian persuasion problem. For the second part, the condition ensures $\mu(\theta)w(\mu,\theta)$ is weakly concave for that type θ, and the result follows from looking at the direction $\lambda(\theta) = 1$ for that θ, and $\lambda(\theta) = 0$ elsewhere. □

Corollary 3 provides conditions for all types to benefit from persuasion or for no-disclosure to be Pareto efficient. Corollaries 6 and 7 below refine the conditions presented in Corollary 3:

Corollary 6 (Benefit from disclosure). *All types benefit from disclosure if $\hat{w}(\cdot, \theta)$ is strictly convex in a neighborhood of the prior for all $\theta \in \Theta$. Moreover, if $\hat{w}(\cdot, \theta)$ is everywhere strictly convex for all $\theta \in \Theta$, full information is uniquely Pareto efficient.*

Proof of Corollary 6. Because the indirect utility $\hat{v}_\lambda$ is a convex combination of the truth-adjusted welfare functions, the first condition in Corollary 6 ensures that for any $\lambda \in \Delta(\Theta)$, $\hat{v}_\lambda$ is strictly convex in a neighborhood of the prior. Hence, all types benefit from disclosure because for any $\lambda \in \Delta(\Theta)$, $\mathrm{cav}\hat{v}_\lambda(\mu_0) > \hat{v}_\lambda(\mu_0)$, and the result follows from Proposition 2. Similarly, when $\hat{w}(\cdot, \theta)$ is everywhere strictly convex for all types, the indirect utility function $\hat{v}_\lambda$ is strictly convex for any $\lambda \in \Delta(\Theta)$, so full information disclosure is the unique solution to the supporting Bayesian persuasion problem. □

Corollary 7 (No disclosure is Pareto efficient). *The no-disclosure profile w^{ND} is in the Pareto frontier if either of the following conditions is satisfied:*

(i) A vector $a \in \mathbb{R}^N_+$ and a concave function $w : \Delta(\Theta) \mapsto \mathbb{R}$ exist such that for all $\theta \in \Theta$, the ex post welfare function is given by $a(\theta)w(\mu) + b(\theta)$;

(ii) For some $\theta \in \Theta$, $\hat{w}(\mu, \theta)$ is concave in μ.

Proof of Corollary 7. The result follows from Proposition 2 by looking at the supporting Bayesian persuasion problems in the directions $\lambda(\theta) = \mu_0(\theta)/a(\theta)$ for (i), and $\lambda(\theta) = 1$ for θ at which $\hat{w}(\mu, \theta)$ is concave and $\lambda(\theta) = 0$ elsewhere for (ii). □

A.3 Omitted proofs in Section 5

Proof of Theorem 3. As explained in the main text, necessity follows from noting

$$\mathrm{w} \in \mathrm{W} \Rightarrow \mathrm{w} = \mathrm{D}_0 \underbrace{\sum_{m=1}^{M} \alpha_m \mu_m \mu_m^T}_{\mathrm{C}} \rho.$$

C is completely positive because it is the convex combination of rank-one non-negative matrices, $\mu_m \mu_m^T$. That $\mathrm{Ce} = \mu_0$ follows from the martingale property of beliefs.

For sufficiency, consider $\mathrm{w} = \mathrm{D}_0\mathrm{C}\rho$, for some completely positive matrix C, such that $\mathrm{Ce} = \mu_0$. Then, $\{\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{M}\} \subseteq \mathbb{R}^N_+$ exist such that

$$\mathrm{C} = \sum_{m=1}^{M} \mathrm{x_m} \mathrm{x_m}^T. \tag{16}$$

Let $\sqrt{\alpha_m} = \sum_{j=1}^{N} \mathrm{x}_{\mathrm{m}j}$ and note $\mathrm{x_m}/(\sqrt{\alpha_m}) \equiv \mu_m \in \Delta(\Theta)$.

$$\mathrm{C} = \sum_{m=1}^{M} \alpha_m \left(\frac{\mathrm{x_m}}{\sqrt{\alpha_m}}\right)\left(\frac{\mathrm{x_m}}{\sqrt{\alpha_m}}\right)^T = \sum_{m=1}^{M} \alpha_m \mu_m \mu_m^T.$$

It remains to show $\sum_{m=1}^{M} \alpha_m = 1$ and $\sum_{m=1}^{M} \alpha_m \mu_m = \mu_0$. Note that for all $i \in \{1, \ldots, N\}$,

$$(\mathrm{Ce})_i = \sum_{m=1}^{M} \alpha_m \mu_{mi} \sum_{j=1}^{N} \mu_{mj} = \sum_{m=1}^{M} \alpha_m \mu_{mi} = \mu_0(\theta_i). \tag{17}$$

Furthermore,

$$\sum_{i=1}^{N} \mu_0(\theta_i) = 1 = \sum_{i=1}^{N} \sum_{m=1}^{M} \alpha_m \mu_{mi} = \sum_{m=1}^{M} \alpha_m. \tag{18}$$

Thus, an information structure exists that generates the distribution of posteriors $\{\alpha_m, \mu_m\}_{m=1}^{M}$. Therefore, $\mathrm{w} \in \mathrm{W}$. □

Proof of Claim 2. If $\Pr(X) = 0$, the statement is trivial. If $\Pr(X) > 0$, denote by $\mathrm{P_{i\cdot}}$ the i-th row of the matrix $\mathrm{P} \equiv \mathrm{D}_0\mathrm{C}$, presented as a row-vector. By Bayes' rule, $\Pr(X) = \mu_0^T \beta$ and $\Pr(\theta_i \mid X) = (\mu_{0i}\beta_i)/(\mu_0^T\beta)$, so

$$\mathbb{E}_\Pi\left[\Pr(X \mid s) \mid \theta_i\right] = \sum_{j=1}^{N} \mathbb{E}_\Pi\left[\Pr\left[\theta_j \mid s\right] \mid \theta_i\right] \Pr(X \mid \theta_j) = \mathrm{P_{i\cdot}}\beta,$$

$$\mathbb{E}_\Pi\left[\Pr\left(X \mid s\right) \mid X\right] = \sum_{i=1}^{N} \Pr\left(\theta_i \mid X\right) \mathbb{E}_\Pi\left[\Pr\left(X \mid s\right) \mid \theta_i\right] = \sum_{i=1}^{N} \frac{\mu_{0i}\beta_i}{\mu_0^T\beta} \mathrm{P_{i\cdot}}\beta.$$

Hence, the truth-drifting condition can be restated as:

$$\sum_{i=1}^{N} \frac{\mu_{0i}\beta_i}{\mu_0^T\beta} \mathrm{P_{i\cdot}}\beta \geq \mu_0^T\beta.$$

Define $\hat{\mathrm{C}} \equiv \mathrm{PD}_0 = \mathrm{D}_0\mathrm{C}\mathrm{D}_0$. By Theorem 3, $\hat{\mathrm{C}}$ is a completely positive matrix such that $\hat{\mathrm{C}}\mu_0 = \mathrm{e}$ and $\mu_0^T\hat{\mathrm{C}}\mu_0 = 1$. Hence, the truth-telling condition can be restated in a matrix form as:

$$\left(\frac{\mu_0 * \beta}{\mu_0^T\beta}\right)^T \hat{\mathrm{C}} \left(\frac{\mu_0 * \beta}{\mu_0^T\beta}\right) \geq \mu_0^T\hat{\mathrm{C}}\mu_0.$$

The term $\zeta \equiv (\mu_0 * \beta)/(\mu_0^T\beta)$ is an element of simplex $\Delta(\Theta)$, equal to μ_0 when $\beta = \mathrm{e}$. Hence, showing that μ_0 is a minimizer of a quadratic form $\zeta^T\hat{\mathrm{C}}\zeta$ among all $\zeta \in \Delta(\Theta)$ is enough to prove the result. Noting that we can rely on the Lagrangian approach, at $\zeta = \mu_0$, the derivative of the quadratic form is collinear to e and hence, collinear to the space of $\Delta(\Theta)$. Thus, first-order conditions are satisfied. At the same time, $\hat{\mathrm{C}}$ is completely positive and thus positive semi-definite. Thus, second-order conditions are satisfied. The result follows. □

Example 4 (w^{ND} and w^{FD} not on the boundary of W.). Consider the case with binary types, $\Theta = \{\theta_1, \theta_2\}$. Denote by $\mu \in [0, 1]$ the probability of type θ_2 and let $\mu_0 = 1/2$. Consider the following welfare function[20]:

$$w(\mu, \theta_1) = \frac{\sin(2\pi\mu)}{2(1-\mu)},\ w(\mu, \theta_2) = \frac{\sin(4\pi\mu)}{2\mu},$$

with $w(1, \theta_1)$ and $w(0, \theta_2)$ defined by continuity as equal to $-\pi$ and 2π respectively. Given this welfare function, the truth-adjusted welfare function is

$$\hat{w}(\mu, \theta_1) = \sin(2\pi\mu),\ \hat{w}(\mu, \theta_2) = \sin(4\pi\mu).$$

The corresponding indirect utility in the supporting Bayesian persuasion problem in the direction $\lambda = (\lambda_1, \lambda_2)$ is equal to

$$\hat{v}_\lambda(\mu) = \lambda_1 \sin(2\pi\mu) + \lambda_2 \sin(4\pi\mu).$$

For any $\lambda \in \mathbb{R}^2 \setminus \{0\}$, $\hat{v}_\lambda(\mu_0) = \hat{v}_\lambda(1/2) = \hat{v}_\lambda(0) = \hat{v}_\lambda(1) = 0$. Hence, providing full disclosure or no disclosure results in zero payoff. At the same time, for any such λ, $\hat{v}_\lambda(\mu)$ is a non-constant continuous function anti-symmetric around $\mu = 1/2$. Hence, it achieves strictly positive values on $[0, 1]$ and $\mathrm{cav}\hat{v}_\lambda(\mu_0) > 0$ so that optimal disclosure outperforms both full disclosure and no disclosure. By Theorem 2–extended to all boundary points,– it follows that w^{ND} and w^{FD} are not on the boundary of the Bayes welfare set.

The proofs of Proposition 4 and Corollary 5 rely on the graph-theoretic approach in Rayo and Segal (2010), the main properties of which we summarize in Remark 5:

[20] Whereas an analogous argument may be applied to other welfare functions, we choose the specific functional forms to simplify exposition.

Remark 5 (Rayo and Segal, 2010). *Rayo and Segal (2010) propose the following graphical depiction of an information structure,* Π. *Given a direction* λ, *let the prospect values* $(\frac{\lambda(\theta_j)}{\mu_0(\theta_j)}, \rho(\theta_j)) = (\gamma_j, \rho_j)$ *for* $j = 1, \ldots, N$ *be vertices of a graph in* $\mathbb{R}^2$. *Connect the points* (γ_j, ρ_j) *and* (γ_k, ρ_k) *by an edge if and only if a signal* s *exists such that* $\pi(s \mid \theta_j) * \pi(s \mid \theta_k) > 0$. *The set of types that have positive probability under* s *is called the* pooling *set of signal* s.

Lemmas 2, 3, 4, and 5 in Rayo and Segal (2010) establish respectively that under any optimal information structure, the following hold:

(a) posterior expectations induced by any two signals are ranked (in vector order),

(b) prospects appear in the support of some signal only if they lie on a straight line with non-positive slope,

(c) if the pooling segments[21] *of two signals do not lie on the same line, they can intersect only if they share an endpoint, and*

(d) if two prospects values are ranked and appear in the support of two signals, then the posterior expectations induced by these signals are ranked in the same way.

Proof of Proposition 4. By the arguments presented in the main text, any optimal information structure solves the instance of the problem of Rayo and Segal (2010) in which the prospect values are $(0, \rho(\theta_j))$ for $j \neq i$ and $(1, \rho(\theta_i))$ for the sender and for the receiver, respectively.

Given the structure of the prospect values in our problem, the property in item (b) implies that θ_i is never pooled with lower-index types. Furthermore, whenever it is pooled with some type, it is pairwise pooled. The property in item (c) implies that whenever θ_i is pooled with some type θ_j, then no types θ_k, θ_l with $k < j < l$ can be pooled. Together with the property in item (d), this observation implies that whenever θ_i is pooled with some type θ_j, it is also pooled with all types θ_k with $k > l$. Moreover, as θ_i is pooled with increasingly higher-index types, the corresponding posterior expectations increase in vector order, which means that higher signals induce higher reputation yet have a relatively higher proportion of θ_i (if all types are equally likely, then the probability of pooling θ_i with θ_k increases in k).

It is left to show that the threshold type, the lowest type with which θ_i is pooled, is not pooled with any type of lower index. However, because the threshold type is of higher index than θ_i, such pooling could clearly be improved by pooling the threshold type exclusively with θ_i. □

[21] By item (b), the pooling set of a signal lies on a *segment*.

Calculations for Example 3. Define the following parameterized family of information structures (rows correspond to types and columns to signals):

$$\Pi_1(\alpha,\beta)=\begin{pmatrix}\alpha & 1-\alpha & 0\\ 1-\beta & \beta & 0\\ 0 & 0 & 1\end{pmatrix},\quad \Pi_2(\alpha)=\begin{pmatrix}\alpha & 1-\alpha\\ 1 & 0\\ 0 & 1\end{pmatrix},$$
$$\Pi_3(\beta)=\begin{pmatrix}1 & 0\\ 0 & 1\\ \beta & 1-\beta\end{pmatrix},\quad \Pi_4(\alpha,\beta)=\begin{pmatrix}1 & 0 & 0\\ 0 & \alpha & 1-\alpha\\ 0 & 1-\beta & \beta\end{pmatrix}.$$

Note information structures $\Pi_1(1,1)$ and $\Pi_4(1,1)$ coincide and correspond to full disclosure. Likewise, information structures $\Pi_2(0)$ and $\Pi_3(1)$ both correspond to full pooling of types θ_1 and θ_3.

By Proposition 3, any solution to the supporting Bayesian persuasion problem solves the instance of the problem of Rayo and Segal (2010) with prospect values $\{(\gamma(\theta_i),\rho(\theta_i)) : i\in\{1,2,3\}\}$ for the sender and for the receiver, respectively. For simplicity, we assume that the types are strictly ranked under ρ, i.e., $\rho(\theta_1)<\rho(\theta_2)<\rho(\theta_3)$.

The property in item (b) in Remark 5 implies that if $(\gamma_i,\rho(\theta_i))<(\gamma_j,\rho(\theta_j))$, then types θ_i and θ_j are never pooled (cf. Corollary 5). We can then immediately establish the properties of the boundary information structures in the following cases:

- If $\gamma_1<\gamma_2<\gamma_3$, then the uniquely optimal information structure is full disclosure.
- If $\gamma_2<\gamma_1<\gamma_3$, then any optimal information structure separates type θ_3 and belongs to class $\Pi_1(\alpha,\beta)$.
- If $\gamma_1<\gamma_3<\gamma_2$, then any optimal information structure separates type θ_1 and belongs to class $\Pi_4(\alpha,\beta)$.
- If $\gamma_2<\gamma_3<\gamma_1$, then an optimal information structure never pools types θ_2 and θ_3 and belongs to class $\Pi_2(\alpha)$.
- If $\gamma_3<\gamma_1<\gamma_2$, then an optimal information structure never pools types θ_1 and θ_2 and belongs to class $\Pi_3(\beta)$.

In the remaining case $\gamma_3<\gamma_2<\gamma_1$, no two prospects are ranked. However, by the property in item (a) in Remark 5, the induced posterior expectations are necessarily ranked. Hence, if all three prospects lie on a straight line, then no disclosure is optimal. In contrast, if the three prospects don't lie on a straight line, then an optimal information structure separates either types θ_1 and θ_2 or types θ_2 and θ_3, and thus belongs to either class $\Pi_2(\alpha)$ or to class $\Pi_3(\beta)$.

Finally, it is easy to see that by the same arguments, an optimal information policy

for the cases in which $\gamma_i = \gamma_j$ for some i and j belongs to one of the same four classes of information structures.

Knowing the classes of boundary information structures, we can plot the Bayes welfare set in Example 3 by direct calculation.

□

A.4 Omitted proofs in Section 6

Derivations for Section 6. The welfare of individuals of type θ under information structure, Π is given by

$$
\begin{aligned}
w_\Pi(\theta) \equiv \mathbb{E}_{\Pi|\theta}[w_\dagger(\tilde{\eta},\theta)] &= \sum_{\eta\in\text{supp}(\Pi)} \sum_{s\in S:\eta_s=\eta} \sum_{d\in D} \nu_0(d|\theta)\pi(s|d) w_\dagger(\eta,\theta) \qquad (19)\\
&= \sum_{\eta\in\text{supp}(\Pi)} \sum_{s\in S:\eta_s=\eta} \sum_{d\in D} \nu_0(d|\theta) \frac{\sum_{d'\in D}\eta_0(d')\pi(s|d')}{\eta_0(d)} \frac{\eta_0(d)\pi(s|d)}{\sum_{d'\in D}\eta_0(d')\pi(s|d')} w_\dagger(\eta,\theta)\\
&= \sum_{\eta\in\text{supp}(\Pi)} \sum_{s\in S:\eta_s=\eta} \Pr_\Pi(s) \sum_{d\in D} \nu_0(d|\theta)\frac{\eta_s(d)}{\eta_0(d)} w_\dagger(\eta,\theta)\\
&= \sum_{\eta\in\text{supp}(\Pi)} \sum_{s\in S:\eta_s=\eta} \Pr_\Pi(s)\hat{w}_\dagger(\eta,\theta).
\end{aligned}
$$

□

Proof of Proposition 5. The proof of the first part is presented in the main text. We now formally prove the second part. Toward a contradiction, assume (D',ν_0') is not a garbling of (D,ν_0). Then, by Blackwell (1953), a $\mu_0 \in \Delta(\Theta)$ and payoff function $u : A\times\Theta\to\mathbb{R}$ exist such that an agent with utility u derives strictly greater value from having access to (D',ν_0') than to (D,ν_0). That is, letting $U(\mu)$ denote the agent's indirect utility, $\max_{a\in A}\mathbb{E}_\mu[u(a,\theta)]$, we have that:

$$
\sum_{\theta\in\Theta}\mu_0(\theta)\mathbb{E}_{(D',\nu_0')}[U(\mu)\mid\theta] > \sum_{\theta\in\Theta}\mu_0(\theta)\mathbb{E}_{(D,\nu_0)}[U(\mu)\mid\theta]. \qquad (20)
$$

Consider now the Bayes welfare sets given the welfare function $w(\mu,\theta)=U(\mu)$ and prior distribution μ_0, under data sources (D,ν_0) and (D',ν_0'). We have that

$$
\begin{aligned}
\sum_{\theta\in\Theta}\mu_0(\theta)\mathbb{E}_{(D',\nu_0')}[U(\mu)\mid\theta] &= \max_{\mathrm{w}\in\mathrm{W}(\cdot,D',\nu_0')}\sum_{\theta\in\Theta}\mu_0(\theta)\mathrm{w}(\theta) \qquad (21)\\
&\le \max_{\mathrm{w}\in\mathrm{W}(\cdot,D,\nu_0)}\sum_{\theta\in\Theta}\mu_0(\theta)\mathrm{w}(\theta) = \sum_{\theta\in\Theta}\mu_0(\theta)\mathbb{E}_{(D,\nu_0)}[U(\mu)\mid\theta],
\end{aligned}
$$

where the equalities follow because the maximal ex ante payoff is obtained by having full access to available data, and the inequality follows from the assumption that $\mathrm{W}(\mu_0, w, D', \nu_0') \subseteq \mathrm{W}(\mu_0, w, D, \nu_0)$ for all w and μ_0.

The comparison of Equations 20 and 21 leads to the desired contradiction and the result follows. □