MEDIATED PERSUASION: OVERCOMING THE LIMITS OF BAYESIAN PERSUASION

Andrew Kosenko*

October 2025

Abstract

I study a game of information design between a sender who chooses state-dependent information structures, a mediator who garbles the signals generated from these structures, and a receiver who acts after observing the signal generated by the first two players. I provide sufficient conditions under which the receiver benefits from mediation: when the mediator prefers more information revelation than the sender, but less than full revelation. I also propose a novel algorithm for computing the set of feasible posterior beliefs that the sender can induce, and illustrate a new link to the Blackwell order.

JEL Classification: D82, D83, C72.

Keywords: persuasion, communication, strategic information transmission, noisy communication, Blackwell informativeness, garbling, information provision.

---

*Associate Professor of Economics, Department of Economics and Finance, School of Management, Marist University. 3399 North Road, Poughkeepsie NY, 12601. kosenko.andrew@gmail.com.
I am deeply grateful to Navin Kartik, for his invaluable help and advice. I would also like to thank Yeon-Koo Che, Alessandro Pavan, and Joseph Stiglitz for guidance and comments, as well as James Best, Ambuj Dewan, Laura Doval, Nate Neligh, Anh Nguyen, Pietro Ortoleva, Daniel Rappoport, Luca Rigotti, Teck Yong Tan, Roee Teper, Richard van Weelden, participants of the Columbia Microeconomic Theory Colloquium, the 2019 international Conference on Game Theory at Stony Brook University, and the Pitt Micro Theory Brown Bag for discussions and input. Any remaining errors are my own. First draft: October 2017.

1 Introduction

How does the presence of a mediator affect the informational interaction between two parties? In a game of persuasion between one side (a *sender*) that is trying to persuade another side (a *receiver*) to take a certain action, I add a *mediator* who is able to alter the recommendation of the sender, before the receiver takes her action. A suprising result is that although the mediator can only destroy information, preference conflict may nonetheless result in mediated persuasion being more informative than unmediated persuasion, and therefore, better for the receiver.

In the main theorem, I show that the presence of a mediator who can only destroy information may be beneficial to the receiver; example 1 elucidates this new mechanism for increasing information provision in strategic settings. This is a counterintuitive finding; how can it be that introducing an additional player (with her own preferences), and an ability to only destroy information benefits the receiver in a very strong sense (Blackwell dominance)? The intuition is this: because the mediator can only destroy information, in equilibrium, the sender will take that into account; meaning, she will provide *more* information to compensate for the presence of the mediator. There results, in a sense, a "tug of war" in information - the sender is providing more information (relative to an environment without a mediator), while the mediator is destroying information. One contribution of this work is to show explicitly that there are preferences for the sender and the mediator, for which this "tug of war" results in an increase in overall information revelation, and therefore, strictly benefits the receiver. This provides a new mechanism for information revelation: if the receiver is unable to do anything else (such changing the preferences of the sender, introducing multiple senders with conflict of interest, or obtaining information herself), she may be able to improve her lot by introducing a mediator that can destroy information, provided that mediator has freedom of action, and has different preferences from the receiver.

This work also provides a framework for understanding when and why informational mediation is beneficial to the receiver.[1] It also provides a foundation for a kind of "ante-persuasion" consideration – if, before playing any game, the receiver were to be able to choose whether to play a game with or without a mediator, what would she choose? It turns out that the receiver benefits from mediation if the mediator prefers more information revelation than the sender, but *less* than perfect revelation (which is the preferred outcome of the receiver). Thus, there is yet another surprising twist - it is a mediator with preferences that are sufficiently *different* from the receiver that can benefit the receiver. If the preferences of the mediator and the receiver coincide, there is no increase in information revelation. Put somewhat differently, the receiver should never want to mediate for herself; she should rather delegate mediation to another party with different preferences.

Notably, while most papers in this literature focus on the sender's most preferred equilibria, taking the view that the sender can "steer" the receiver into the appropriate equilibrium, I also consider the welfare of the receiver across different equilibria, taking the view that in most applications, it is the *receiver's* welfare that one ultimately cares about, because it is the receiver who takes the socially relevant action. This work adds an institutional aspect to the information design research program, investigating the effects of different informational-organizational topologies on information revelation and welfare.

The sender and the receiver are restricted to communicate indirectly, via the mediator (perhaps more than one), due to technical or institutional constraints. For example, when a financial firm issues certain kinds of financial products, some large (institutional) investors are prohibited from purchasing them, unless they have been rated by a third party, and have achieved a certain rating. Similarly, in many organizations (including many firms, the military, and the intelligence community) the flow of information is directed, with the direction exogenously predetermined, with various agents having the ability to alter (or perhaps not pass

[1] Obviously, it is never strictly beneficial to the sender.

on) the information passed up to them. The introduction of the mediator, and this specific organization, are a stylized representation of real-world examples of information flows.

This is a difficult problem because one strategic player (the sender) can both create and destroy information, and the other strategic player (the mediator) can only destroy information that the first player provided, but cannot create any new information; in addition, both of these players only have probabilistic control over evidence realization, and of course there is the question of which receiver beliefs can be induced. One of the difficulties is that when the mediator changes her action, not only does the sender's best response change, but in effect, the *choice set* of the sender, understood as the set of receiver posterior beliefs that can be induced, changes as well. To address this difficulty I introduce a novel way of explicitly the feasible sets for the mediator and the sender as functions of the sender's and the mediator's actions, respectively. Given the choice of the sender, the mediator can deviate to anything less informative in the Blackwell sense, while given the choice of the mediator the sender can deviate to *some* things that are less Blackwell informative (than the implied final experiment), but not everything. This procedure, and the resulting computations, appear in the appendix.

This is also what makes this work different. The major thrust of the literature on multi-sender Bayesian persuasion has focused on players only being able to add (in a certain sense) information.[2] On the other hand, the contemporaneous work on persuasion with noise (where some information is exogenously destroyed) has studied nonstrategic settings. I consider an environment where some players can add information, some can subtract information, and in addition, I study a game,

[2]For example, in the Gentkow and Kamenica (2017a) work on this problem they identify a condition - Blackwell-connectedness - which ensures that full revelation is always an equilibrium outcome in their multisender game. The condition says that given the information provided by the others, any individual player can always deviate to something *more* informative. Li and Norman (2021) similarly assume that each sequential player has access to signals that are arbitrarily correlated, so that a player can improve upon the information provided by others realization by realization. In both cases any one player can unilaterally increase the amount of information provided.

not a decision problem. Furthermore, I compare outcomes of the game along two dimensions; first I vary the preferences of the sender and the mediator. The most prominent result is that (perhaps unsurprisingly) preference divergence "quickly" leads to the only equilibrium being uninformative. The final object of this exercise is to compare outcomes (for fixed preferences), in terms of information revelation and welfare, between standard Bayesian persuasion, and Bayesian persuasion with an informational mediator. I show that although the mediator can only destroy information (in an appropriate sense), the presence of the mediator can still result in a strict increase in the amount of information revealed in the Blackwell sense.

The paper is organized as follows: section 2 provides a leading motivating example and discussion. Section 4 presents an explicit example where mediated persuasion benefits the receiver; the example illustrates several features critical for the general characterizations that follow, and are generalized in the main theorems of Section 5. Section 3 presents the model, the literature is reviewed in section 6, and section 7 briefly concludes. Appendices A and B illustrate additional features of the computations used throughout the paper, and their connection to the Blackwell order over information structures. All proofs appear in appendix C.

2 The Function of a Ratings Agency

The focus will be on the amount of information revealed in various organizational setups and the effect of competition and preference (mis)alignment on information revelation and outcomes. Although the basic model is quite general, I have in mind one particular application - the function of a ratings agency. A rating assigned to a financial product can be thought of as an expression of likelihood of default or expected economic loss. A firm (in the parlance of the present setting, the sender) chooses strategically what evidence to submit to a rater (here, the mediator). The mediator, perhaps driven by concerns that may not be identical to those of the firm, then rates the evidence submitted by the firm, and issues a recommendation to the client or the public. I analyze the effect on informativeness and welfare of the mediator's presence in this informational-organizational topology.

There are several features of this real-world example that deserve mention. First note that the issuing firm itself cannot rate its own financial products; it does, however, *design* its products (or at least gets to choose the products that it submits for review at a particular instance). The ratings agency cannot choose the products - it is constrained to rate the package it has been submitted, but it can choose its ratings process and criteria. It also exhibits the criteria according to which it issued its conclusions. Finally, the purchaser of the financial products (the receiver) is often required to only buy products that have been rated by a reputable firm - in other words, there is an *institutional* constraint at work.

To take a specific example consider structured finance products that consisted of various repackagings of individual loans (mortgages were by far the most important component) into so-called structured investment vehicles, or SIVs. The financial firms issued products that consisted of bundles of individual mortgages, along with rules for obtaining streams of payments from those mortgages. These streams were correlated with each other (since two nearby houses were in the same area, the local economic conditions that affected the ability of one lender to repay, also affected the ability of the other lender to repay), as well as with the overall economy. The firms chose the specific mortgages that went into each SIV strategically. The ratings agencies then rated these SIVs; however, one key element in their ratings (and one that was later shown to be partially responsible for the revealed inaccuracy of those ratings) is that the ratings agencies did not provide their ratings based on the correlations of the returns with the overall economy. Rather, their ratings consisted (mostly) of evaluations of correlations of individual financial products in an SIV with each other. The issuer clearly wants to achieve as high a rating as possible[3], but if the preferences of the mediator are to "collude" with the seller, this essentially means that there may be very little information revelation in equilibrium.

In this example the state of the world is a complete, fully specified joint distribution of returns; an experiment is a mapping from states of the world into a set that specifies only partial information about the correlations (for example, individual correlations).[4] The mediator then designs a signal (a rating procedure) that maps information about individual correlations into a scaled rating. The precise ratings methodologies are proprietary, so it makes sense to assume that the sender does not know the strategy of the mediator. This example, although it is meant to be suggestive, is not completely analogous to the situation I study. I view the model presented in this paper as a normative exercise, descriptive of interesting features of a problem, but not identical to actual ratings process.

In single-issuer bonds, ratings are mute about correlations with other bonds or with the market. In 2007, less than 1% of corporate issues but 60% of all structured products were rated AAA. 27 of 30 AAA issues underwritten by Merril Lynch in 2007, were by 2008 rated as speculative ("junk") (See Coval, Jurek and Stafford (2008)). I suggest that a possible explanation for this is that if the mediator is unable to provide new information, and is only able to "garble" or rely on the information provided to it by the issuer, then the equilibria in general will not be very informative (and in fact, as the preferences of the mediator and the sender diverge, the only equilibrium that survives is uninformative). This reasoning suggests a policy proposal - requiring the ratings agencies to perform independent analysis (say, additional "stress tests") on the products they are rating, to increase the informativeness of the rating; another way of increasing information revelation is to ensure that the preferences of the mediator are what they are prescribed to be by

[3] And in fact, there is evidence in structured finance that the firms did design their products so that the senior tranches would be as large as possible, while still getting the highest possible rating.

[4] The "big three" firms all utilize fairly coarse scales for ratings.

this work.

3 Model

I study a game with $n \geq 3$ players; the first player is called the *sender* and the last player is called the *receiver*. The remaining players are the *mediators*; if there are more than one of them, I also specify the order in which their probabilistic strategies are executed.

Let Ω be the state space, and let E and S be generic signal realization spaces. A Blackwell experiment is $\mathcal{B} = (B, \mu)$ where B is a signal realization space and β is a probability measure defined on B. The (usual) interpretation is that when the state is $\omega \in \Omega$, the distribution of signal realizations in B is $\beta(\omega)$. To keep matters simple, in the initial analysis I do not allow players to choose the signal realization spaces as part of the Blackwell experiments.

Fix a finite state space Ω (consisting of n_Ω elements), and a finite realization space[5] E (consisting of n_E elements), where to avoid unnecessary trivialities, the cardinality of the set of signals is weakly greater than that of the set of states. An *experiment* for the sender is a distribution over the set E, for each state of the world: $X : \Omega \to \Delta(E)$; denote by $\boldsymbol{X}$ the set of available experiments. I assume that X contains both the uninformative experiment (one where the probabilities of all experiment realizations are independent of the state) and the fully revealing experiment (where each state is revealed with probability one). To distinguish between the choices of the sender and those of the mediator, I define a *signal* for the mediator to be a function $\Sigma : E \to \Delta(S)$ where S is the space of signal realizations containing n_S elements; let $\boldsymbol{\Sigma}$ denote the set of available experiments. Put differently, the mediator is choosing distributions of signal realizations conditional on realizations of experiments.

[5]Typically, the realization space is part of the choice of the sender; here I fix this space (while keeping it "rich enough") to isolate the effects of mediated persuasion.

All experiments and signals are available, and all have the same zero cost. In the present work I focus exclusively on pure strategies for all players. I also refer to either an experiment, or a signal, or their product, generically as an *information structure*. Since the state space and all realization spaces are finite, I represent information structures as column-stochastic matrices with the (i, j)'th entry being the probability of realization i conditional on j. Finally, the receiver takes an action from a finite set A (with n_A elements; I assume that $n_A \geq n_S = n_E \geq n_\Omega$ to avoid trivialities associated with signal and action spaces not being "rich" enough). The utility of the sender is denoted by $\tilde{u}^S(\omega, a)$, that of the mediator by $\tilde{u}^M(\omega, a)$ and that of the receiver by $\tilde{u}^R(\omega, a)$. I assume for concreteness that if the receiver is indifferent between two or more actions given some belief, she takes the action that is best for the sender.

This setup is capturing one of the key features of the model - the space of realizations of experiments for one player is the state space for the other player. In other words, both the sender and the mediator are choosing standard Blackwell experiments, but with different state and realization spaces.

For clarity: I use the convention that capital Greek letters (X, Σ) refer to the distributions, bold capital Greek letters ($\boldsymbol{X}, \boldsymbol{\Sigma}$) refer to sets of distributions, capital English letters (E, S) refer to spaces of realizations for information structures, and small English letters (e, s) refer to particular realizations.

The timing of the game is simple: the sender and the mediator choose their actions *simultaneously*, while the receiver observes the choices of the experiment, the signal, and the signal realization, but not the experiment realization. The mediator does not observe the choice of the sender when choosing her own action; if she did observe the choice (but not the experiment realization), this would be a special case of the model of sequential persuasion of Li and Norman (2021).[6] Note that

[6]If the mediator in addition could observe the experiment realization (and could therefore condition her own action upon it), this would be similar to the models of persuasion with private information by Hedlund (2017) and Kosenko (2022) since then the mediator would have an infor-

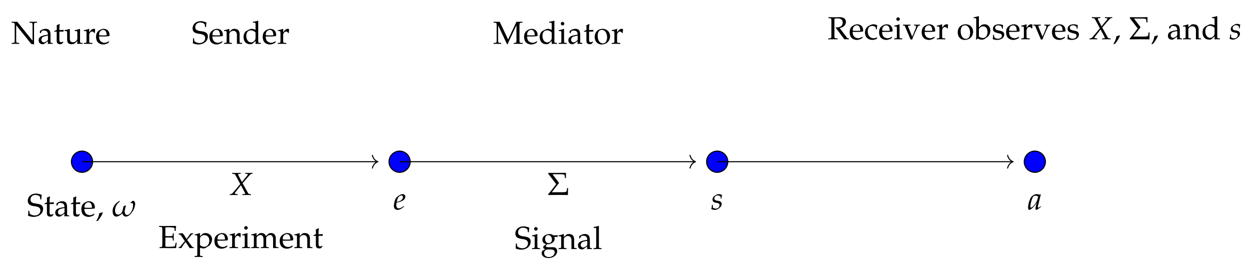

Figure 1: Illustration of the Model: Flow of Information

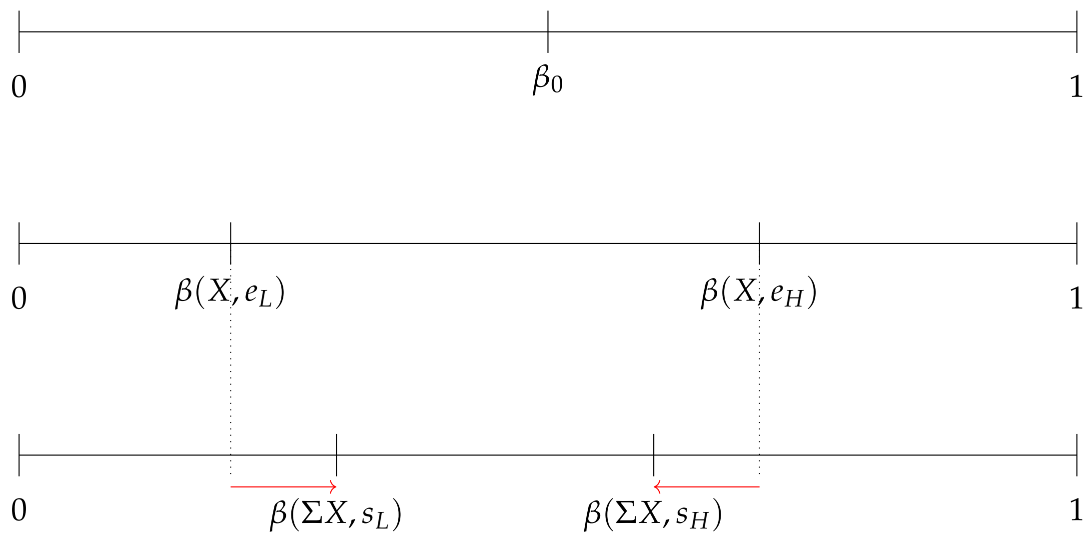

Figure 2: Effect of Garbling on Beliefs in a Dichotomy

no player observes the realization of the experiment, yet that realization clearly still plays a role in determining outcomes. I focus on pure strategies for all players in the present work; a diagram of the main features, nomenclature, timing, and notational conventions of the model is in figure 1.

The following definition will be useful in what follows:

Definition 1. *Let f and g be two probability mass functions on a finite set $X = \{x_1, x_2, ..., x_k\} \in \mathbb{R}^n$. f is a mean-preserving spread of g if there is a (k × k) Markov matrix $T_{K \times K} = (t(x_i|x_J))_{ij})$ such that*

1. *Tg=f.*

mational "type".

2. *For each* $j = 1, ...k$, $\sum_i T(x_i|x_j)x_i = x_j$

We can also illustrate the effect of a garbling of the experiment by the signal on the beliefs (as seen in figure 2). In that figure all players start with a common prior, β_0. When the sender chooses her experiment X, the two possible beliefs (one for each possible realization of the experiment) are a mean-preserving spread of the prior. Following that mediator's choice of signal, M brings beliefs back in in a mean-preserving contraction. In other words, in terms of figure 2, the mediator chooses the length (but not the location) of the two arrows, and the sender chooses the outer endpoint for each arrow. The inner point of each arrow represents the final beliefs.

Denote by $\beta_A(\omega|s)$ the posterior belief of the receiver that the state of the world is ω, computed after observing information structure A, and a signal realization s and denote by $\beta_A(s)$ the full distribution. It is also convenient to refer to distributions of distributions, which I will denote by τ so that $\tau_A(\beta)$ is the expected distribution of posterior beliefs given some generic information structure A:

$$\tau_A(\beta) \triangleq \sum_{\{s \in \text{supp}(A) | \beta_A(s) = \beta\}} \sum_{\omega \in \Omega} A(s|\omega)\beta_0(\omega) \tag{1}$$

Given a receiver posterior belief (suppressing the arguments for notational compactness) β, let $a^*(\beta)$ denote the optimal action of the receiver. Analogously to KG, if two actions for a sender or a mediator result in the same final belief for the receiver, they are equivalent. The number of arguments in the utility functions can be reduced: write $v^R(\beta)$, $v^M(\beta)$, $v^S(\beta)$ (with $u^i(\beta) \triangleq \mathbb{E}_\beta u^i(a^*(\beta), \omega)$, as is customary), and also, $V^i(\tau) \triangleq \mathbb{E}_{\beta\sim\tau}\Big(v^i(a^*(\beta), \omega)\Big)$.

We can begin by observing that an equilibrium exists, and in particular, there is an equilibrium analogous to the "babbling" equilibria of cheap talk models. If the sender chooses a completely uninformative experiment. Then the mediator is indifferent between all possible signals, since given the sender's choice, they cannot affect the action of the receiver; in particular she can choose the uninformative signal as well. Clearly, no player can profitably deviate, given the other's choices, and thus this is an equilibrium, which I note in the following Similarly, if either u^S or u^M (or both) is globally strictly concave over the set of $\beta \in \Delta(\Omega)$. Then the unique equilibrium is uninformative.

Given any X, the mediator's problem is now similar to the one faced by the sender in KG: choose a Σ such that the distribution of beliefs induced by B is optimal:

$$\Sigma^* \in \arg\max_{\{\Sigma \in \boldsymbol{\Sigma}\}} \mathbb{E}_\tau u^M(\beta) \tag{2}$$

$$\tau = p(B) \tag{3}$$

$$\text{s.t.} \sum_{s \in \text{supp}(\Sigma)} \beta^R(s)\mathbb{P}_B(s) = \beta_0 \tag{4}$$

Similarly, for the sender the problem is

$$X^* \in \arg\max_{\{X \in \boldsymbol{X}\}} \mathbb{E}_\tau u^S(\beta) \tag{5}$$

$$\tau = p(B) \tag{6}$$

$$\text{s.t.} \sum_{s \in \text{supp}(\Sigma)} \beta^R(s)\mathbb{P}_B(s) = \beta_0 \tag{7}$$

Let $p : \mathcal{M}_{n_S,n_\Omega} \to \Delta(\Delta(\Omega))$ where $\mathcal{M}_{n_S,n_\Omega}([0,1])$ denotes the set of $n_S \times n_\Omega$ column-stochastic matrices be the mapping between an information structure and the space of posterior beliefs. In other words, p maps a column-stochastic matrix into a distribution over posteriors: $p(B) = \tau$.

The solution concept is (weak) perfect Bayesian equilibrium. Call a pair of experiments (X, Σ), and a belief system that solve the above problems simply an equilibrium.

Notice that the matrix equation $\Sigma X = B$ is precisely the definition for X to be more Blackwell-informative than B, with Σ being the garbling matrix. I will rely on this fact (as well as the different and related implications of this fact) throughout what is to follow.

Given a particular choice of X by the sender, the mediator effectively chooses from a set of information structures that are Blackwell-dominated by the experiment. The set of feasible beliefs for the mediator, given a particular choice of the sender is illustrated in figure 3. This set is effectively a proportional "shrinking" of the Bayes-plausible set,[7] since all garblings of X are available to the mediator; the only constraint is that the mediator is not able to induce something more informative (by assumption) than the sender's choice. Foreshadowing the discussion to come, I note that the *sender's* feasible set, given a mediator action Σ is not going to be a simple "shrinking", and will involve other nontrivial constraints.

4 Mediated Persuasion Benefits the Receiver: Examples

Here we show by example that a mediated persuasion setting can be strictly better for the receiver than the Bayesian persuasion one.

4.1 Example 1

Consider a sender and a mediator with preferences that are illustrated in figure 4; in this figure the sender's utility is in red and that of the mediator is in blue. The sender's utility vanishes for beliefs below 0.2, then jumps up at 0.2, jumps back down to a value of $-k$, for k positive and "large", for beliefs $\beta \in (0.2, 0.955)$, except for another jump up at 0.5 then jumps up at 0.955, and then returns to

[7] The language of "guilty" and "innocent" to align with the language of KG.

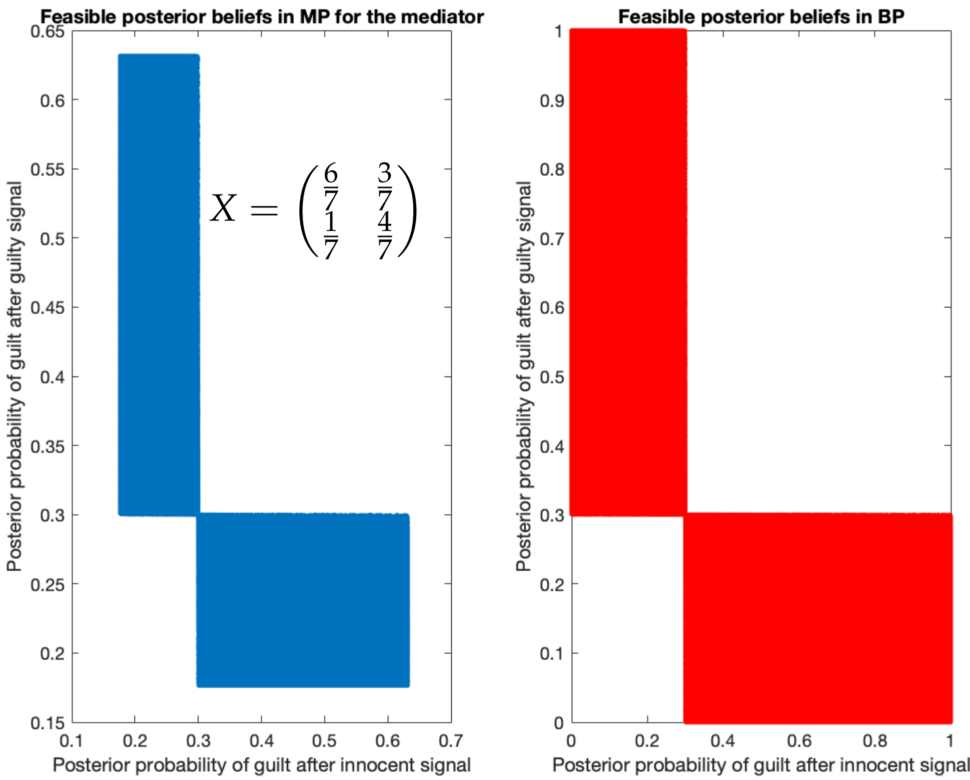

Figure 3: Feasible set for the mediator

0. The mediator's utility is M-shaped and peaks at 0.17 and 0.955. The common prior is β_0; without a mediator the sender would clearly choose the posteriors $\{\beta_1^{BP}, \beta_2^{BP}\} = \{0.2, 0.5\}$. They are certainly Bayes-plausible, and give the sender her highest possible utility. Suppose however that the mediator chose to play the following garbling:

$$\Sigma = \begin{pmatrix} \frac{1}{100} & \frac{1}{2} \\ \frac{99}{100} & \frac{1}{2} \end{pmatrix} \tag{8}$$

The $F(\Sigma, 0.3)$ set for this garbling is depicted in blue in figure 5. If the sender were to simultaneously play a fully revealing experiment, the outcome would be $\{\beta_1^{MP}, \beta_2^{MP}\} = \{0.17, 0.955\}$, yielding her a payoff of 0; note also that this is the most preferred outcome of the mediator. Given this garbling, the only way in

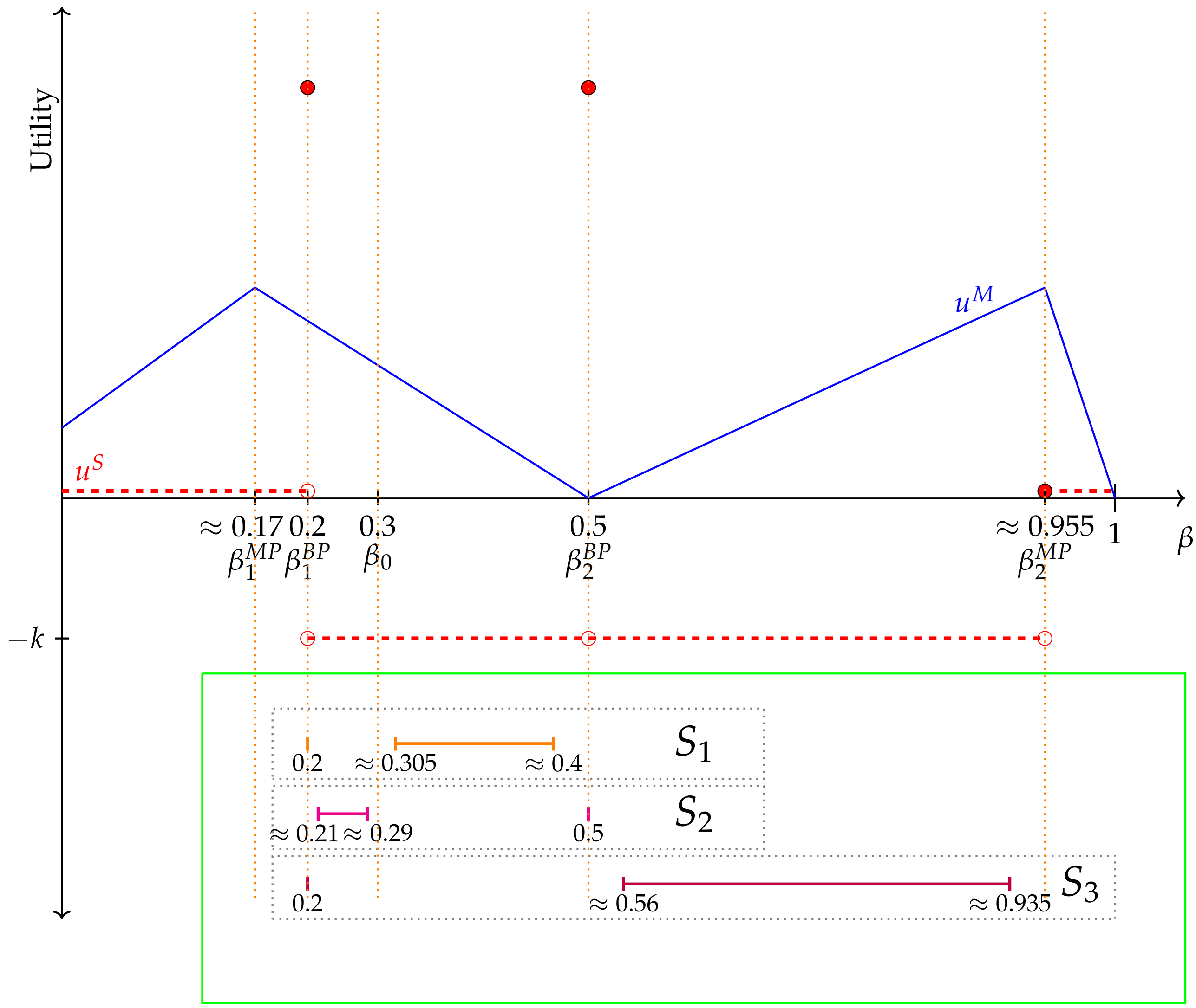

Figure 4: An MP Equilibrium That is Strictly More informative Than the BP Equilibrium. The utility of the mediator is in blue, the utility of the sender is in red.

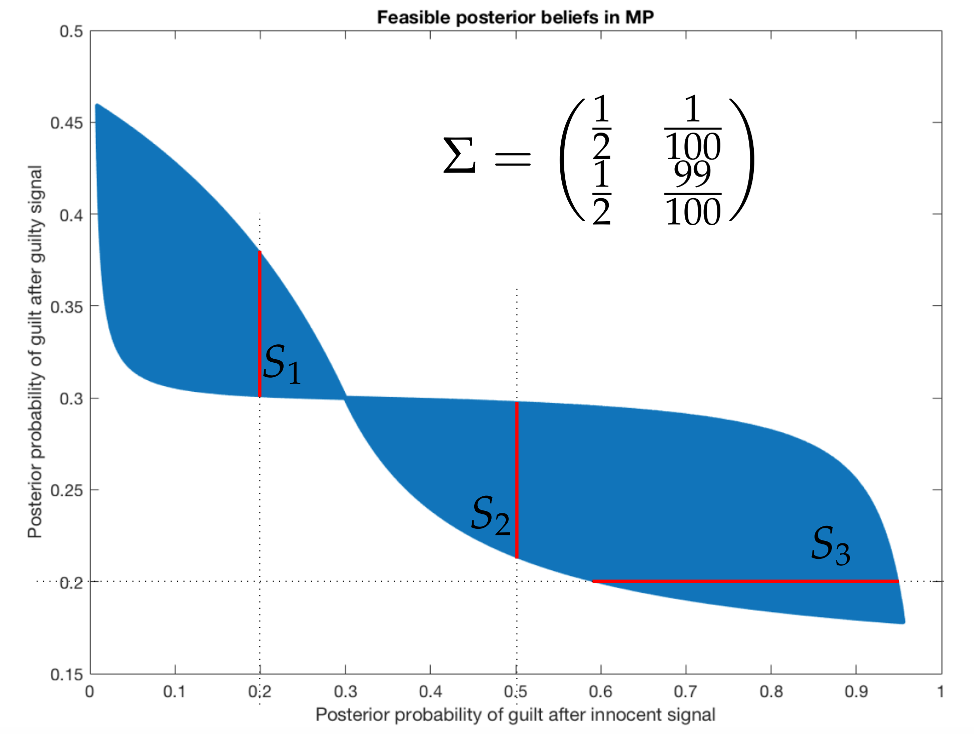

Figure 5: Feasible set $F(\Sigma, 0.3)$

which the sender can improve her payoff is by deviating to something that induces a posterior of 0.2, or 0.5. Suppose she deviates to something that results in one posterior (say, the first one) begin $\beta_1^{deviation} = 0.2$. Then the second posterior must lie in the S_1 set also illustrated in figure 5; given this constraint, the sender would get a payoff of 1 (with some probability) when the posterior is 0.2, and the payoff of $-k$ upon the other signal realization. For k large enough this will be negative for all beliefs in S_1, and thus this cannot be a profitable deviation. A similar logic applies to beliefs in S_2 and S_3 - inducing one belief that makes the sender better off also necessitates inducing another belief which makes her worse off, and the deviation is unprofitable. (The S sets are also illustrated in the green box in figure 4; it is helpful to see them on the same figure as the utility to understand the logic.) Obviously the mediator would not choose to deviate either, since she gets her first best outcome. Thus, the sender is "forced" to provide more information she would otherwise

$$(X, \Sigma) = \left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} \frac{1}{100} & \frac{1}{2} \\ \frac{99}{100} & \frac{1}{2} \end{pmatrix} \right) \tag{9}$$

is an equilibrium in which the outcome supported on $\{\beta_1^{MP}, \beta_2^{MP}\} \approx \{0.17, 0.955\}$ is strictly more Blackwell-informative than the BP outcome $\{\beta_1^{BP}, \beta_2^{BP}\} = \{0.2, 0.5\}$.

This construction explicitly shows that there are examples where the outcome of a game where a player can only decrease the amount of information turns out to be more informative in a very strong sense.

Say that *the receiver benefits from mediation* or that *mediation is beneficial* if the receiver's utility in the mediated persuasion game is strictly greater than her utility in the Bayesian persuasion game, where of course I compare games where the preferences of the sender and the receiver do not change between games.

We make a surprising, but important observation: when the preferences of the mediator are perfectly aligned with those of the sender (or alternatively, the receiver simply *is* the mediator), the receiver cannot benefit from mediation. The import is that the receiver benefits from mediation when the mediator's preferences are such that the mediator prefers more information revelation, in the sense of Blackwell, than the sender, but less information revelation than the receiver (which in a canonical environment, is of course, full revelation).

In other words, if $u^M = u^R$, the receiver cannot strictly benefit since if that were the case, the mediator would have a profitable deviation. And that profitable deviation - no garbling at all - cannot be part of an equilibrium where $\tau^* \succ \tau^{BP}$ since then the sender would have a profitable deviation, bringing beliefs back to the BP outcome, which must now be feasible, since the mediator is not garbling the information at all, and thus all Bayes-plausible beliefs are feasible. Thus, if the sender were the mediator, there can exist a garbling that makes her better off, but that cannot be an equilibrium, since there is also always a better one still until I get to no garbling at all, which again cannot be part of an equilibrium. In fact, remarkably enough, it is easy to see that the outcome when $u^M = u^R$ is either uninformative (i.e. $\beta_1^* = \beta_2^* = \beta_0$), or it coincides with the BP outcome. In other words, in the dichotomy setting, the nontrivial outcome when $u^M = u^S$ or when $u^M = u^R$ is exactly the same.

There are two important points: 1) the receiver can be better off in a very strong sense with a mediator, even though the mediator can only destroy information, and 2) for this to happen the preferences of the mediator cannot be the same as the preferences of the receiver (or the sender, for that matter). It must be the case that the mediator prefers more information revelation than the sender, but not perfect revelation (which of course is the preferred outcome of the receiver by assumption).

4.2 Example 2

I now show by example that even in this simple and canonical setting the MP outcome can be strictly more informative than the BP outcome. Suppose that the common prior is $\beta_0 = \frac{1}{2}$, that $A = \{a_1, a_2, a_3\}$, and that the optimal strategy of the receiver is to take action a_1 for $\beta(\sigma) \in [0, \frac{1}{3})$, take action a_2 for $\beta(\sigma) \in [\frac{1}{3}, \frac{2}{3})$, and take action a_3 for $\beta(\sigma) \in [\frac{2}{3}, 1]$. By monotonicity of the sender's utility it follows that she would induce posterior beliefs $\tau^{BP} = \begin{cases} \frac{1}{3} \text{ with probability } \frac{1}{2} \\ \frac{2}{3} \text{ with probability } \frac{1}{2} \end{cases}$, using the experiment $X^{BP} = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}$.

However, with preferences as in eq. 11, the pair $\{X^*, \Sigma^*\}$ can be an equilibrium, where

$$X^* = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \Sigma^* = \begin{pmatrix} \frac{6}{7} & \frac{3}{7} \\ \frac{1}{7} & \frac{4}{7} \end{pmatrix} \tag{10}$$

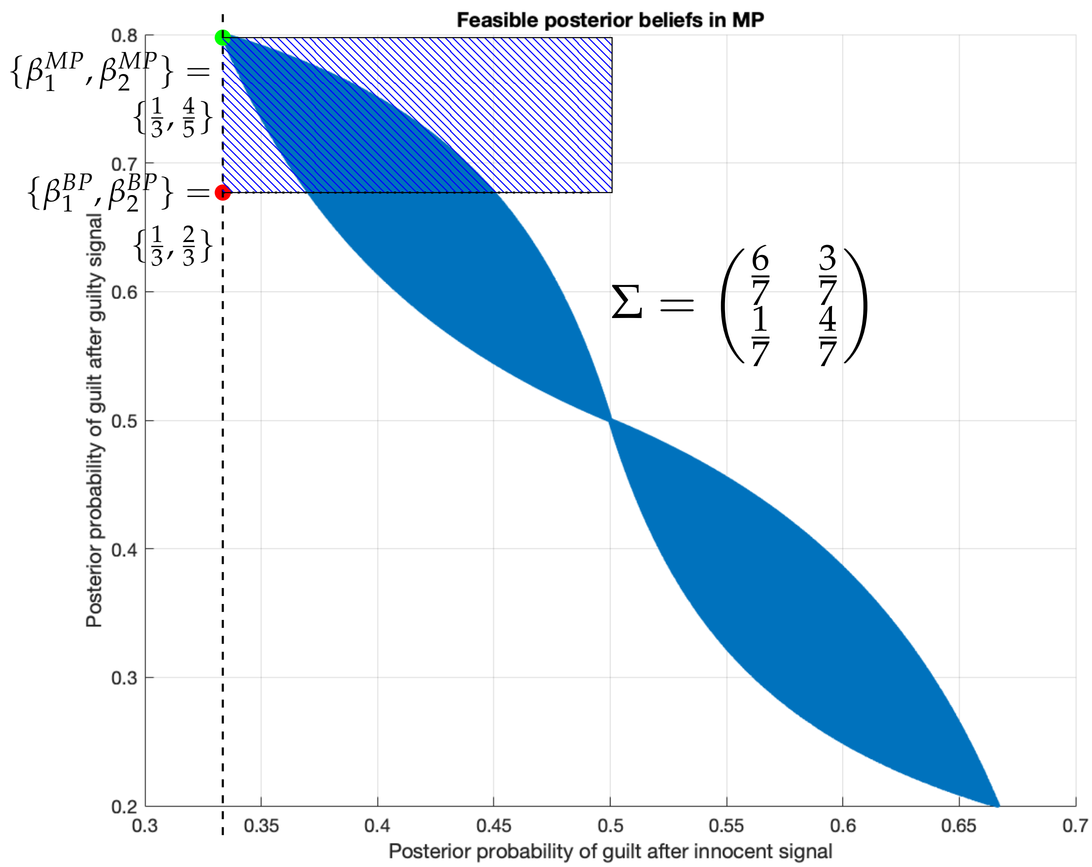

Figure 6: An informative MP equilibrium in a canonical environment.

yielding the following distribution of posterior beliefs: $\tau^{MP} = \begin{cases} \frac{1}{3} \text{ with probability } \frac{9}{14} \\ \frac{4}{5} \text{ with probability } \frac{5}{14} \end{cases}$.
It can be checked (and is in fact, graphically apparent from figure 20) that τ^{MP} is a mean-preserving spread of τ^{BP}, so that the MP outcome is Blackwell more informative than the BP outcome. To show that this is an equilibrium, suppose first that the mediator is choosing Σ^*; $F(\Sigma^*, \frac{1}{2})$ is depicted in figure 6. In the shaded region the sender's utility is increasing in the southwestern direction, so for k_3 (the level of the highest flat segment) high enough the best that she could do is induce the green point (using a fully revealing information structure). And given that the sender is choosing full revelation, it is easy to construct mediator preferences that would result in Σ^* being the optimal choice.

$$v^M(\beta) = \begin{cases} 0.15, & \beta \leq \frac{1}{3} \\ 0.05, & \frac{1}{3} < \beta < 0.8 \\ 0, & \beta \geq 0.8 \end{cases} \tag{11}$$

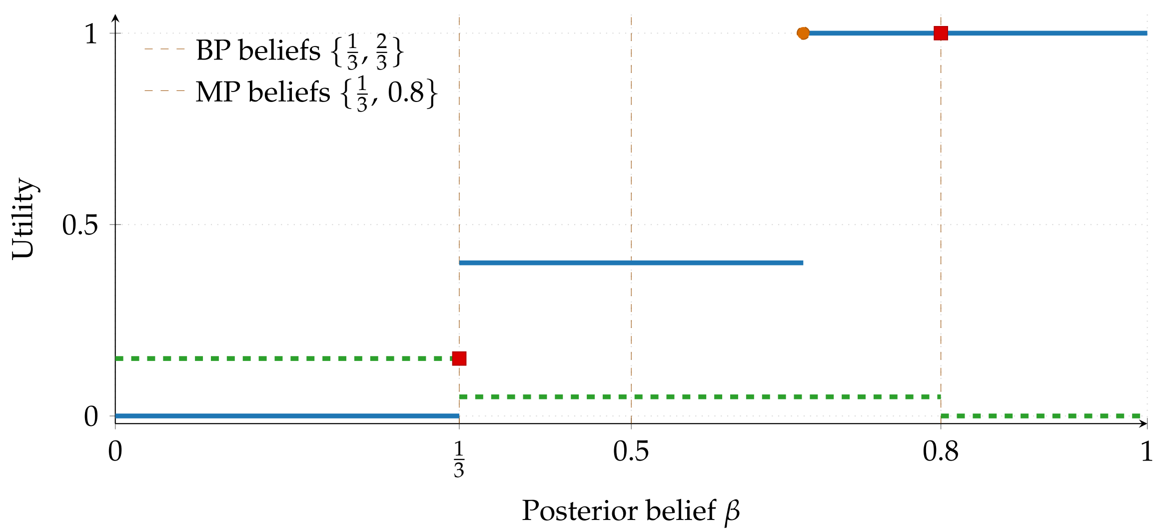

Figure 7: Step functions for sender and mediator utilities.

5 Mediated Persuasion Equilibria: Characterizations

Let $\Omega = (L, H)$, let $\beta_0 = \mathbb{P}(H)$ be an interior prior, suppose that there is only one mediator, and that $v^R(\beta)$ is convex in β.

A major, and surprising, observation from Example 1 has been that the receiver benefits from mediation when the mediator's preferences are more dispersed than those of the sender. Interestingly, if the mediator's preferences are the same as the receiver (i.e. also convex in β), then the receiver does *not* benefit from mediation. An implication of this finding, is, of course, that if one were to decide whether or not to include a mediator in a setting such as ours, and one were inclined to choose outcomes that are preferred by the receiver, then one would choose a mediator only if such a mediator's preferences were sufficiently *different* from the receiver's.

The second important observation is that if there is to be any information revelation, the utilities of the players have to be convex over a set that includes the prior. To that end, consider the following:

Definition 2 (*Convex basin*). *A convex basin of agent i ($i = S, M_1, M_2, \ldots, M_n, R$), $CB^i \subseteq \Delta(\Omega)$ is a convex set of beliefs such that $\{\beta \in \Delta(\Omega) : v^i(\beta) = \text{cav}\, v^M(\beta)\}$.*

A second critical definition is the following:

Definition 3 (*Coincident beliefs*, Gentzkow and Kamenica (2016)). *Let $cav(u)$ denote the concavification of a function u: the upper concave envelope of u. A belief $\beta \in \Delta(\Omega)$ is said to be coincident, if $u(\beta) = cav(u(\beta))$.*

An *extreme point* of a convex set O is a point $\beta \in O$ that cannot be represented as a convex combination of two other points in O; the set of extreme points is denoted by $\mathcal{E}(O)$. We can now state a necessity result: what properties equilibrium beliefs have to satisfy it mediated persuasion is to benefit the receiver. Consider two states $\{L, H\}$ with prior β_0. Assume $v^R(\beta)$ is convex.

Let $\beta_0 \in (0, 1)$ be the prior, $v^R, v^S, v^M : [0, 1] \to \mathbb{R}$ continuation utilities with v^R convex.

Observe that, as the above examples make clear, if mediation is to be beneficial, then $CB^S \subset CB^M$, and furthemore, as is also clear from the examples, the mediator's beliefs are (in a beneficial equilibrium) coincident at the boundaries of CB^M. In all of the above examples, a beneficial MP equilibrium was implemented in the following way:

1) Find CB^M

2) Let $\{\beta_-, \beta_+\}$ be the mediator coincident beliefs over CB^M; in these examples we had $\left[\beta_-^{BP}, \beta_+^{BP}\right] \subset \left[\beta_-, \beta_+\right]$

3) Find garbling Σ^c, a candidate equilibrium gabrling, that yields Bayes-plausible $\{\beta_-, \beta_+\}$ when $X = I$

4) Then find concavification of v^S over $F(\Sigma^c, \pi)$

5) If this concavification (over a particular, fixed Σ^c) yields that the best response to Σ^c is $X^* = I$, we are done. Implement the informative MP equilibrium, $\Sigma^* = \Sigma^c, X^* = I$.

Define an informative (i.e. beneficial) MP equilibrium of this form (that is, $\left(I, \Sigma^*\right)$) a *canonical* equilibrium.

5.1 Theorems 1 and 2: Sufficient and Necessary Conditions in the Case of a Dichotomy and one Mediator

Using the concepts and intuition of the preceding sections, it is straightforward to construct a complete characterization of canonical equilibria in the case of a dichtomy (in states, signals, and actions), and one mediator.

Theorem 1 (Sufficient Primitive Conditions for Beneficial Mediation)**.** *Let $v^S, v^M : \Delta(\Omega) \to \mathbb{R}$ be the sender's and mediator's indirect utilities, $v^R : \Delta(\Omega) \to \mathbb{R}$ be the receiver's utility (assumed convex), and $\pi \in \Delta(\Omega)$ the common prior.*

Suppose that:

i) Beneficial mediation is possible:

$$\pi \in CB^M \tag{12}$$

ii) Extreme point existence: There exist extreme points $\beta_-^{ext}, \beta_+^{ext} \in \mathcal{E}(CB^M)$ with $\beta_-^{ext} < \pi < \beta_+^{ext}$ such that:

$$\exists \lambda \in (0,1): \quad \lambda \beta_-^{ext} + (1-\lambda)\beta_+^{ext} = \pi \tag{13}$$

iii) Mediator optimality:

$$\lambda v^M(\beta_-^{ext}) + (1-\lambda) v^M(\beta_+^{ext}) = \max_{\substack{\beta_-, \beta_+ \in \mathcal{E}(CB^M) \\ \mu\beta_- + (1-\mu)\beta_+ = \pi}} \left[\mu v^M(\beta_-) + (1-\mu) v^M(\beta_+) \right] \tag{14}$$

iv) Canonical garbling: Define Σ^c, the canonical signal, by:

$$\Sigma_{ij}^c = \frac{\beta_j^{ext}(i) \cdot \lambda_j}{\pi(i)} \tag{15}$$

where $j \in \{-,+\}$, $\lambda_- = \lambda$, $\lambda_+ = 1-\lambda$, and Σ^c has full rank (i.e., two).

v) Sender optimality (over associated canonical feasible set):

$$\lambda v^S(\beta_-^{ext}) + (1-\lambda) v^S(\beta_+^{ext}) = \max_{\substack{(\beta_1, \beta_2) \in F(\Sigma^c, \pi) \\ \mu\beta_1 + (1-\mu)\beta_2 = \pi}} \left[\mu v^S(\beta_1) + (1-\mu) v^S(\beta_2) \right] \tag{16}$$

where $F(\Sigma^c, \pi)$ is the feasible set given Σ^c.

vi) Strict convexity:

$$v^M \text{ is strictly convex at } \beta_-^{ext} \text{ or } \beta_+^{ext} \tag{17}$$

When these conditions hold, the canonical equilibrium (I, Σ^c) *exists with outcome:*

$$\tau^{MP} = \{(\lambda, \beta_-^{ext}), (1-\lambda, \beta_+^{ext})\} \tag{18}$$

Moreover, $\tau^{MP} \succeq_m \tau^{BP}$ *and* $V_R(\tau^{MP}) > V_R(\tau^{BP})$*; that is the MP equilibrium is beneficial.*

Note that this theorem greatly simplifies the search for equilibria, because instead of considering all possible deviations for th sender, theorem 1 gives just *one* garbling to check against.

While non-canonical equilibria may exist when players have multiple optimal strategies, the restriction is without loss for three reasons: (i) they arise naturally when the mediator's optimization problem has a unique solution, (ii) they represent focal communication protocols where the sender provides maximal information and the mediator applies optimal garbling, and (iii) for generic[8] preferences, all equilibrium posterior distributions can be implemented by a canonical equilibrium.

Theorem 2 (Necessary Conditions for Beneficial Mediation)**.** *If a beneficial mediation equilibrium exists with outcome* $\tau^* = \{(\lambda^*, \beta_1^*), (1-\lambda^*, \beta_2^*)\}$ *where* $\beta_1^* < \beta_2^*$*, then:*

(i) Beneficial mediation is possible: $\pi \in CB^M$

(ii) Extremality: $\beta_1^*, \beta_2^* \in \mathcal{E}(CB^M)$ *and* $\beta_1^* < \pi < \beta_2^*$

(iii) Mediator optimality:

$$\lambda^* v^M(\beta_1^*) + (1-\lambda^*) v^M(\beta_2^*) = \max_{\substack{\beta_-, \beta_+ \in \mathcal{E}(CB^M) \\ \mu\beta_- + (1-\mu)\beta_+ = \pi}} \left[\mu v^M(\beta_-) + (1-\mu) v^M(\beta_+)\right] \tag{19}$$

[8] I.e. those for which the mediator's convex basin has multiple representations of the prior as convex combinations of its extreme points

(iv) Implementability: There exist (X^*, Σ^*) *with* $p(\Sigma^* X^*) = \tau^*$ *and* $\text{rank}(\Sigma^*) = |\Omega|$

(v) Non-degeneracy: v^M *is strictly convex at* β_1^* *or* β_2^*

These conditions are necessary but not sufficient (without the additional sender optimality condition).

The gap between theorems 1 and 2 is condition $v)$ in theorem 1 about sender optimality. To ensure (I, Σ^c) is an equilibrium, we must verify the sender has no profitable deviation given Σ^c. This requires checking condition $v)$. If one only knows an equilibrium exists (but don't know which strategies), one can't directly state "the sender has no profitable deviation" without referring to the specific equilibrium strategies.

The necessary conditions $i)$ thorugh $v)$ characterize the outcome τ^* of any beneficial equilibrium, but don't characterize the strategies. To get sufficiency, we focus on the canonical equilibrium and add the explicit sender optimality check.

Proposition 1 (Complete Characterization via Canonical Equilibria)**.** *If we restrict attention to canonical equilibria (of the form* (Σ^c, I)*), then conditions in Theorem 1 are necessary and sufficient.*

That is: a beneficial canonical equilibrium exists if and only if the premises of theorem 1 hold.

5.2 Theorem 3 : Sufficient Conditions in the Case of n States

Theorem 3 (Sufficient conditions with mixed strategies; $|\Omega| = n \geq 2$)**.** *Let the state space be* $\Omega = \{1, \dots, n\}$ *with full-support prior* $\pi \in \Delta(\Omega)$*. The sender (S) chooses an experiment* X *from the compact convex set* $\mathcal{X}$ *of column-stochastic* $k \times n$ *matrices; the mediator (M) chooses a garbling* Σ *from the compact convex set* $\mathcal{S}$ *of column-stochastic* $m \times k$ *matrices; the receiver (R) has a finite action set* A*. Let* $v^S, v^M, v^R : \Delta(\Omega) \to \mathbb{R}$ *be the induced continuation utilities; assume* v^R *is convex. Define the mediator's convex* *basin*

$$\mathrm{CB}^M \triangleq \{\beta \in \Delta(\Omega) : v^M(\beta) = \operatorname{cav} v^M(\beta)\} \tag{20}$$

Suppose that:

(i⋆) *Beneficial mediation is possible:* $\pi \in \mathrm{CB}^M$

(ii⋆) *Extreme support: There exist* $r \in \{2, \ldots, n\}$*, affinely independent* $\beta_1^{\text{ext}}, \ldots, \beta_r^{\text{ext}} \in \mathcal{E}(\mathrm{CB}^M)$ *and* $\lambda \in \Delta(\{1, \ldots, r\})$ *with* $\pi = \sum_{i=1}^r \lambda_i \beta_i^{\text{ext}}$

(iii⋆) *Mediator optimality on the convex basin:*

$$\sum_{i=1}^r \lambda_i v^M(\beta_i^{\text{ext}}) = \max_{\substack{\{\tilde{\beta}_j\} \subset \mathcal{E}(\mathrm{CB}^M), \\ \mu \in \Delta, \sum_j \mu_j \tilde{\beta}_j = \pi}} \sum_j \mu_j v^M(\tilde{\beta}_j) \tag{21}$$

(iv⋆) *Define* Σ^c *with* r *effective signals* $s_1, \ldots, s_r$ *by*

$$\Pr(s_i \mid \omega) = \frac{\lambda_i \, \beta_i^{\text{ext}}(\omega)}{\pi(\omega)} \qquad (\omega \in \Omega, \; i = 1, \ldots, r), \tag{22}$$

padding with rows of zeros if needed. Then $\Sigma^c I_n$ *induces posteriors* $\{\beta_i^{\text{ext}}\}$ *with probabilities* $\{\lambda_i\}$*. Assume* $\operatorname{rank}(\Sigma^c) = n$ *(full column rank).*[9] *A sufficient condition is* $r = n$ *with* $\{\beta_i^{\text{ext}}\}$ *affinely independent.*

(v⋆) *Sender optimality against* Σ^c*: Let* $F(\Sigma^c, \pi) \triangleq \{\, p(\Sigma^c X) : X \in \mathcal{X} \,\}$ *be the set of implementable posterior distributions against* Σ^c*. Then*

$$\sum_{i=1}^r \lambda_i v^S(\beta_i^{\text{ext}}) = \max_{\tau \in F(\Sigma^c, \pi)} \mathbb{E}_{\beta \sim \tau}[v^S(\beta)] \tag{23}$$

(vi⋆) v^M *is strictly convex along some exposed direction at at least one* β_j^{ext} *(equivalently: the supporting hyperplane of* $\operatorname{cav} v^M$ *at* β_j^{ext} *is unique).*

[9] This is easy to check in the case of $m = k$; otherwise one can use an pseudoinverse and confirm the column-stochasticity of the result.

Then the canonical profile $(X^*, \Sigma^*) = (I_n, \Sigma^c)$ is a perfect Bayesian equilibrium even when S and M are allowed mixed strategies. Its outcome

$$\tau^{MP} = \{(\lambda_i, \beta_i^{\text{ext}})\}_{i=1}^{r} \tag{24}$$

is Blackwell-more informative than the Bayesian-persuasion outcome τ^{BP}; if the improvement is strict, then $V_R(\tau^{MP}) > V_R(\tau^{BP})$ by convexity of v^R.

The proof appears in Appendix C, and relies on the concepts developed earlier, essentially generalizing the idea of supporting the beneficial equilibrium on the extreme points of the convex basin of the mediator.

6 Literature

This work is at the intersection of several literatures - strategic information transmission (cheap talk), persuasion, and noisy communication. It is in the spirit of the celebrated "Bayesian persuasion" approach of Kamenica and Gentzkow (2011) (referred to simply as "KG" for brevity hereafter) who consider a simpler version of this problem, and discuss an application of a certain concavification result first considered in chapter 1 of Aumann and Maschler (1995). Sah and Stiglitz (1986) introduced the analysis of economic systems organized in parallel and in series; hierarchies and polyarchies of persuasion via provision of information have already been explored in previous work (Gentzkow and Kamenica (2017a), referred to as "GK" henceforth, not to be confused with "KG"),

Critically, I study a game where the players move simultaneously (this is just a modeling trick of course - they do not have to actually act at the same time - the reason for this is because typically one party is not aware of the ratings mechanism or the choice of the financial instruments of the other party when committing to an action; it could also be simply because a player is unable to detect deviations in time to adjust their own strategy); the key point is that the mediator does not see the choice of the sender before making her own choice as in many other models. In other words, I assume "simultaneous double commitment" in a sense - commitment to an information structure for the sender and the mediator, along the lines discussed in KG. This feature generates an interesting possibility of having a kind of prisoner's dilemma not in actions, but in information.[10] The flow of information is path-dependent (as in Li and Norman (2018)), yet not quite sequential while action choices for the sender and the mediator are simultaneous.

There are several recent papers related to the present model. The closest are papers by Arieli, Babichenko, and Sandomirskiy (2022), Zapechelnyuk (2022), Li and Norman (2018), Ambrus, Azevedo and Kamada (2013) and Ivanov (2010). Arieli, Babichenko, and Sandomirskiy (2022) consider a very similar model with a different order of moves (in their model the mediator(s) before after the receiver). This key difference, as they note, makes their work more tractable. Zapechelnyuk (2022) studies a model where the mediator can also create information; Ivanov (2010) considers a uniform-quadratic cheap talk model with a strategic mediator, where the difference in preferences is the bias of the three parties.

In five of the six papers discussed in the previous paragraph, the mediator moves *after* the sender. This key difference between these papers and the present work is that the mediator observes the action of the sender. This feature allows the mediator to condition her action on that of the sender; something which I disallow. This, in turn, introduces a serious technical difficulty - without knowing what the sender is choosing, the same action choice for the mediator may lead to different outcomes, because the distribution of the final beliefs will depend on both choices. I discuss a solution to this problem in a special case in section 4.

Ambrus, Azevedo and Kamada (2013) consider a cheap talk model where the sender and receiver also communicate via chains of intermediators. My work is

[10]This is also discussed in GK.

similar in that talk is "cheap" here as well, meaning that the specific choices of the sender and the mediators do not enter their utility functions directly and only do so through the action of the receiver; in addition, we, too, have an analogous communication sequence. The difference is that the sender is not informed about the state. Li and Norman (2021)'s work on sequential persuasion serves as another stepping stone - they have a very similar model of persuasion, except that the senders move sequentially, observing the history of actions of the senders who moved before them (unlike in our model), and can provide arbitrarily correlated experiments. The other relevant work is Gentzkow and Kamenica (2017)'s work on competition in persuasion where the senders move simultaneously (like in our model), but all senders are trying to provide information about the state of the world, whereas I study an environment where the mediator is trying to provide information about the realization of the sender's experiment. Lipnowski, Ravid, and Shishkin (2022) also study a related environment where a "weak institution" in their parlance plays the role of a kind of informational mediator, although the setup is considerably different and there is no role for the interplay of preferences which I focus on here. The subject of introducing a mediator to potentially improve outcomes has also been studied in contract theory (see, inter alia, Pollrich (2017) and Rahman and Obara (2010)).

Perez-Richet and Skreta (2018) present a complementary model that differs in one key respect - the mediator (using my nomenclature) moves first and her choice is observed by the sender before the sender acts. My focus is on analyzing outcomes of a particular game as one changes preferences for the mediator (and fixing the signal realization spaces in advance), while they focus on equilibria of a game where the preferences of the mediator are always fully aligned with those of the receiver. More specifically, they construct a "test" where the sender/persuader employs a continuum of signal realizations to pass or fail different types of sender. Plainly, the difference between our work and theirs is that I fix the signal realization space and vary the preferences of the players, while they fix the preferences and derive the optimal signal realization space (and signal realization probabilities). Notably, the contrast with Perez-Richet and Skreta (2018) immediately shows that it is strictly with loss of generality to restrict the space of signal realizations, as I do in the paper. This assumption, however, greatly simplifies our problem.[11]

Strulovici (2017) in his "Mediated Truth" work explores a somewhat related environment where a "mediator" - an expert of some sort or a law enforcement officer - has access to information that is "costly to acquire, cheap to manipulate and produced sequentially". He shows that when information is reproducible and not asymptotically scarce (for example, one can perform many scientific experiments) then societies will learn the truth, while when information is limited (such as evidence from a crime) the answer is negative. In our work I consider a one-shot game, but his insight provides an interesting contrast. For example, a repeated version of the game considered here would satisfy the condition for evidence to not be asymptotically scare, however, it is not clear that this is enough to overcome the incentive problem when the mediator can only garble the signals; certainly there will be no learning is the unique equilibrium in our model is uninformative, as can be the case.

Le Treust and Tomala (2019) study a very similar, but simpler setting. They consider persuasion with an additional constraint - exogenous noise - and show that while the sender generically suffers a loss as a result of the noise, information-theoretic tools show that the sender can do as well as possible, provided she plays the game enough times (i.e. enough independent copies of the same basic problem are available). Their model can be viewed as a (possibly repeated) special case of the model studied in this paper, with a nonstrategic mediator who chooses a garbling structure that results in the exogenous noise structure.

[11] indeed, if one were to consider a problem of which both this paper and Perez-Richet and Skreta (2018) are special cases, one would have a strategic problem with an unrestricted domain of utilities with complicated infinite-dimensional action spaces.

Tsakas and Tsakas (2021) also study the problem of Bayesian persuasion subject to exogenous noise. They show that while it is in principle possible for the sender to benefit from noise, they obtain analogous results to ours (that the sender is always worse off with more noise) when comparing similar noise structures. The reason for why in our model the sender is always worse off, and in their model the sender can be better off is that they consider additional noise structures (which they refer to as "partitional" channels), and the sender may be better off when faced with noise structures that are both canonical and partitional. Thus, our work agrees with theirs along the dimension along which the environments are comparable, but I also consider strategic interaction.

Ichihashi (2017) studies a model in which the sender's information may be limited; he focuses on the cases where doing so might benefit the receiver. In our model a similar role is played by the mediator who modifies the information produced by the sender, and can only modify it by garbling (i.e. only decreasing the amount of information). Thus, while Ichihashi (2017) limits the sender's information, I limit what the sender can do with that information.

7 Concluding Remarks

This work has presented the first (to my knowledge) explicit example of a situation where one party, moving simultaneously, can add information, one party can only subtract information, and preferences for which this mechanism benefits the final receiver of information. To reiterate; the receiver benefits from mediation when the preferences of the mediator are sufficiently different; delegation is thus beneficial only in certain specific situations.

Models of informational mediation that are derived from real-world information flows shed light on the welfare consequences of these arrangements. Most of the time they are not beneficial, but in certain cases - as was shown earlier -

they may be. This provides one critical way in which the receiver may be able to increase the amount of information revelation, if the receiver is unable to do anything else (such changing the preferences of the sender, introducing multiple senders, or obtaining information herself) - she may be able to improve her lot by introducing a party that can destroy information, provided that party has freedom of action, and has different preferences from the receiver. This work has illustrated this new mechanism of increasing information revelation.

Appendix A: The Set of Feasible Posteriors: A Characterization

For tractability I work with a binary model where there are two states of the world and two experiment and signal realizations. This is with loss of generality, but will serve well to illustrate the basic idea of how to compute a best response for the sender given the choice of the mediator.

7.1 A Non-strategic Setting: Feasible Posteriors for a Fixed Garbling

Setting aside the issues of strategic behavior for now, I first ask a simpler question: given a *fixed*[12] signal (or equivalently, a fixed garbling), or a fixed experiment, what are all the posterior distributions that can be induced? The problem narrowed thus, an important connection with the cheap talk and communication literature can be made. Blume, Board and Kawamura (2007) discuss a model of cheap talk where the signal sent by the sender is subject to random error - with a small probability the message observed by the receiver is not the message sent by the sender, but rather, a message sent from some other distribution that does

[12] I.e. not strategically chosen by a player as a function of her preferences.

not depend on the sender's type or the message chosen. I make this connection to note that choosing an information structure that will be subjected to a fixed, non-strategically-chosen garbling is exactly equivalent to choosing a random signal that will be subject to noise. Thus, this model subsumes a model on Bayesian persuasion with noisy communication, similar to those studied by Le Treust and Tomala (2019) and Tsakas and Tsakas (2021).

In the (different but related) setting of cheap talk, as noted by Ambrus, Azevedo and Kamada (2013) as well as Blume, Board and Kawamura (2007) stochastic reports make incentive compatibility constraints *easier* to satisfy. This will not quite be the case here, but this will nevertheless be an illuminating exercise.

As mentioned above, for tractability[13] I work in the simplest possible environment of binary signal and state spaces for both the sender and the mediator. In addition to being the simplest nontrivial example of the problem, working with two-by-two square matrices has a very important additional advantage. The rank of such a stochastic[14] matrix can be only two things - one or two. If the rank of a two-by-two stochastic matrix is one, that means that not only the columns (and rows) are linearly dependent, but they must, in fact be identical. In that case the garbling is fully uninformative - it can be readily checked that this results in the same posteriors as the canonical complete garbling; namely, the posterior (after either signal realization) is equal to the prior. The other possible case is that the matrix has rank two - but that automatically means that such a matrix is invertible. I will rely on the existence of an inverse throughout.[15]

More specifically, let ϵ be a small positive number, set the space of experiment realizations to be $E = \{e_L, e_H\}$ and suppose that the sender and receiver play a game exactly identical to KG (that is, there is no mediator), except that with probability ϵ the signal observed by the receiver (denoted by e^o) is not the signal

[13] And with loss of generality, discussed later.

[14] Which of course, rules out the zero matrix, which has rank zero.

[15] I comment on the interpretation of the rank of a garbling matrix later in the discussion, and in related contemporaneous work ("Beyond the Blackwell Order in Dichotomies").

sent (which I denote by e^s), but a signal chosen from the following distribution

$$e^o = \begin{cases} e_H & \text{with probability } p \\ e_L & \text{with probability } 1-p \end{cases} \tag{25}$$

The key thing is that this distribution is independent of both the type and the signal realized. The probabilities of observed signals as functions of the parameters and realized signals can be computed as usual:

$$\mathbb{P}(e^o = e_H | e^s = e_H) = 1 - \epsilon + \epsilon p \tag{26}$$

$$\mathbb{P}(e^o = e_L | e^s = e_H) = \epsilon - \epsilon p \tag{27}$$

$$\mathbb{P}(e^o = e_L | e^s = e_L) = 1 - \epsilon p \tag{28}$$

$$\mathbb{P}(e^o = e_H | e^s = e_L) = \epsilon p \tag{29}$$

Then this is equivalent to having a garbling

$$\Sigma = \begin{pmatrix} \sigma_1 & \sigma_2 \\ 1-\sigma_1 & 1-\sigma_2 \end{pmatrix} = \begin{pmatrix} \epsilon p - \epsilon + 1 & \epsilon p \\ \epsilon - \epsilon p & 1 - \epsilon p \end{pmatrix} \tag{30}$$

with realization space $S = \{e^o_L, e^o_H\}$.

Denote by $X = \begin{pmatrix} x & y \\ 1-x & 1-y \end{pmatrix}$ the experiment chosen by the sender so that

$$B = \Sigma X = \begin{pmatrix} x(\epsilon p - \epsilon + 1) - \epsilon p(x-1) & y(\epsilon p - \epsilon + 1) - \epsilon p(y-1) \\ (\epsilon p - 1)(x-1) + x(\epsilon - \epsilon p) & (\epsilon p - 1)(y-1) + y(\epsilon - \epsilon p) \end{pmatrix} \tag{31}$$

is the resulting distribution of signal observations given states. Letting $\Omega = \{\omega_H, \omega_L\}$

be the set of states and setting prior belief of $\omega_L = \pi$ the posterior beliefs are

$$\beta(s_H) = \mathbb{P}(\omega_L|s_H) = \frac{\pi\left[y(\epsilon p - \epsilon + 1) - \epsilon p(y-1)\right]}{\pi\left[y(\epsilon p - \epsilon + 1) - \epsilon p(y-1)\right] + (1-\pi)\left[x(\epsilon p - \epsilon + 1) - \epsilon p(x-1)\right]} \tag{32}$$

and

$$\beta(s_L) = \mathbb{P}(\omega_L|s_L) = \frac{\pi\left[(\epsilon p - 1)(y-1) + y(\epsilon - \epsilon p)\right]}{\pi\left[(\epsilon p - 1)(y-1) + y(\epsilon - \epsilon p)\right] + (1-\pi)\left[(\epsilon p - 1)(x-1) + x(\epsilon - \epsilon p)\right]} \tag{33}$$

Define the set of feasible beliefs to be a pair

$$F(\Sigma, \pi) \triangleq \{\beta_{\Sigma X}(s_H), \beta_{\Sigma X}(s_L)|X \in \mathbf{X}\} \tag{34}$$

An immediate observation is that the set of feasible beliefs with a garbling is a strict subset of the set of feasible beliefs without one, simply due to the fact that there are extra restrictions in computing $F(\Sigma, \pi)$. To illustrate, let $\epsilon = \frac{1}{100}$ and $p = \frac{1}{4}$ so that there is a 1% chance that the signal will be a noise signal, and if that happens, there is a 75% probability that the signal will be correct. The set of Bayes-plausible beliefs is depicted in red in the figure 5, while the set of feasible beliefs given this particular Σ is in blue.

Clearly the "butterfly" set of feasible beliefs (left) is a strict subset of the Bayes-plausible set on the right, verifying the observation made above. Thus, *for a fixed garbling, not all Bayes-plausible posterior beliefs can be induced.*

Perhaps another illustration can make this point more starkly - suppose the probability of error increases tenfold, so that there is a much greater chance that the signal is a noise signal. The resulting sets are depicted in figure 6.

Thus, increasing the probability of error (or noise signal) shrinks the set of feasible beliefs. This is consistent with intuition - if the signal is pure noise, then there should not be any update of beliefs (and thus the set would shrink to a single point at the prior), and with a larger probability of noise one would update "less".

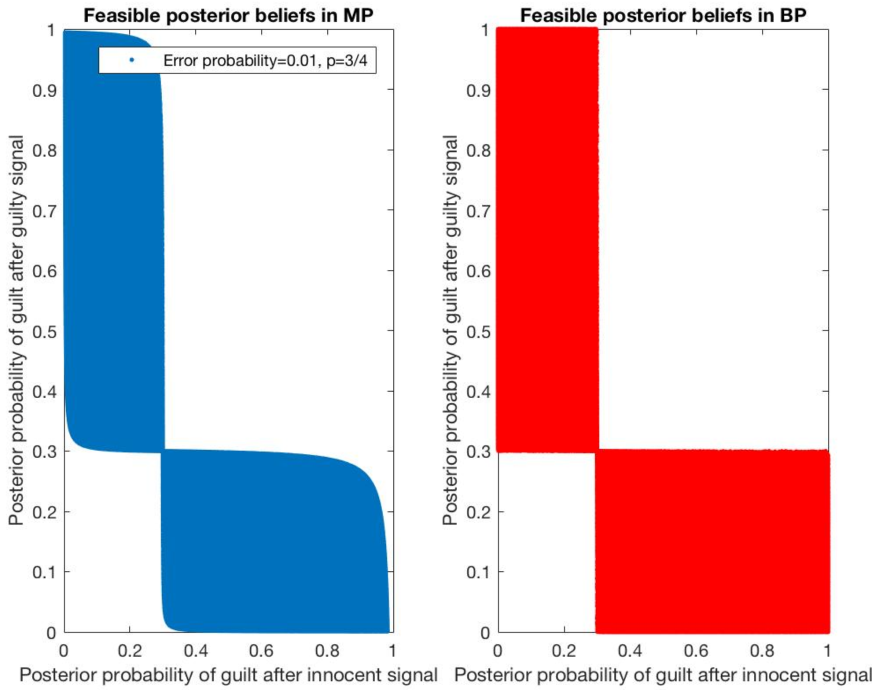

Figure 8: Comparing the feasible sets of posteriors.

I make precise the idea that with a less informative garbling "fewer" posteriors are available below.

This discussion leads to the following question: What is the set of feasible posterior beliefs given a garbling (without computing whether or each belief is feasible one by one as was done in computing the figures above, which were generated by simulating random matrices with the appropriate stochasticity constraints)? One way of answering this question is to trace out the confines of the feasible set. As luck would have it, there is an observation I can make that simplifies this a great deal. Fixing one posterior belief (say, β_1 the posterior after the innocent signal) and then asking what would the elements X need to be to either maximize or minimize the other posterior belief, it turns out that either x or y (or both) will always be 1 or 0. Fix $\Sigma = \begin{pmatrix} \sigma_1 & \sigma_2 \\ 1-\sigma_1 & 1-\sigma_2 \end{pmatrix}$, let π be the prior belief and con-

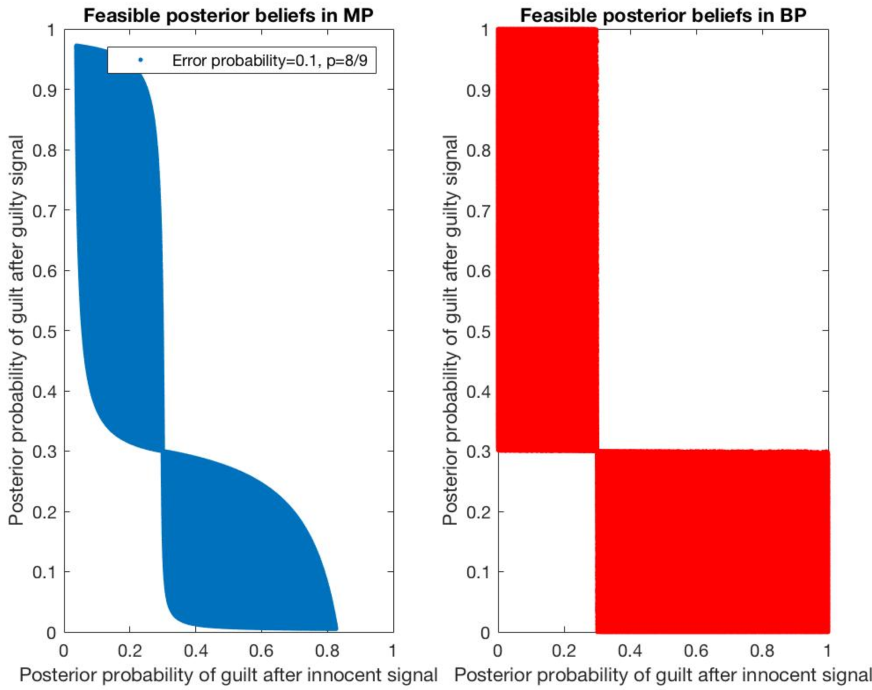

Figure 9: Increasing noise shrinks the set of feasible posteriors.

sider $X = \begin{pmatrix} x & y \\ 1-x & 1-y \end{pmatrix}$. Computing outer limits of $F(\Sigma, \pi)$ is equivalent to the following program:

$$\max_{x,y} \beta_1 = \frac{\pi[\sigma_1 y - \sigma_2(y-1)]}{\pi[\sigma_1 y + \sigma_2(y-1)] + (1-\pi)[\sigma_1 x - \sigma_2(x-1)]} \tag{35}$$

$$s.t. \quad \beta_1 = const. \tag{36}$$

$$0 \leq x \leq 1; 0 \leq y \leq 1 \tag{37}$$

The solution (which I do not exhibit, as it is straightforward but somewhat tedious) shows that either x, or y or both will be 0 or 1 (and of course, one could also have fixed β_2 and let that be the parameter; the answer would be the same). The result is intuitive (maximizing a posterior belief requires maximizing the probability of one of the signals in the first place), but this verifies the intuition formally.

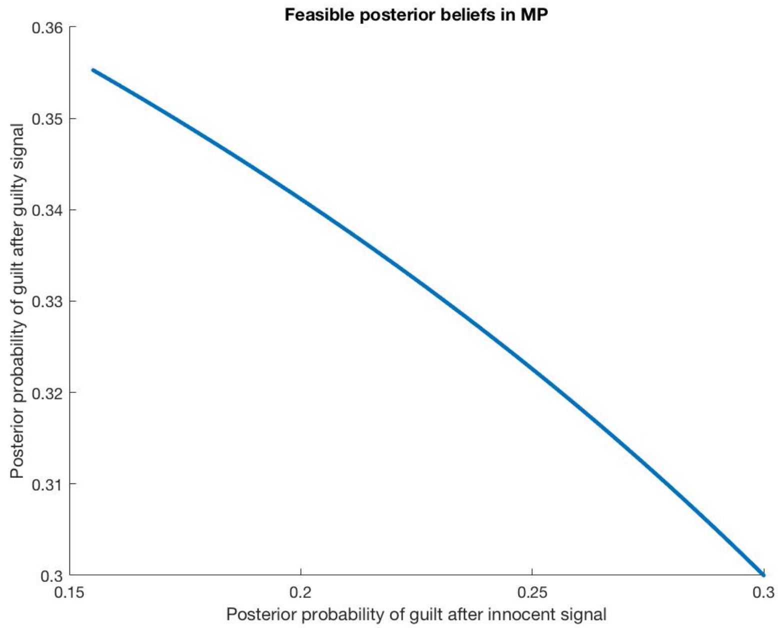

Figure 10: Tracing the Outer Limit of $F(\Sigma, \pi)$: First Boundary.

This observation can be operationalized in the following way: fix first one of four extreme points of the X matrix, and then trace out the corresponding possible beliefs by systematically varying the other probabilities in the experiment, which yields a curve (or a path, in topological terms) parametrized by a single number - the probability of one of the signals.

Let us illustrate this approach using $M = \begin{pmatrix} \frac{1}{3} & \frac{1}{7} \\ \frac{2}{3} & \frac{6}{7} \end{pmatrix}$. The question is, what is $F(\Sigma, \pi)$ for this garbling? I use the algorithm just prescribed: first fix a perfectly revealing part of the experiment, and then vary the corresponding distribution.

Letting $X^1 = \begin{pmatrix} 1 & p \\ 0 & 1-p \end{pmatrix}$ and varying p from 0 to 1 yields the following (blue) curve in figure 10.

Now fix the next extreme point: $X^2 = \begin{pmatrix} 0 & p \\ 1 & 1-p \end{pmatrix}$ and again vary p, which yields the following (reddish-brown) boundary in figure 11.

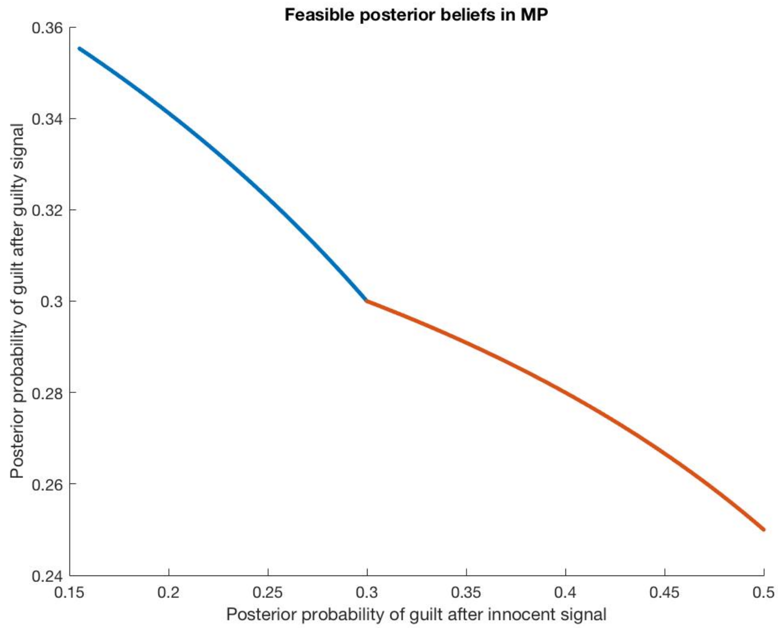

Figure 11: Tracing the Outer Limit of $F(\Sigma, \pi)$: Second Boundary.

Next fix the third extreme point: $X^3 = \begin{pmatrix} p & 1 \\ 1-p & 0 \end{pmatrix}$ and trace the corresponding (yellow) curve, illustrated in figure 12.

And finally trace out the last (purple) curve by using $X^4 = \begin{pmatrix} p & 0 \\ 1-p & 1 \end{pmatrix}$ in figure 10.

This procedure is a simple way of computing the set of $F(\Sigma, \pi)$; *this procedure is a complete characterization of the set of feasible beliefs for* 2×2 *signals and experiments*. Now, for a belief in this set one can ask: does there exist an experiment that yields this belief, and if so, how can it be computed?

One of the implications of proposition 1 in KG is that for every Bayes-plausible posterior distribution there exists an experiment that induces that distribution; they also give an explicit formula for computing such an experiment. In mediated persuasion this fails - an experiment inducing a particular Bayes-plausible distribution may not exist, if it is garbled. However, for beliefs that are feasible

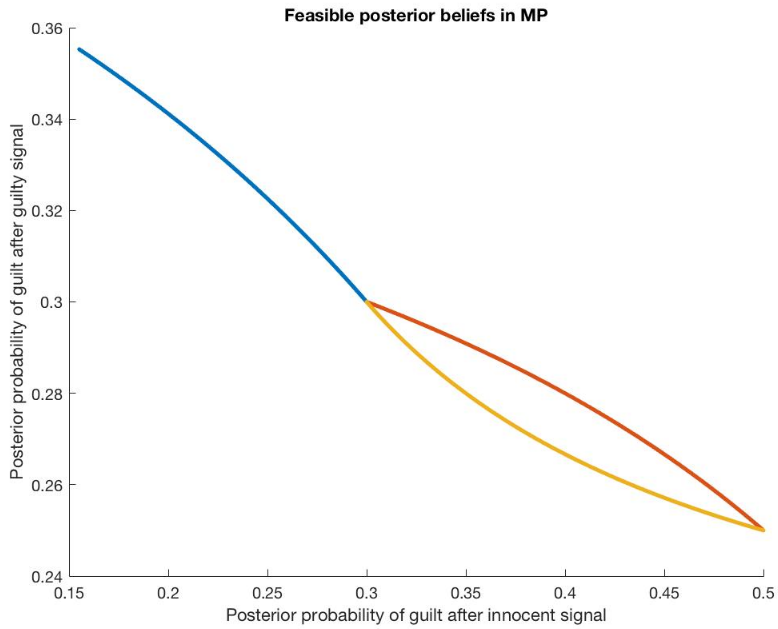

Figure 12: Tracing the Outer Limit of $F(\Sigma, \pi)$: Third Boundary.

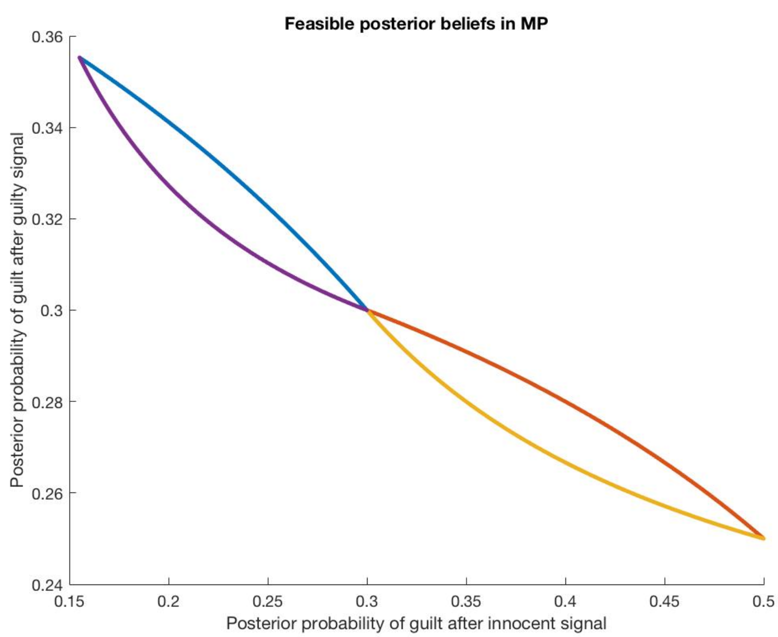

Figure 13: Tracing the Outer Limit of $F(\Sigma, \pi)$: Fourth Boundary.

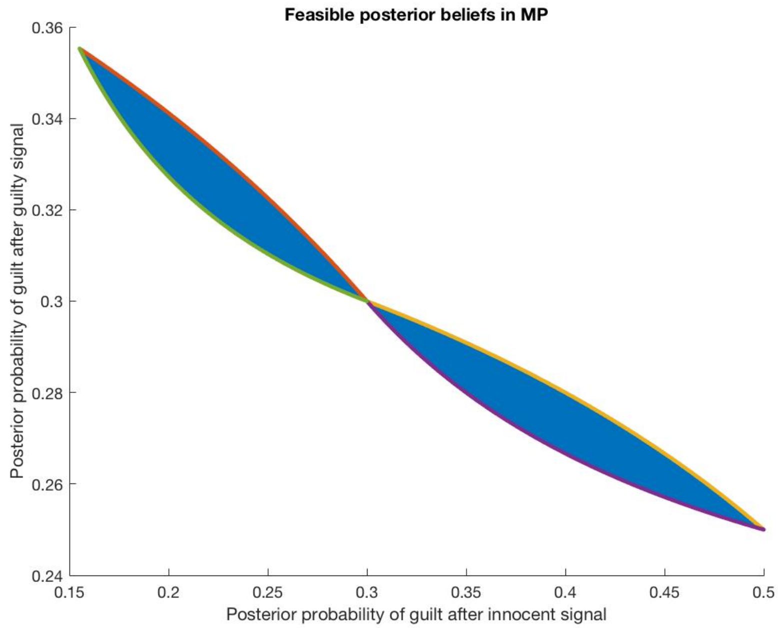

Figure 14: $F(\Sigma, \pi)$: an illustration.

given Σ I have a simple formula for computing the experiment that induces those beliefs.

Definition 4. *Fix Σ. A distribution of posterior beliefs τ is said to be Σ-plausible if there exists a stochastic matrix X such that $p(\Sigma X) = \tau$.*

Theorem 4. *Fix Σ. Suppose that τ is a Bayes-plausible and Σ-plausible distribution of posterior beliefs. There exists an experiment X such that $p(\Sigma X) = p(B) = \tau$.*

We construct the entries in B by setting $b(s|\omega) = \frac{\beta(\omega|s)\tau(\beta)}{\pi(\omega)}$ as in KG; simple algebra shows that this yields a Bayes-plausible distribution that results in the necessary beliefs. The experiment yielding B is then simply $X = \Sigma^{-1}B$.[16] The fact that X is, in fact, an experiment is guaranteed by the fact that the beliefs were feasible in the first place. This is, in a sense, a tautological statement, but it does provide an analogue to proposition 1 in KG by exhibiting an explicit formula for constructing

[16]Note that I am using the existence of Σ^{-1}, which is ensured owing to the discussion at the beginning of the section.

B and then X and showing that both do, in fact, exist.

The above example and proposition suggest a general way of solving the problem with two states, two signal realizations and two experiment realizations with a fixed garbling Σ. First I compute the four outer limits of $F(\Sigma, \pi)$ as above. Then I ask how the sender's utility varies over the feasible set, and having found a maximum point, I construct the optimal experiment yielding those posteriors. And then, given the feasible set of a garbling, one can compute the sender's utility from choosing each posterior in that set (simply plot the sender's utility as a function of the posterior beliefs), find the maximal beliefs and construct the experiment yielding those beliefs. *This procedure shows how to find a best response for the sender.*

We can write this problem and its solution more formally. Let κ be the constant and denote the maximization program by P. Suppose that the program has a solution[17] and denote by $x^*(\sigma_1, \sigma_2, \pi, \kappa)$ the solution. Suppose for now that $\kappa \leq \pi$. This produces a (second posterior belief) function $\beta_2^{max}(y; x^*(\sigma_1, \sigma_2, \pi, \kappa), \sigma_1, \sigma_2, \pi) : [0,1] \rightarrow [0,1]$ I write it to emphasize that all arguments of the β_2^{max} function after the semicolon are parameters, and only the y argument is varying from 0 to 1. Analogously I can compute $\beta_2^{min}(y; x^*(\sigma_1, \sigma_2, \pi, \kappa), \sigma_1, \sigma_2, \pi) : [0,1] \rightarrow [0,1]$. Let $Gr(\beta_2^{max})$ and $Gr(\beta_2^{min})$ be the graphs of the two functions, and let $Co(A)$ be the convex hull of an arbitrary nonempty set A. I then define $F^1(\Sigma, \pi) \triangleq Co(Gr(\beta_2^{max}) \cup Gr(\beta_2^{min}))$; the reason this is possible is that the set of posterior beliefs is convex (because the set of information structures is convex, and Bayes rule is monotonic). Similarly, for $\kappa \geq \pi$ I can compute analogous objects, and define $F^2(\Sigma, \pi)$. Finally, let $F(\Sigma, \pi) \triangleq F^1(\Sigma, \pi) \cup F^2(\Sigma, \pi)$.

There are a number of important and interesting observations about the Σ-feasible set that I can make at this point. Consider the F set illustrated in figure

[17] This amounts to assuming that the κ can actually be a posterior belief, which is not always the case - take for example the belief $\beta_1 = 0.9$ in figure 5. Such a belief is clearly infeasible for that Σ, and thus the program would not have a solution.

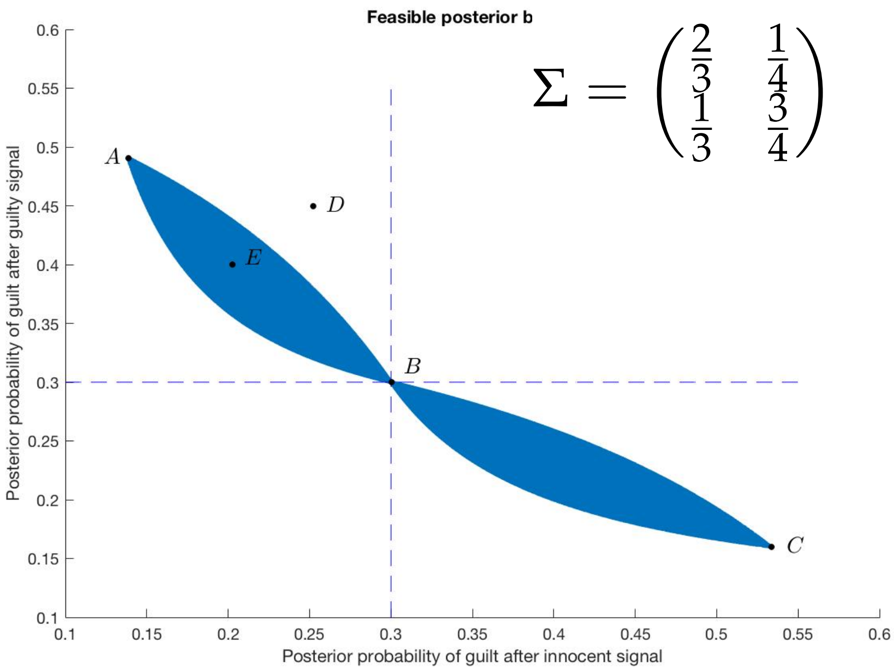

Figure 15: Major features of the feasible set $F(\Sigma, \beta_0)$

15, using the garbling matrix $\begin{pmatrix} \frac{2}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{3}{4} \end{pmatrix}$. In this set each point corresponds to an experiment for the sender. The first thing to notice is that the so-called "butterfly" has two "wings". The "left" wing - the one including point A, i.e. the wing up and to the left from the "origin" (i.e. the point where the posteriors are equal to the prior), is the set that would result if the sender were using "natural" signals - i.e. a guilty signal is more likely in the guilty state and an innocent signal is more likely in the innocent state. The right wing is the set that would result if the sender were instead using "perverse" signals - a *guilty* signal that is more likely in the *innocent* state, and vice versa.[18] This is also equivalent to flipping the labels on the signals.

Consider point B, the point where both posteriors are equal to the prior (with the obvious motivation, I call that the "origin"). Observe that moving weakly northwest meaning decreasing the first posterior while increasing the second - in other words, a mean-preserving spread.[19] Thus, points that are northwest of B are posteriors that are Blackwell-more informative than B. Equivalently, they correspond to signals that Blackwell dominate the uninformative signals. Iterating this, point A is Blackwell-most informative among all the points in the left wing. It can also be verified that point A is *precisely* the two posteriors that correspond to the sender using the fully informative (and "natural") signal. The exact opposite logic applies to the right wing, so that C is the extreme posterior corresponding to the Blackwell-most informative "perverse" signal. Importantly, this logic works only within each wing, (or quadrant by quadrant, which are delineated by the dashed lines), and not on the figure as a whole.

The other observation is that while F seems symmetric around the "origin", in general, it is not. The lack of symmetry comes from the constraints (and biases) imparted by the garbling; $F(\Sigma, \pi)$ is symmetric if and only if M is symmetric.

[18]Note that if the sender were to choose a signal, say, guilty, that is more likely in both states, that would quickly bring beliefs back to the prior, and whether it would be in the right or the left wing would be dictated by the relative probabilities.

[19]The fact that the spread is mean preserving comes from Bayes rule.

Definition 5. *F is said to be symmetric if for each* $\{\beta_1, \beta_2\}$ *if the ordered pair* $\{\beta_1, \beta_2\} \in F$ *then the ordered pair* $\{\beta_2, \beta_1\}$ *is also in F.*

The next observation is that each wing of the butterfly is convex, but the butterfly itself is not. This comes from the fact that for normal (and respectively, for perverse) signals, if two posteriors can be induced, than so can any convex combination (since the set of the relevant stochastic matrices is convex). On the other hand, for the entire set to be convex, taking a point from the left wing, a point from the right and requiring that a mixture would also be in the set would require each signal to be weakly more likely in either state - which is impossible, except for the degenerate case. This is why I can take the convex hull of the extreme beliefs and outer limits for each wing, but not the convex hull of the entire butterfly.

The final observation that I can make is the following: the sender is certainly capable of choosing the identity experiment, and inducing $\Sigma I = B$ (in figure 14 this would correspond to point A); this is the best (in the sense of being Blackwell-maximal) that the sender can induce. Since the sender can also choose any less informative experiment, it would seem that the sender may be capable of inducing *any* Blackwell-inferior distribution to A. Figure 11 shows that this intuition is false. A point like D is certainly Blackwell-inferior to A, being a mean-preserving contraction, yet it is outside the feasible set. The question then arises, why can I not simply "construct" the required experiment X as follows: suppose $\Sigma I \succeq B'$ and $p(B') = D$. If there exists an X with $\Sigma X = B'$, we would be done. What about simply putting $X = \Sigma^{-1} B'$? The answer is that *if* $p(\Sigma\Sigma^{-1}B')$ is in F, this would work. It turns out that if that it not true, then $\Sigma^{-1}B'$ will not yield a stochastic matrix X and therefore would not be a valid experiment (this can be seen by example). In other words, the sender is not capable of inducing any posterior belief that is Blackwell-inferior to ΣI.

Appendix B: Feasible Sets and The Blackwell Order

There are a number of interesting results that can illustrated using this technique of characterizing the feasible sets. To give but one example, I give a simple proof of a result first described in Bohnenblust, Shapley and Sherman (1949), and alluded to in Blackwell's original work (Blackwell (1951), Blackwell (1953)):

Theorem 5. *Suppose Σ_1 and Σ_2 are two garblings with $\Sigma_1 \succeq_B \Sigma_2$. Then $F(\Sigma_2, \pi) \subseteq F(\Sigma_1, \pi)$.*

Proof. Fix any π. I must show that for any τ if $supp(\tau) \in F(\Sigma_2, \pi)$, then $supp(\tau) \in F(\Sigma_1, \pi)$. By assumption I have that $p(\Sigma_2 X) = \tau$ for some X. The question is, does there exist a Y such that $\tau = p(\Sigma_1 Y)$? in other words, does there exist a Y such that $\Sigma_2 X = \Sigma_1 Y$? The answer is yes; by assumption I have that $\Gamma\Sigma_1 = \Sigma_2$ for some Γ. Thus,

$$\Sigma_2 X = \Sigma_1 Y \Rightarrow \Gamma\Sigma_1 X = \Sigma_1 Y \tag{38}$$

and therefore the required Y is given by

$$Y = \Sigma_1^{-1}\Gamma\Sigma_1 X \tag{39}$$

Note that Y does depend on both Σ_1 and X, as intuition would suggest. □

In other words, using a strictly more Blackwell-informative garbling results in a strictly larger set of feasible receiver posterior beliefs. Of course, this is obvious with trivial garblings (an identity, which would leave the feasible set unchanged from the Bayes-plausible one, and a completely uninformative garbling which would reduce the set to a single point - just the prior), but this theorem shows that the same "nesting" is true for nontrivial Blackwell-ranked garblings.

Figure 13 illustrates this observation using $\Sigma_1 = \begin{pmatrix} \frac{9}{10} & \frac{1}{100} \\ \frac{1}{10} & \frac{99}{100} \end{pmatrix}$ and $\Sigma_2 = \begin{pmatrix} \frac{2}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{3}{4} \end{pmatrix}$; it can be readily checked that $\Sigma_1 \succeq_B \Sigma_2$.

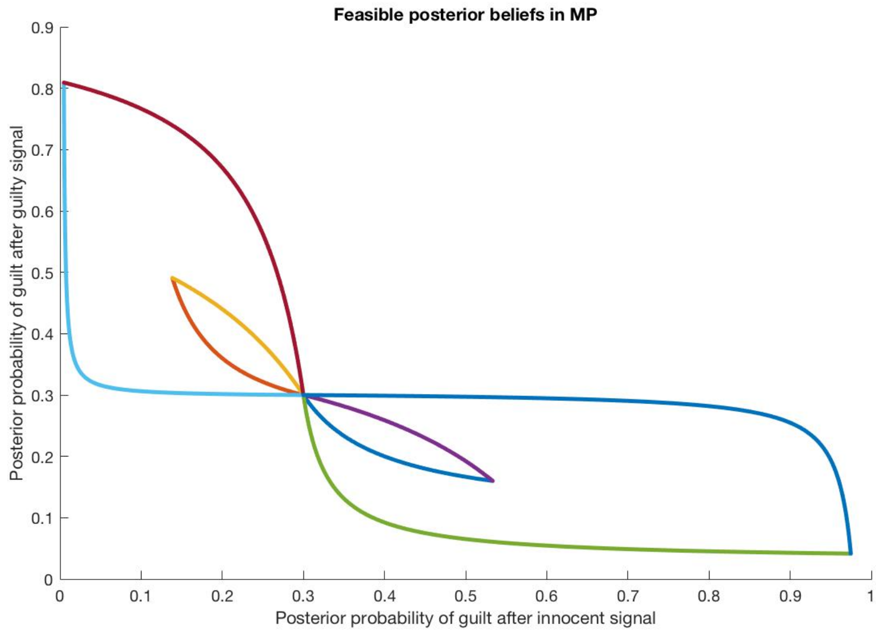

Figure 16: Blackwell's order implies set inclusion for feasible sets.

With "filled in" convex hulls the same idea is represented in figure 14.

Similarly, if Σ_1 and Σ_2 are not ranked by Blackwell's criterion, the F sets are not nested. I illustrate this by an example: consider $\Sigma_1 = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}$ and $\Sigma_2 = \begin{pmatrix} \frac{4}{5} & \frac{1}{2} \\ \frac{1}{5} & \frac{1}{2} \end{pmatrix}$.[20] The F sets are illustrated in figure 16.

With two states and three signals beliefs that were not feasible with two signals, become feasible.

Not shown all of the possible beliefs are shown (because the sets overlap), but rather the outer limits of the feasible sets and some of the feasible interior beliefs. The key observation from this experiment is that with three beliefs there are beliefs that can be induced, that cannot be induced with two signals. Namely, these are beliefs below 0.3 (this can be seen by comparing the relevant figures).

[20]it can be readily checked that these matrices are not ranked.

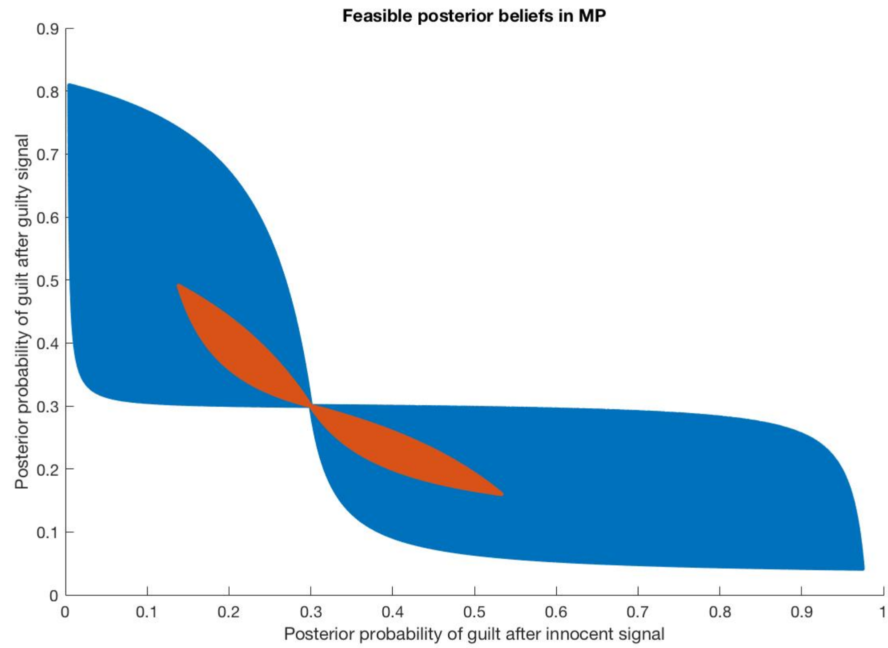

Figure 17: Further illustration of set inclusion.

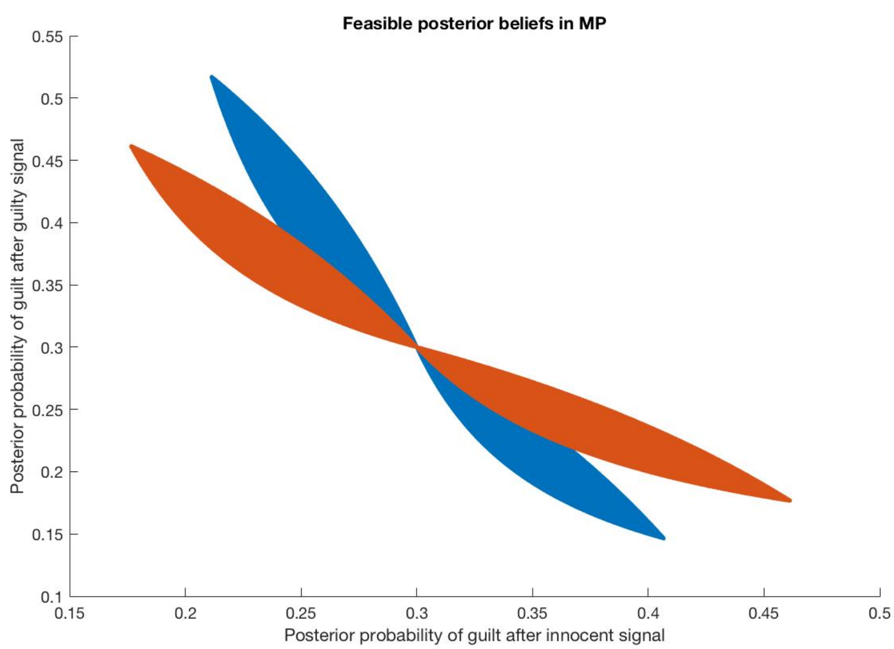

Figure 18: Unranked feasible sets.

7.2 Interpretation of the Rank of a Garbling Matrix

We now turn to a discussion of one of the key conditions established above - the necessity for Σ to be of full rank. This is a fairly straightforward question, yet it has never come up in the literature - what is the economic interpretation of the rank of a garbling matrix?

For simplicity suppose that the matrix is square, so that full rank guarantees invertibility. I first start with a discussion of what is means for a garbling matrix to *not* be invertible. By definition of rank, the column rank and the row rank of a matrix are always identical; recall also the convention that the columns of a garbling matrix represent signal realizations in each state of the world. If a matrix is not invertible, it means that there is at least one columns (a profile of signals in a given state) that is a linear combination of the other columns. In other words, one can *replicate the distribution of signals in a state without knowing anything about the state*. This embeds the definition of a Blackwell garbling.

The corresponding (row) point of view offers the same insight. If a garbling matrix is not invertible, then the distribution of a particular signal in all possible states is a linear combination of the distributions of the signals in the other states, and hence, one can replicate the distribution of a signal. In other words, a singular garbling contains within itself a sort of Blackwell garbling. Whether or not this internal garbling can be "undone", perhaps by constructing a new one, remains an open question[21]

This discussion sheds some light on the invertibility condition. The fact that the garblings used in the discussion of the feasible sets were all invertible means that they carry "as much information as possible", given their dimensional constraints.

Finally, suppose that the garbling is not square, i.e. Σ is a m-by-n matrix with m signals, n states and $m \geq n$.[22] Σ being full rank means that the rank is equal

[21] For example, given a garbling suppose that the receiver constructs another garbling from that, one that has full rank. What are the properties of this artificial garbling relative to the original one?

[22] Recall that the assumption that there are at least as many signals as states is made to avoid

to n (the most it can be), the number of states, which in turn implies that there always exists a *left* inverse. Observe that all of the inverses discussed so far were always used in left-multiplying the relevant matrices, so for non-square garblings the logic and algebra of being full rank is the same as the logic of invertibility for square matrices.

Appendix C: Proofs

Several of the results below rely on the Carathéodory theorem, a standard tool in economics, which I nevertheless state here without proof.

Theorem 6 (Carathéodory)**.** *Let* $X \subset \mathbb{R}^d$*, and let* $x \in \text{conv}(X)$*. Then there exist points* $x_0, \ldots, x_d \in X$ *and nonnegative coefficients* $\lambda_0, \ldots, \lambda_d$ *with* $\sum_{i=0}^{d} \lambda_i = 1$ *such that*

$$x = \sum_{i=0}^{d} \lambda_i x_i. \tag{40}$$

In particular, every point in the convex hull of X *can be expressed as a convex combination of at most* $d+1$ *points of* X*.*

Proof of proposition 1, on the complete characterization of canonical equilibria. Sufficiency: Proven in Theorem 1.

Necessity: suppose a canonical equilibrium (I, Σ^c) exists with outcome $\tau^{MP} = \{(\lambda, \beta_-^{ext}), (1-\lambda, \beta_+^{ext})\}$.

Conditions $i)$ through $iv)$ and $vi)$ follow from Theorem 2 (they are necessary for any equilibrium).

For condition $v)$: Since (I, Σ^c) is an equilibrium, the sender has no profitable deviation from $X = I$ given Σ^c. Therefore:

$$V^S(I, \Sigma^c) \geq V^S(X', \Sigma^c) \quad \text{for all } X' \in \mathcal{X} \tag{41}$$

some trivialities which arise when the signal space is not "rich enough".

Since $V^S(I, \Sigma^c) = \lambda v^S(\beta_-^{ext}) + (1-\lambda) v^S(\beta_+^{ext})$ and $V^S(X', \Sigma^c)$ depends only on $p(\Sigma^c X')$, which is exactly condition $v)$.

Thus, conditions $i)$ through $vi)$ are necessary for a canonical equilibrium. □

Lemma 1. *Fix any sender experiment X whose* entire *posterior-support $\{\beta_e\} \subset CB^M$. Then:*

1. The mediator's best response to X is any garbling that leaves each β_e unchanged (in particular, the identity), because for every $\beta \in CB^M$, cav $v^M(\beta) = v^M(\beta)$ *so no mean-preserving split can strictly increase his payoff.*

2. Consequently, every perfect-Bayesian equilibrium of the mediated game with sender's experiment X has $\Sigma^* = I$, and thus induces the same distribution of posteriors (and the same receiver-payoff) as the pure Bayesian-persuasion equilibrium with experiment X.

In particular, there is no strictly beneficial mediated equilibrium unless the sender can induce posteriors outside CB^M.

Proof. Since $\beta_e \in CB^M$ for every raw posterior of X,

$$v^M(\beta_e) = \text{cav } v^M(\beta_e). \tag{42}$$

Thus for each column e, any mean-preserving splitting τ_e of β_e achieves $\mathbb{E}_{\tau_e}[v^M] =$ cav $v^M(\beta_e) = v^M(\beta_e)$. In particular, the "no-splitting" choice (i.e. leaving β_e itself with probability 1) is a best-response in every column. Piecemealing these columns shows that the identity garbling $\Sigma(s \mid e) = \mathbf{1}\{s = e\}$ is a global best-response.

Hence in any mediated equilibrium with that X, the mediator uses $\Sigma^* = I$. But then the receiver sees exactly the same experiment as in the underlying Bayesian-persuasion game, so his equilibrium payoff is identical. No strict gain from mediation is possible unless the sender can produce some $\beta_e \notin CB^M$, so that splitting could strictly raise v^M. □

Lemma 2 (Column–wise decomposition, 2×2 case). *Let*

$$\Sigma \triangleq \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \tag{43}$$

be a column–stochastic matrix ($\sigma_{11} + \sigma_{21} = 1$, $\sigma_{12} + \sigma_{22} = 1$) and let the prior be $\beta_0 \in [0, 1]$. Let X be a 2×2 column–stochastic experiment that induces posterior beliefs $\{\beta_1, \beta_2\}$ with probabilities $\{\lambda, 1 - \lambda\}$, i.e.

$$\tau = \{(\beta_1, \lambda), (\beta_2, 1 - \lambda)\}, \qquad \lambda\beta_1 + (1 - \lambda)\beta_2 \triangleq \beta_0 \tag{44}$$

Then the composite experiment ΣX induces a posterior distribution supported on at most two beliefs $\{\beta_1', \beta_2'\}$, and these posteriors are given by column–wise Bayes updating:

$$\beta_k' = \frac{\sigma_{1k}\,\beta}{\sigma_{1k}\,\beta + \sigma_{2k}\,(1 - \beta)}, \qquad k = 1, 2, \tag{45}$$

for any posterior β produced by X. Moreover, the resulting posterior distribution is a Bayes–plausible convex combination:

$$\tau' \triangleq \{(\beta_1', \lambda'), (\beta_2', 1 - \lambda')\}, \qquad \lambda'\beta_1' + (1 - \lambda')\beta_2' \triangleq \beta_0 \tag{46}$$

In particular, every column of Σ acts as a fractional–linear map on posterior beliefs, and the composite ΣX can be represented as a convex combination of at most two such images.

Proof. Obvious. □

Lemma 3. *Let X be a sender experiment whose raw–posterior support*

$$S_X \triangleq \{\beta_e : e \in \operatorname{supp}(E)\} \subseteq [0, 1] \tag{47}$$

satisfies

$$\beta_0 \in \operatorname{conv}(S_X) \qquad \textit{and} \qquad CB^M \subseteq \operatorname{conv}(S_X), \tag{48}$$

where β_0 is the prior and CB^M denotes the mediator's convex basin. Then, under any mediator best-response garbling Σ^*, the support of the induced posterior distribution $\Sigma^* X$ lies in CB^M; i.e.

$$\operatorname{supp}(\Sigma^* X) \ \subseteq\ CB^M \tag{49}$$

Proof. Fix X and let $S_X = \{\beta_e : e \in \operatorname{supp}(E)\}$ be the set of posterior beliefs induced by X prior to any mediator garbling. Since $\beta_0 \in \operatorname{conv}(S_X)$ and $CB^M \subseteq \operatorname{conv}(S_X)$, every belief $\beta \in CB^M$ can be written as a convex combination of points in S_X.

Now let Σ be any mediator garbling. In the two–state case, each column of Σ transforms posteriors by a fractional–linear mapping, so (by Lemma 2) the composite ΣX induces a posterior distribution supported on at most two beliefs, each lying in $\operatorname{conv}(S_X)$; thus

$$\operatorname{supp}(\Sigma X) \subseteq \operatorname{conv}(S_X) \tag{50}$$

Suppose, to obtain a contradiction, that Σ assigns positive probability to a posterior $\beta' \notin CB^M$. Write $\operatorname{supp}(\Sigma X) = \{\beta_1', \beta_2'\}$ (support size at most two). Since β_0 is preserved under garbling,

$$\lambda' \beta_1' + (1 - \lambda')\beta_2' = \beta_0 \tag{51}$$

for some $\lambda' \in [0, 1]$.

Because $CB^M \subseteq \operatorname{conv}(S_X)$, and v^M is convex on CB^M, any probability mass placed outside CB^M strictly reduces the mediator's expected utility relative to pushing that mass back into CB^M while preserving β_0. Hence there exists a garbling Σ' such that

$$\operatorname{supp}(\Sigma' X) \subseteq CB^M \qquad \text{and} \qquad \mathbb{E}_{\Sigma' X}[v^M] > \mathbb{E}_{\Sigma X}[v^M] \tag{52}$$

contradicting Σ being a best reply. Therefore any mediator best–reply garbling Σ^*

must satisfy

$$\mathrm{supp}(\Sigma^* X) \subseteq CB^M \tag{53}$$

□

Proof of theorem 3. We proceed in several steps; first stating and proving lemmata 4-7, and then proceding with the proof.

Any sender experiment X and mediator garbling Σ compose to a final information structure $B \triangleq \Sigma X$ (column-stochastic $m \times n$). Given prior π, each realized signal s induces a Bayesian posterior $\beta_B(s) \in \Delta(\Omega)$ and the induced *distribution of posteriors* is $\tau_B \triangleq \sum_s \Pr_B(s)\, \delta_{\beta_B(s)}$, which is Bayes-plausible: $\mathbb{E}_{\beta \sim \tau_B}[\beta] = \pi$. Write $V^i(B) \triangleq \mathbb{E}_{\beta \sim \tau_B}[v^i(\beta)]$.

Lemma 4 (Column-wise decoupling)**.** *Fix X; let its columns be indexed by $e \in E$ with probabilities $w_e = \Pr(e)$ and raw posteriors β_e. The mediator's problem*

$$\max_{\Sigma \text{ column-stochastic}} V^M(\Sigma X) \tag{54}$$

decouples across columns: for each e, M chooses a mean-preserving split τ_e with mean β_e that maximizes $\mathbb{E}_{\tilde{\beta} \sim \tau_e}[v^M(\tilde{\beta})]$. Thus

$$\max_{\Sigma} V^M(\Sigma X) = \sum_{e \in E} w_e\, (\mathrm{cav}\, v^M)(\beta_e) \tag{55}$$

Proof. Because Σ is column-stochastic, each column e maps to a distribution over mediator signals *independently* of other columns. Bayes plausibility within a column enforces $\mathbb{E}[\tilde{\beta} \mid e] = \beta_e$. The objective is linear across columns, yielding the sum of independent concavifications. □

Lemma 5 (Support bound)**.** *For Ω with $|\Omega| = n$, the simplex $\Delta(\Omega)$ has affine dimension $n-1$. Hence every mean-preserving split τ_e with mean β_e can be supported on at most n posterior points in $\Delta(\Omega)$.*

Proof. Carathéodory's theorem: any point in the convex hull of a set in $\mathbb{R}^d$ is a convex combination of at most $d+1$ points. Here $d = n-1$. □

Lemma 6 (Extreme-point support of concavification)**.** *Fix a continuous $u : \Delta(\Omega) \to \mathbb{R}$. For any β, $(\operatorname{cav} u)(\beta)$ equals the maximum of $\mathbb{E}[u(\tilde{\beta})]$ over splits τ with mean β, and there exists an optimal τ supported on extreme points of the upper contact set $\{\gamma : u(\gamma) = (\operatorname{cav} u)(\gamma)\}$. Moreover, by Lemma 5 the support can be chosen of size at most n.*

Proof. This is the standard concavification result. Existence of an optimal split supported on the contact set follows from the upper supporting hyperplane (that supports the feasible posterior set) characterization: where $u = \operatorname{cav} u$, touching points suffice. Reduction to extreme points follows by an application of the Carathéodory theorem. □

Lemma 7 (Extension of mixed strategies)**.** *Any mixed strategy of S (a lottery over experiments) and M (a lottery over garblings) is outcome-equivalent to a single pair $(\bar{X}, \bar{\Sigma})$ that is column-stochastic on enlarged signal spaces, i.e., there exists $\bar{X} \in \mathcal{X}$, $\bar{\Sigma} \in \mathcal{S}$ such that $\bar{\Sigma}\,\bar{X} = \sum_{j,\ell} \eta_j \theta_\ell \Sigma^j X^\ell$.*

Proof. Let S mix over $\{X^\ell\}_\ell$ with probabilities $\{\theta_\ell\}$ and M over $\{\Sigma^j\}_j$ with $\{\eta_j\}$. Define a block-diagonal experiment $\bar{X}$ formed by $(X^1, \ldots, X^L)$, and the concatenated garbling $\bar{\Sigma} \triangleq [\eta_1 \Sigma^1 \ \cdots \ \eta_J \Sigma^J]$. Both are column-stochastic by construction and $\bar{\Sigma}\bar{X} = \sum_{j,\ell} \eta_j \theta_\ell \Sigma^j X^\ell$. Hence the induced distributions of posteriors coincide. □

We now finally come to the proof of the theorem. Step 1 (Mediator's best response to $X = I_n$ is Σ^c). Let $X = I_n$. Then the raw posterior support is $\{\delta_\omega\}_{\omega \in \Omega}$ with weights $\pi(\omega)$ (degenerate posteriors at each state). By Lemma 4 and Lemma 6, M attains $\sum_\omega \pi(\omega)(\operatorname{cav} v^M)(\delta_\omega)$ by choosing, for each state ω, a split supported on $\mathcal{E}(\mathrm{CB}^M)$. Condition (C2*) gives a *common* family of extreme points $\{\beta_i^{\mathrm{ext}}\}_{i=1}^r$

and weights λ with $\sum_i \lambda_i \beta_i^{\text{ext}} = \pi$. The canonical likelihood specification in (C4*)

$$\Pr(s_i \mid \omega) = \frac{\lambda_i \, \beta_i^{\text{ext}}(\omega)}{\pi(\omega)} \tag{56}$$

is column-stochastic (for each fixed ω, $\sum_i \Pr(s_i \mid \omega) = 1$), and Bayes' rule yields the posterior β_i^{ext} upon observing s_i with unconditional probability λ_i:

$$\Pr(\omega \mid s_i) = \frac{\Pr(s_i \mid \omega) \, \pi(\omega)}{\sum_{\omega'} \Pr(s_i \mid \omega') \, \pi(\omega')} = \frac{\lambda_i \beta_i^{\text{ext}}(\omega)}{\lambda_i} = \beta_i^{\text{ext}}(\omega). \tag{57}$$

Thus $\Sigma^c I_n$ induces exactly $\tau^{MP} = \{(\lambda_i, \beta_i^{\text{ext}})\}$.

By (C3*) this mixture maximizes M's value among all representations of π by extreme points of CB^M. By (C6*) (strict convexity along some exposed direction / uniqueness of a support) there is no alternative π-representation (nor an infinitesimal perturbation) that preserves the mean and strictly raises $\mathbb{E}[v^M]$. Therefore Σ^c is a best-response to I_n.

Finally, if M were allowed to mix over garblings, Lemma 7 implies this is equivalent to a single (larger) garbling, which cannot beat the concavification value either; hence Σ^c remains optimal under mixing.

Step 2 (Sender's best response to Σ^c is $X = I_n$). Fix Σ^c. By Lemma 7 it suffices to consider a single experiment X (possibly with more rows). The induced feasible set of posterior distributions is $F(\Sigma^c, \pi) = \{p(\Sigma^c X) : X \in \mathcal{X}\}$, which is convex and compact.[23] Condition (C5*) posits that the sender's expectation $\mathbb{E}[v^S(\beta)]$ is maximized over $F(\Sigma^c, \pi)$ exactly at $\tau^{MP} = \{(\lambda_i, \beta_i^{\text{ext}})\}$, which is implemented by $X = I_n$; hence I_n is a best response. Again, allowing S to mix does not enlarge the value by Lemma 7.

Step 3 (Equilibrium and belief consistency). Steps 1–2 show mutual best replies at (I_n, Σ^c). Beliefs at each signal are Bayesian by construction, so sequential ratio-

[23]Convex because $\mathcal{X}$ is convex and the sender can randomize across experiments; compact by closedness and finite dimensionality.

nality of R reduces to her stage payoff taking $a^*(\beta)$; these are already embedded into $v^i(\beta)$. Therefore (I_n, Σ^c) is a perfect Bayesian equilibrium.

Step 4 (Blackwell improvement). Because M only garbles S's signal, the final structure $\Sigma^c X$ is a Blackwell contraction of X. Conversely, in the canonical profile S uses $X = I_n$, which is Blackwell-maximal, so τ^{MP} is at least as informative[24] as any two-stage (ΣX) with the same π; in particular, $\tau^{MP} \succeq_B \tau^{BP}$. □

[24]This is essentially by asumption, because we are considering cases where the information structure under consideration Blackwell-dominates the BP outcome.

References

[1] Alonso, Ricardo and Odilon Camara. (2016). "Bayesian Persuasion with Heterogeneous Priors." *Journal of Economic Theory*, Volume 165, pp 672-706, http://dx.doi.org/10.1016/j.jet.2016.07.006.

[2] Ambrus, A., Azevedo, E. M. and Kamada, Y. (2013). "Hierarchical Cheap Talk." *Theoretical Economics*, 8: 233-261. doi:10.3982/TE1038

[3] Ariely, Itai, Yakov Babichenko, and Fedor Sandomirskiy. (2022). Bayesian Persuasion with Mediators. arXiv:2203.04285v2. https://doi.org/10.48550/arXiv.2203.04285

[4] Aumann, Robert and Michael Maschler. (1995). *Repeated Games with incomplete information.* MIT Press.

[5] Ben-Porath, Elchanan and Dekel, Eddie. (1992). "Signaling Future Actions and the Potential for Sacrifice". *Journal of Economic Theory*, 57(1), pp. 36-51.

[6] Blackwell, David. (1951). "Comparison of Experiments." Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 93-102, University of California Press, Berkeley, Calif.

[7] Blackwell, David. (1953). "Equivalent Comparisons of Experiments." *Ann. Math. Statist*. 24, no. 2, 265-272.

[8] Blume, Andreas, Oliver J. Board and Kohei Kawamura. (2007). "Noisy Talk." *Theoretical Economics*, 2(4), pp.395-440.

[9] Blume, Andreas. (2012). "A class of strategy-correlated equilibria in sender-receiver games." *Games and Economic Behavior*, 75(2), pp. 510-517.

[10] Bohnenblust, H.F., L.S. Shapley and S. Sherman. (1949). "Reconnaissance in Game Theory". US Air Force Project RAND Research Memorandum 208.

[11] Coval, Joshua, Jakub Jurek, and Erik Stafford. (2009). "The Economics of Structured Finance." *Journal of Economic Perspectives*, 23 (1): 3-25. DOI: 10.1257/jep.23.1.3

[12] Ely, Jeffrey C. (2017). "Beeps." *American Economic Review*, 107 (1): 31-53.

[13] Geanakoplos, John. (1989). "Game Theory without Partitions, with Applications to Speculation and Consensus". Cowles Foundation Discussion Paper No.914.

[14] Gentzkow, Matthew and Emir Kamenica. (2017a). "Competition in Persuasion." *Review of Economic Studies*, 84(4), pp. 300-322.

[15] Gentzkow, Matthew and Emir Kamenica. (2017b). "Bayesian Persuasion with Multiple Senders and Rich Signal Spaces." *Games and Economic Behavior*, 104, pp. 411-429, https://doi.org/10.1016/j.geb.2017.05.004.

[16] Hedlund, Jonas, (2017). "Bayesian Persuasion by a Privately informed Sender." *Journal of Economic Theory*, 167(C), pp. 229-268

[17] Ivanov, Maxim. (2010). Communication via a strategic mediator. *Journal of Economic Theory* 145(2), pp. 869-884. https://doi.org/10.1016/j.jet.2009.08.001

[18] Kamenica, Emir, and Matthew Gentzkow. (2011). "Bayesian Persuasion." *American Economic Review,* 101(6): 2590-2615.

[19] Kosenko, Andrew. (2022). "Constrained Persuasion with Private information." Forthcoming at *The B.E. Journal of Theoretical Economics*, https://doi.org/10.1515/bejte-2021-0017

[20] Kleiner, A., Moldovanu, B. and Strack, P. (2021), Extreme Points and Majorization: Economic Applications. *Econometrica*, 89: 1557-1593. https://doi.org/10.3982/ECTA18312

[21] Li, Fei and Norman, Peter. "Sequential Persuasion." *Theoretical Economics* 16, 639–675. DOI: 1555-7561/20210639

[22] Lipnowski, Elliot, Doron Ravid, and Denis Shishkin. (2022). "Persuasion via Weak Institutions" *Journal of Political Economy* 130:10, 2705-2730, https://doi.org/10.1086/720462 (2018). "Persuasion via Weak institutions". Working paper. Available at SSRN: https://ssrn.com/abstract=3168103 or http://dx.doi.org/10.2139/ssrn.3168103

[23] Pollrich, Martin. (2017). "Mediated audits." *RAND Journal of Economics*, 48: 44-68. doi:10.1111/1756-2171.12167

[24] Rahman, David and ichiro Obara. (2010). "Mediated Partnerships." *Econometrica*, 78: 285-308. doi:10.3982/ECTA6131

[25] Rubinstein, Ariel. (1997). *Modeling Bounded Rationality.* MiT Press, 1997.

[26] Samuelson, Larry. (2004). "Modeling Knowledge in Economic Analysis," *Journal of Economic Literature*, vol. 42, pp. 367-403.

[27] Sah, Raaj Kumar, and Joseph E. Stiglitz. (1986). "The Architecture of Economic Systems: Hierarchies and Polyarchies." *American Economic Review*, 76(4), pp. 716-727

[28] Tsakas, Elias and Nicolas Tsakas. (2021). "Noisy Persuasion" *Games and Economic Behavior* 130, pp. 44-61, https://doi.org/10.1016/j.geb.2021.08.001

[29] Le Treust, Maël and Tristan Tomala (2019). "Persuasion with Limited Communication Capacity" *Journal of Economic Theory* 184, https://doi.org/10.1016/j.jet.2019.104940

[30] Zapechelnyuk, Andriy. (2022). Sequential Obfuscation and Toxic Argumentation. Working Paper.