EXPERIMENTAL PERSUASION

By

Ian Ball and José-Antonio Espín-Sánchez

August 2021

COWLES FOUNDATION DISCUSSION PAPER NO. 2298

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

Experimental Persuasion

Ian Ball* José-Antonio Espín-Sánchez†

August 16, 2021

Abstract

We introduce experimental persuasion between Sender and Receiver. Sender chooses an experiment to perform from a feasible set of experiments. Receiver observes the realization of this experiment and chooses an action. We characterize optimal persuasion in this baseline regime and in an alternative regime in which Sender can commit to garble the outcome of the experiment. Our model includes Bayesian persuasion as the special case in which every experiment is feasible; however, our analysis does not require concavification. Since we focus on experiments rather than beliefs, we can accommodate general preferences including costly experiments and non-Bayesian inference.

*Department of Economics, MIT, ianball@mit.edu.
†Department of Economics, Yale University, jose-antonio.espin-sanchez@yale.edu.

1 Introduction

In Bayesian persuasion (Kamenica and Gentzkow, 2011, henceforth BP), Sender can design any information structure for Receiver. An information structure is a Blackwell experiment, which specifies in each state a probability distribution over signal realizations. Sender publicly selects one of these experiments and Nature draws the realization. Receiver sees the realization, updates his beliefs about the state, and then chooses an action. The interpretation of these experiments varies with the application, but Sender must have access to all experiments. Otherwise, the elegant BP analysis does not apply.

In many applications, however, the set of feasible experiments is institutionally restricted. Pharmaceutical companies seeking FDA approval must follow standard clinical trial protocols, though they do have flexibility in selecting the sample size, allowing them to perform a "sample size samba" (Schulz and Grimes, 2005, p. 1351). When schools select grading standards (Boleslavsky and Cotton, 2015; Ostrovsky and Schwarz, 2010), they may choose among coarse letter grades and various numerical systems, not arbitrary stochastic maps from performance to grades. Even in court—the leading example in Kamenica and Gentzkow (2011)—the prosecutor is restricted: only certain evidence is admissible, and each witness will be cross-examined by the defense.

We introduce *experimental persuasion* (henceforth EP). Sender has access to a feasible set of experiments and selects one experiment from this feasible set to perform. Receiver observes the realization of this experiment and chooses an action. In our model, the feasible set of experiments is a primitive that reflects institutional and technological constraints.

With restricted experiments, Sender cannot induce every Bayes-plausible distribution of posteriors, so concavification (Aumann et al., 1995; Kamenica and Gentzkow, 2011) does not apply. We therefore focus directly on the space of experiments, rather than the space of beliefs. Each experiment determines Receiver's posterior beliefs and hence his optimal actions, which generate Sender's payoffs. Composing this sequence, we obtain an indirect utility function over experiments. We believe this presentation provides a new intuition for persuasion, particularly in the binary setting, where we illustrate the solution by plotting the indifference curves of the indirect utility function. Our approach also accommodates more general preferences such as costly

experiments for Sender or non-Bayesian decision rules for Receiver.

We analyze two different regimes. In the first regime, Sender cannot garble the outcome of the experiment. This regime is appropriate if the outcome of the experiment is publicly observed. In this regime, the feasible set does not have a special structure. We illustrate the solution graphically in examples and observe that Sender's preferences are not monotone with respect to Blackwell's order. This setting reduces to Bayesian persuasion if every experiment is feasible.

In the second regime, Sender can garble the outcome of the experiment. Receiver observes only the realization of the garbled experiment. This garbling regime reduces to Bayesian persuasion if a perfectly revealing experiment is available, for then Sender can garble the state arbitrarily. In the garbling regime, we first show that Sender's problem has a solution (Theorem 1). Then we establish properties of the solution. Unlike in the no-garbling setting, Sender weakly prefers Blackwell more-informative primitive experiments since they can always be garbled into less informative experiments. Therefore, Sender can restrict attention to the subclass of Blackwell undominated primitive experiments. If there are multiple undominated experiments, then the set of synthetic experiments Sender can induce via garbling is not convex (Theorem 2). Nevertheless, we show that optimal persuasion satisfies a form of Kamenica and Gentzkow's (2011) indifference result: optimal persuasion generally makes Receiver indifferent between multiple actions (Theorem 3).

Bayesian persuasion has been extended in many different directions, as surveyed in Kamenica (2019). These extensions generally retain the assumption that all experiments are feasible. An exception is Haghtalab et al. (2021), in which Sender has access to multiple noisy signals about the state and she can disclose exactly one of these signals. Henry and Ottaviani (2019) show that observing a particular Brownian motion until a stopping time can achieve the BP optimum in a binary setting. In contrast, we analyze the general problem of persuasion from a restricted set of experiments.

2 Model

2.1 Setting

Sender (she) and Receiver (he) have state-dependent preferences over the action chosen by Receiver. The action space A and the state space Ω are both finite, with typical elements a and ω, respectively. Payoffs are given by $v(a, \omega)$ for Sender and $u(a, \omega)$ for Receiver.

Sender and Receiver share a common prior p over the state space Ω. Labeling the states $\omega_0, \ldots, \omega_{I-1}$, we can express the prior p as a vector $(p_0, \ldots, p_{I-1})$. Neither player observes the state realization. Sender can generate information about the state by performing an experiment. Without loss, we restrict attention to experiments that generate at most J signal realizations, where $J \geq \max\{|\Omega|, |A|\}$; see footnote 3 for a discussion of this restriction. Signal realizations are denoted $s_0, \ldots, s_{J-1}$. An *experiment* π is a stochastic $I \times J$ matrix,[1] where π_{ij} is the probability that signal s_j is realized, given that the state is ω_i. Let Π_0 denote the set of all experiments. Sender has access to a closed subset Π of Π_0. She selects from the feasible set Π one experiment to perform.

Receiver observes the realization of the experiment and then updates his beliefs. Given posterior belief $q \in \Delta(\Omega)$, Receiver chooses an action from the set

$$a^*(q) = \operatorname*{argmax}_{a \in A} \sum_i q_i u(a, \omega_i).$$

Assume that Receiver breaks ties in Sender's favor: if there are multiple actions in $a^*(q)$, Receiver chooses one that maximizes Sender's utility. Denote this action by $\hat{a}(q)$.[2]

2.2 Induced preferences over experiments

We now define Sender's induced preferences over experiments. First we introduce notation for Bayesian updating. Under experiment π, the ex ante probability of

[1] A matrix is *stochastic* if its entries are nonnegative and each row sums to 1.

[2] If there are multiple actions in $a^*(q)$ that maximize Sender's utility, then $\hat{a}(q)$ can be chosen among them arbitrarily.

signal realization s_j is given by

$$\bar{q}_j(\pi) = \sum_i p_i \pi_{ij}.$$

Upon seeing s_j, Receiver assigns to state ω_i probability

$$q_i^j(\pi) = \frac{p_i \pi_{ij}}{\bar{q}_j(\pi)}. \tag{1}$$

Denote the full belief vector following realization s_j by $q^j(\pi) = (q_0^j(\pi), \ldots, q_{I-1}^j(\pi))$. The beliefs in (1) depend on the experiment π and also on the prior p, but our notation suppresses the dependence on p when the prior is understood.

Sender's indirect utility function over experiments, $V\colon \Pi_0 \to \mathbf{R}$, is defined by

$$V(\pi) = \sum_{i,j} p_i \pi_{ij} v(\hat{a}(q^j(\pi)), \omega_i). \tag{2}$$

This expression is an expectation over realized state-signal pairs (ω_i, s_j). Sender's *direct* utility for each pair depends on the state ω_i and Receiver's chosen action $\hat{a}(q^j(\pi))$. Receiver's utility over experiments, $U\colon \Pi_0 \to \mathbf{R}$, is defined analogously, with u in place of v in (2).

While we focus directly on experiments, Kamenica and Gentzkow (2011) focus on beliefs. They define Sender's utility over beliefs, $\hat{v}\colon \Delta(\Omega) \to \mathbf{R}$, by

$$\hat{v}(q) = \sum_i q_i v(\hat{a}(q), \omega_i).$$

Receiver's posterior belief q determines his action $\hat{a}(q)$ and also pins down the conditional distribution of the state (since Receiver's beliefs are correct). By grouping the summation in (2) by the signal realizations s_j, we can express Sender's experiment utility function V as an expectation over the belief utility function $\hat{v}$:

$$V(\pi) = \sum_j \bar{q}_j(\pi) \hat{v}(q^j(\pi)).$$

Kamenica and Gentzkow (2011) assume that Sender can perform any experiment. Thus, Sender can induce every Bayes-plausible distribution over posteriors. Sender's optimal value, as a function of the prior, is given by the concavification of $\hat{v}$—the

smallest concave function larger than $\hat{v}$. In our setting, Sender is restricted to performing experiments in the feasible set Π, so she cannot necessarily induce every Bayes-plausible distribution over posteriors. Hence, the concavification payoff is not necessarily achievable.

2.3 Sender's problem with and without garbling

We consider two different regimes. In the *no-garbling* regime, the timing is as follows. Sender selects one experiment π from the feasible set Π. Receiver observes the realization of π, updates his beliefs, and chooses an action. In the *garbling* regime, Sender commits upfront to the garbling that she will apply to her chosen experiment. After nature draws the signal realization, Receiver observes only the garbled signal.

We now formally state Sender's problem and we observe that a solution exists.

No garbling Sender performs one feasible experiment π from the set Π. Therefore, Sender's problem is

$$\begin{aligned} &\text{maximize} && V(\pi) \\ &\text{subject to} && \pi \in \Pi. \end{aligned} \tag{3}$$

This problem reduces to standard Bayesian persuasion if every experiment is feasible, i.e., $\Pi = \Pi_0$.

With garbling Sender faces two choices: (i) which experiment to select, and (ii) how to garble the chosen experiment. A *garbling* is a $J \times J$ stochastic matrix G. If signal s_j is realized in the chosen experiment, then the distribution of the garbled signal is given by the j-th row of G.

Let $\succeq$ denote Blackwell's (1951) informativeness order over experiments. For any experiments π, π' in Π_0, we have $\pi \succeq \pi'$ if and only if there exists a garbling matrix G such that $\pi G = \pi'$. For any experiment π in Π_0, the *upper set* $\uparrow \pi$ and *lower set* $\downarrow \pi$ are given by

$$\uparrow \pi = \{\psi \in \Pi_0 : \psi \succeq \pi\}, \qquad \downarrow \pi = \{\psi \in \Pi_0 : \psi \preceq \pi\}.$$

If Sender chooses a *primitive* experiment π from Π, then garbling allows her to generate any *synthetic* experiment in the lower set $\downarrow \pi$. Let $\downarrow \Pi = \bigcup_{\pi \in \Pi} \downarrow \pi$. This set contains every experiment that is Blackwell dominated by some experiment in the

feasible set Π. Sender's problem is

$$\begin{aligned} &\text{maximize} && V(\pi) \\ &\text{subject to} && \pi \in \downarrow\Pi. \end{aligned} \tag{4}$$

This setting reduces to Bayesian persuasion if $\downarrow\Pi = \Pi_0$. This equality holds if and only if the feasible set Π contains an experiment that fully reveals the state. In this garbling regime, we are implicitly applying the revelation principle to the set $\downarrow\Pi$ rather than the set Π.[3]

Theorem 1 (Existence)
Each problem (3) *and* (4) *has a solution.*

The tie-breaking assumption ensures that V is upper semicontinuous. The space Π_0 of all experiments is compact. By assumption, Π is closed, and the proof amounts to checking that $\downarrow\Pi$ is closed as well. The details are in Appendix A.

3 Binary persuasion: A geometric approach

We illustrate our setting through Kamenica and Gentzkow's (2011) leading binary example of a prosecutor persuading a judge to convict a defendant. There are two states: ω_0 (innocent) and ω_1 (guilty). Receiver (judge) has two actions: a_0 (acquit) and a_1 (convict). Receiver gets utility 1 from the right decision and utility 0 from the wrong decision. He convicts if he believes the defendant is more likely to be guilty than innocent. Sender (prosecutor) wants the defendant to be convicted. She gets utility 1 from conviction and 0 from acquittal, regardless of the state. The common prior is that the defendant is guilty with probability 0.3. In this binary setting we identify every belief with the probability of ω_1 (guilty), so $p = 0.3$.

[3] In the no-garbling regime, the restriction to experiments with at most $|A|$ signal realizations is without loss by the standard revelation principle. In the garbling regime, we apply the revelation principle to the set $\downarrow\Pi$ rather than the set Π. Consider any primitive set Π of experiments with arbitrarily many signal realizations. This primitive set induces a set $\downarrow\Pi$ of synthetic experiments, also with arbitrarily many signal realizations. By the revelation principle, we can replace $\downarrow\Pi$ with a set of synthetic experiments $\operatorname{rev}\downarrow\Pi$ such that experiments in $\operatorname{rev}\downarrow\Pi$ each have at most $|A|$ signal realizations and the set of outcome distributions that Sender can induce is the same under $\downarrow\Pi$ and $\operatorname{rev}\downarrow\Pi$. We must apply the revelation principle after including the garblings since it is not necessarily true that $\operatorname{rev}\downarrow\Pi = \downarrow\operatorname{rev}\Pi$.

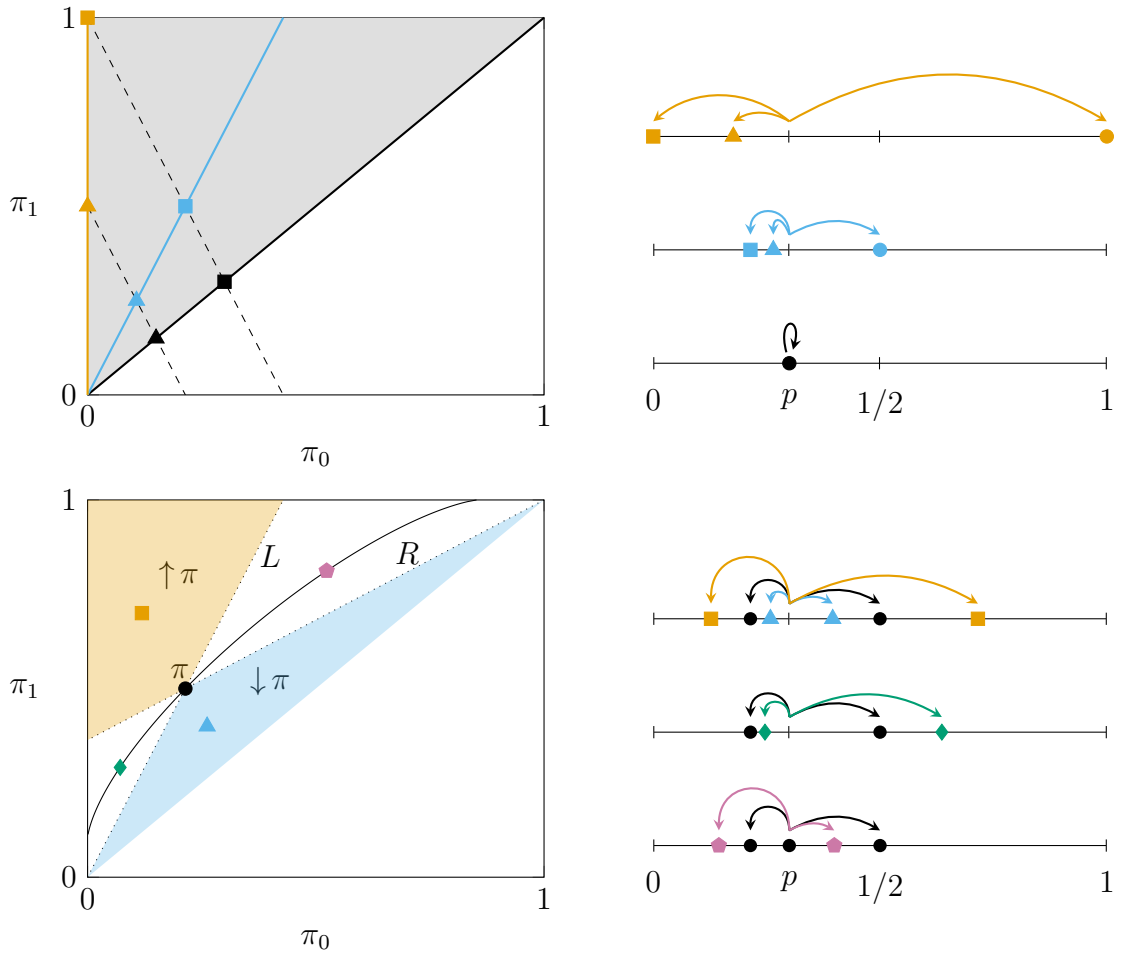

Figure 1. *Top*: Space of experiments (left) and induced beliefs (right). *Bottom*: Blackwell upper and lower sets for experiment π (left) and induced beliefs (right).

3.1 Experiments and induced beliefs

There are two actions and two states, so we restrict attention to experiments with $J = 2$ realizations. An experiment with two realizations, s_0 and s_1, is characterized by the probability of realization s_1 in each state. Denote by π_0 and π_1 the probabilities of realization s_1 in state ω_0 and state ω_1, respectively.[4] Without loss, we assume $\pi_1 \geq \pi_0$; otherwise, swap the labels of the two signal realizations. In (π_0, π_1)-space, the set of experiments is the subset of the unit square on or above the 45-degree line. This region is shaded in the top left panel of Figure 1. In the language of simple

[4]In the notation of the general setting, the vector (π_0, π_1) is column π^1 of the 2×2 stochastic matrix π.

hypothesis testing, if state ω_0 is interpreted as the null hypothesis, state ω_1 is the alternative hypothesis, and signal s_1 is rejecting the null, then the experiments are parameterized by the significance level $\pi_0 = \alpha$ and the power $\pi_1 = 1 - \beta$.

In this binary setting, Bayesian updating can be expressed cleanly in terms of the likelihood ratio between state ω_1 and ω_0. The likelihood ratios after signal realizations s_1 and s_0 are given by[5]

$$\ell_1 = \frac{p}{1-p} \cdot \frac{\pi_1}{\pi_0}, \qquad \ell_0 = \frac{p}{1-p} \cdot \frac{1-\pi_1}{1-\pi_0}. \tag{5}$$

Since $\pi_1 \geq \pi_0$, we have

$$\ell_1 \geq \frac{p}{1-p} \geq \ell_0.$$

Signal s_1 is evidence weakly in favor of state ω_1, while signal s_0 is evidence weakly against state ω_1.

In Figure 1, the top left panel shows the set of experiments and the top right panel shows the posteriors induced by a few highlighted experiments, for the fixed prior $p = 0.3$. In an experiment π, signal s_1 is realized with ex ante probability $p\pi_1 + (1-p)\pi_0$. The two dashed lines are level curves of this expression. For experiments (π_0, π_1) along any fixed line through the origin, the ratio π_1/π_0 is constant and equal to the slope of the line.[6]

We highlight three lines in particular. The black line has slope 1. Experiments along this line are uninformative and do not move Receiver's belief away from the prior p, as indicated in the bottom interval of the right panel. The blue line has slope $(1-p)/p = 7/3$. For any experiment along this line, signal realization s_1 induces posterior $q = 0.5$, denoted by the blue circle. On this line, the triangle (square) experiment splits the prior between the posterior $q = 0.5$ and a posterior q below p that is denoted by a blue triangle (square). The orange line is vertical. For any experiment along this line, signal realization s_1 induces posterior $q = 1$, denoted by the orange circle.[7] On this line, the triangle (square) experiment splits the prior between the posterior $q = 1$ and a posterior q below p that is denoted by the orange triangle (square).

[5]The beliefs can be expressed in terms of the likelihood ratios: $q^j(\pi) = \left(\frac{1}{1+\ell_j}, \frac{\ell_j}{1+\ell_j}\right)$ for $j = 0, 1$.

[6]Similarly, along any line through the upper right endpoint $(1, 1)$, the ratio $(1-\pi_1)/(1-\pi_0)$ is constant and equal to the inverse of the slope.

[7]Similarly, for experiments in the upper line ($\pi_1 = 1$), the signal realization s_0 indicates that the state is ω_0 with certainty.

In Figure 1, the bottom panels illustrate the Blackwell order on the space of experiments. The Blackwell order is independent of the prior p. For the indicated experiment π, the upper set $\uparrow\pi$ is shaded orange and the lower set $\downarrow\pi$ is shaded blue. Any experiment in the upper set can be garbled into π; any experiment in the lower set is a garbling of π. The boundaries of the upper and lower sets are formed by two lines—line L passing through $(0,0)$ and line R passing through $(1,1)$. For experiments ψ above line L, signal s_1 induces a stronger belief in state 1 under experiment ψ than under π. For experiments ψ above line R, signal s_0 induces a stronger belief in state 0 under experiment ψ than under π. Each experiment above both lines Blackwell dominates π; each experiment below both lines is Blackwell dominated by π. The experiments that lie above one line and below the other are not Blackwell comparable with π.

The right panel shows the corresponding posterior splittings induced by the experiments. The top line segment illustrates the splitting from a Blackwell dominating experiment (orange square) and a Blackwell dominated experiment (blue triangle). The middle line segment shows the splitting from the green diamond experiment, and the bottom line segment shows the splitting from the pink pentagon experiment. Both the green diamond and pink pentagon experiments lie on the curve in the left panel, which traces out experiments that have the same mutual information with the state as does experiment π.[8]

3.2 Preferences over experiments

Each experiment determines Receiver's posterior beliefs, which in turn determine Receiver's actions, and hence the payoffs for Sender (and Receiver). By composing these effects, we obtain Sender's (and Receiver's) indirect preferences over experiments. From the likelihood ratios in (5), we see that under experiment π, signal realization s_1 induces Receiver to choose action a_1 if and only if

$$\frac{\pi_1}{\pi_0} \geq \frac{1-p}{p}. \tag{6}$$

[8]For each experiment ψ along this curve, we have $\mathbf{E}\, H(q(\pi)) = \mathbf{E}\, H(q(\psi))$, were $q(\pi)$ denotes the random posterior that takes value $q^j(\pi)$ with probability $\bar{q}_j(\pi)$, and H denotes the Shannon entropy:

$$H(q_0, \ldots, q_{I-1}) = \sum_i -q_i \log(q_i),$$

with log denoting the natural logarithm.

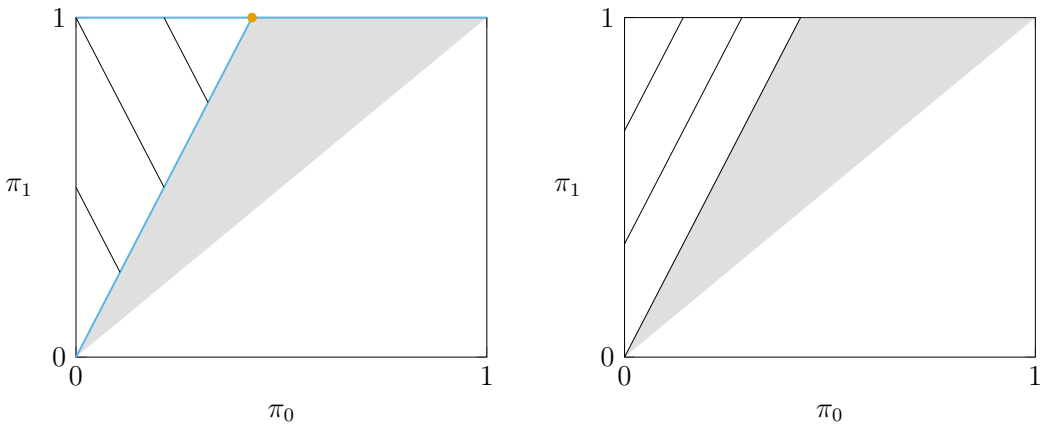

Figure 2. Indifference curves over experiments for Sender (left) and Receiver (right)

Sender's indirect utility function is

$$V(\pi) = \begin{cases} p\pi_1 + (1-p)\pi_0 & \text{if } \pi_1 p \geq \pi_0(1-p), \\ 0 & \text{otherwise.} \end{cases}$$

Receiver's indirect utility function is

$$U(\pi) = \begin{cases} p\pi_1 + (1-p)(1-\pi_0) & \text{if } \pi_1 p \geq \pi_0(1-p), \\ 1-p & \text{otherwise.} \end{cases}$$

In general, these utility functions are piecewise linear, with each piece defined by the vector of Receiver's chosen actions, $\big(\hat{a}(q^0(\pi)), \ldots, \hat{a}(q^{J-1}(\pi))\big)$.

Figure 2 (left) plots the level curves of this function V, with prior $p = 0.3$. Below the line with slope $(1-p)/p = 7/3$ is a thick indifference set, shaded in gray. Experiments in this region do not change Receiver's action. Above this line, Sender's utility V is linear. The indifference curves are parallel lines with slope $-(1-p)/p = -7/3$, with utility increasing towards the northeast. Figure 2 (right) plots the level curves of the function U, with prior $p = 0.3$. Again, there is a thick indifference set below the line of slope $7/3$. Above this line, indifference curves are parallel lines with utility increasing towards the perfectly revealing experiment $(0, 1)$.

3.3 Optimality conditions

Recall that standard BP maximizes Sender's utility V over the entire set of experiments. In Figure 2 (left), we can immediately read off the BP optimum as the orange point $(p/(1-p), 1) = (3/7, 1)$. This experiment always recommends conviction if the defendant is guilty and recommends conviction with probability $p/(1-p) = 3/7$ if the defendant is innocent. The total probability of conviction is therefore

$$(1-p)\left(\frac{p}{1-p}\right) + p = 2p = 0.6.$$

The general expressions in terms of p remains correct for any $p < 1/2$.

In Figure 2 (left), we indicate in blue the two necessary conditions for optimality from Kamenica and Gentzkow (2011).

1. *Certainty condition.* This condition states that Receiver is certain of the state whenever he takes Sender's worst action (Kamenica and Gentzkow, 2011, Proposition 4). Experiments satisfying this condition lie on the horizontal blue line defined by $\pi_1 = 1$. If the certainty condition is violated, Sender can get strictly higher payoffs by shifting the experiment upward, that is, recommending conviction slightly more often when the defendant is guilty. This increases the probability of the conviction signal and makes Receiver more confident in the defendant's guilt upon seeing that signal.

2. *Indifference condition.* This condition states that Receiver is indifferent between multiple actions whenever he takes Sender's preferred action (Kamenica and Gentzkow, 2011, Proposition 5). Experiments satisfying this condition lie on the blue line with slope 7/3. If this indifference condition is violated, Sender can get strictly higher payoffs by shifting the experiment slightly to the right, that is, recommending conviction slightly more often when the defendant is innocent. This increases the probability of the conviction signal without changing Receiver's willingness to convict upon receiving the signal.

The unique intersection of these two lines is the BP solution $(3/7, 1)$.

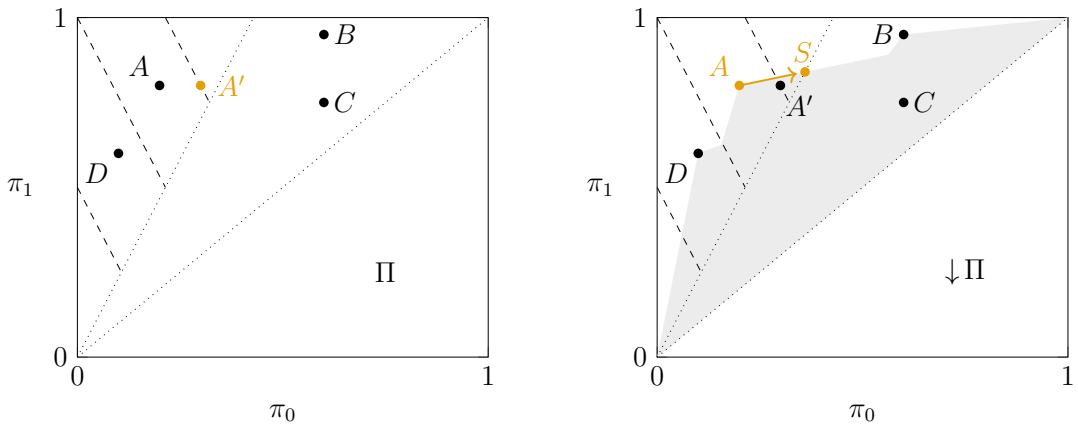

Figure 3. Optimal persuasion without garbling (left) and with garbling (right)

4 Optimal experimental persuasion

In this section, we describe optimal experimental persuasion with and without garbling.

4.1 No garbling

With no garbling, Sender maximizes her utility function V over the set Π of feasible experiments. The feasible set Π is not assumed to have any special structure (besides closedness), so we cannot hope to obtain a sharp characterization of the optimum.

As a running example, consider the binary environment from Section 3, now with the feasible set $\Pi = \{A, A', B, C, D\}$. This set is shown in Figure 3 (left) along with the indifference curves of Sender's indirect utility function. Experiments B and C lie below the indifference line, so they do not change Sender's default action. Experiments A, A', and D change Sender's action after realization s_1. The optimal experiment is A', which is highlighted in orange in the left panel. Experiment A' violates both the indifference condition and the certainty condition. The improving perturbations described in Section 3.3 are not feasible within the set Π.

Experiment A' is Blackwell dominated by experiment A, yet experiment A' gives Sender higher utility. Here, A and A' have the same value of π_1, but experiment A' would still give higher utility if it were slightly perturbed.

4.2 With garbling

Figure 3 (right) shows the same primitive set $\Pi = \{A, A', B, C, D\}$ together with the set $\downarrow\Pi$ of feasible synthetic experiments, shaded in gray. The Blackwell undominated experiments are A, B, and D. The dominated experiments A' and C do not affect the set $\downarrow\Pi$. The garbling solution is to perform the primitive experiment A and then garble the result to generate the synthetic experiment S. In the garbling regime, there is no loss of generality in removing all Blackwell dominated experiments. Any garbling of a Blackwell dominated experiment can be replicated by a suitable garbling of a Blackwell dominating experiment.

Throughout our analysis, the state space is fixed. Kamenica and Gentzkow (2011, p. 2598) explain that the state can be redefined as the outcome of an experiment:

> [I]t may seem restrictive to assume that Sender can generate signals that are arbitrarily informative. This assumption, however, is innocuous under an appropriate interpretation of the state ω. We can define ω to be the realization of the *most informative signal Sender can generate*. Then, his choice of π is equivalent to a choice of how much to garble information that is potentially available. [emphasis added]

This redefinition assumes that there exists a most informative experiment. Of course, if all the feasible experiments can be performed simultaneously, then there is necessarily a most informative (compound) experiment. But as long as there is some constraint on Sender's time or resources, there will likely be multiple Blackwell undominated experiments.[9]

Without a most informative experiment, Sender's domain of optimization is not necessarily convex, as is clear in Figure 3 (right). The next result gives an essential equivalence in the 2 by 2 setting. (Beyond the 2 by 2 setting, there is no canonical way to order the signal realizations, so mixtures of experiments are difficult to interpret.)

Theorem 2 (Convexity)
Suppose $I = J = 2$. If the set Π has a most informative experiment, then $\downarrow\Pi$ is convex. If the set Π does not have a most informative experiment and Π is finite, then $\downarrow\Pi$ is not convex.

[9] An alternative approach would be to define a sequence of auxiliary persuasion problems, one for each Blackwell undominated experiment. We could solve each problem separately and then compare the values of the solutions. This approach becomes unwieldy if the Blackwell frontier has many experiments, and it would not provide graphical intuition for the main persuasion problem.

Proof. First, suppose that Π has a most informative experiment $\hat{\pi}$. Then $\downarrow\Pi = \downarrow\hat{\pi}$. This set is convex because the space of garbling matrices is convex. (This direction of the theorem holds for arbitrary I and J.)

For the converse, suppose $I = J = 2$. Suppose that Π is finite and Π does not have a most informative experiment. Then there exist two experiments π and π' that are Blackwell undominated and Blackwell incomparable. Consider the open line segment (π, π') connecting π and π'. We claim that $\downarrow\pi \cap (\pi, \pi')$ and $\downarrow\pi' \cap (\pi, \pi')$ are both empty.

Now we complete the proof, assuming the claim. We have

$$(\pi, \pi') \cap \downarrow\Pi = \bigcup_{\pi'' \in \Pi} (\pi, \pi') \cap \downarrow\pi'' = \bigcup_{\substack{\pi'' \in \Pi \\ \pi'' \neq \pi, \pi'}} [\pi, \pi'] \cap \downarrow\pi'',$$

where the second equality follows from the claim. The right side is a closed set, being a finite union of closed sets, so it cannot equal (π, π'). Therefore, $\downarrow\Pi$ is not convex.

Now we prove the claim. By symmetry, it suffices to prove that $\downarrow\pi \cap (\pi, \pi')$ is empty. Suppose for a contradiction that there exists $\lambda \in (0, 1)$ such that $\pi \succeq \lambda\pi + (1 - \lambda)\pi'$. It is clear geometrically (and straightforward to check algebraically) that in the 2 by 2 setting, we have $\pi \succeq \psi$ if and only if ψ can be expressed as a convex combination of the three experiments $\mathbf{0} = (0, 0)$, $\mathbf{1} = (1, 1)$, and π. Therefore, there exist nonegative α and β with $\alpha + \beta \leq 1$ such that

$$\lambda\pi + (1 - \lambda)\pi' = \alpha\mathbf{0} + \beta\mathbf{1} + (1 - \alpha - \beta)\pi.$$

Solving for π' we have

$$\pi' = \frac{\alpha}{1 - \lambda}\mathbf{0} + \frac{\beta}{1 - \lambda}\mathbf{1} + \frac{1 - \alpha - \beta - \lambda}{1 - \lambda}\pi. \tag{7}$$

The coefficients on the right side sum to 1, but we must check that they are nonnegative. By our signal-labeling convention, $\pi'_1 \geq \pi'_0$ and $\pi_1 \geq \pi_0$. Since π' and π are undominated, both inequalities must be strict. Therefore, the last coefficient in (7) must be positive, and we conclude that $\pi \succeq \pi'$, which is a contradiction. □

To understand the structure of optimal persuasion with garbling, we seek necessary conditions for optimality. We could observe that the optimum is the solution to an auxiliary BP problem in which the state is redefined as the realization of a particular

primitive experiment. But then the hypotheses and conclusions would be stated in terms of the redefined state space. The interpretation would depend on which primitive experiment was performed to achieve the optimal synthetic experiment. Instead, we provide a different result in terms of the true state space.

To state the characterization result, we need a few definitions. Sender's ordinal preferences are *state-independent* if for all actions a and a' and states ω and ω', we have

$$v(a,\omega) \geq v(a',\omega) \iff v(a,\omega') \geq v(a',\omega').$$

If Sender's ordinal preferences are state-independent, then we can refer to better and worse actions for Sender, without reference to the state. Receiver's *default action* is his action $\hat{a}(p)$ at the prior p. An action a is *induced* by an experiment π if there exists j with $\bar{q}_j(\pi) > 0$ such that $\hat{a}(q^j(\pi)) = a$.

Theorem 3 (Indifference condition)
Suppose that Sender's ordinal preferences are state-independent and that Receiver does not take a best action by default. Under experimental persuasion with garbling, after each signal realization, Receiver either (i) takes the worst action among all the induced actions, or (ii) is indifferent between multiple actions.

Proof. Let π in $\downarrow \Pi$ be an optimal experiment in the garbling regime. Write $\pi = \bar{\pi}G$ for some primitive experiment $\bar{\pi}$ in Π and some garbling matrix G. To simplify notation, write $a_j = \hat{a}(q^j(\pi))$ for each j.

Suppose for a contradiction that for some j with $\bar{q}_j(\pi) > 0$, action a_j is not the worst among the induced actions and Receiver is not indifferent between multiple actions after signal s_j. Then there exists j' with $\bar{q}_{j'}(\pi) > 0$ such that Sender strictly prefers a_j to $a_{j'}$. Fix ε in $(0,1)$ and consider the garbling matrix $G' = \mathrm{id} + \varepsilon(e_{j'j} - e_{j'j'})$.[10] Set $\pi' = \pi G' = \bar{\pi}(GG')$. Since GG' is a garbling matrix, we know π' is in $\downarrow \pi$. For all ε sufficiently small, this modification leaves the Receiver's best response unchanged, and shifts positive probability from action $a_{j'}$ to action a_j, strictly increasing Sender's utility, contrary to the optimality of π. □

If no signal realization from any experiment in Π makes Receiver indifferent between multiple actions, then Theorem 3 implies that optimal persuasion must involve nontrivial garbling.

[10] Here, e_{ij} denotes the standard basis $J \times J$ matrix with (i,j)-element equal to 1 and all others equal to 0.

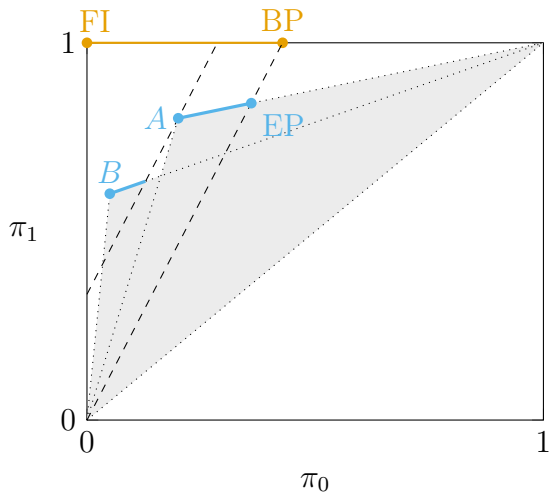

Figure 4. Pareto frontier in BP (orange) and EP (blue)

The conditions of Theorem 3 hold in the leading binary example of the prosecutor and the judge. Therefore, under optimal persuasion Receiver must be indifferent when he convicts the defendant. By contrast, the certainty condition does not hold, as can be seen in Figure 3. Whenever Receiver acquits, however, he is certain of the realization of the primitive experiment.

5 Concluding discussion

We conclude with a few applications that we can analyze because of our focus on experiments.

5.1 Welfare

We discuss the welfare implications of our analysis in the leading binary example. When the state is ω_1 (guilty), both Sender and Receiver prefer a_1 (convict), so there is no conflict of interest. When the state is ω_0 (innocent), Sender prefers a_1 (convict) and Receiver prefers a_0 (acquit). The social planner's preferred action in this state is determined by which player gets a higher Pareto weight.

Figure 4 shows the Pareto frontier in two different settings. Under BP, when all experiments are available, the Pareto frontier is the orange segment connecting the fully informative experiment (FI) to the BP optimum. In every experiment on

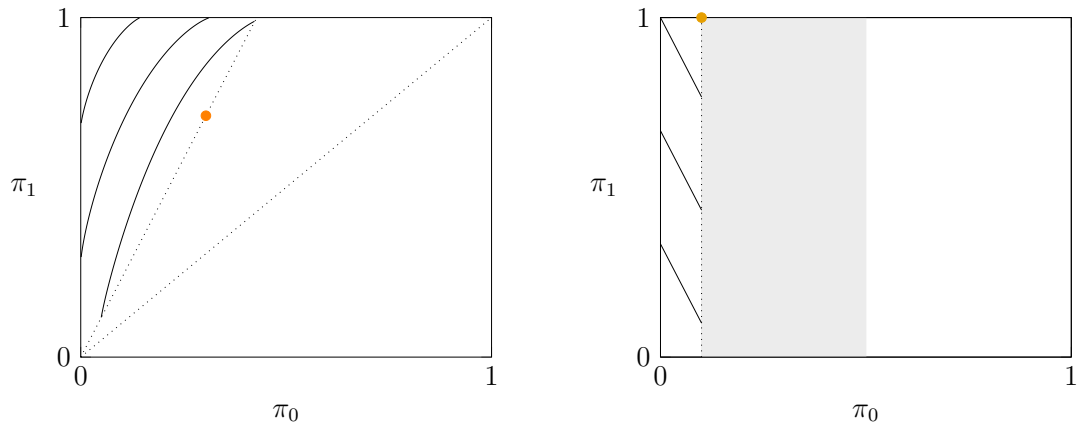

Figure 5. Sender's indifference curves for costly experiments (left) and frequentist decisions (right).

the Pareto frontier, action a_1 is taken in state ω_1. Sender prefers the BP optimum; Receiver prefers the fully informative experiment. If a social planner maximizes a social welfare function, then the optimum will be (i) full disclosure if Receiver gets more weight; (ii) the BP optimum if Sender gets more weight; and (iii) the entire Pareto frontier if the two players get the same weight.

Next, consider experimental persuasion (EP) with garbling when there are two primitive experiments, A and B. The Pareto frontier for this regime is shaded in blue. It consists of two different line segments. One segment consists of garblings of B. The other segment consists of garblings of A and it contains the optimum EP. The EP optimum is *second-best* efficient, but it is not *first-best* efficient because $\pi_1^{\text{EP}} < 1$. Here we see a potential pitfall of redefining the state—it can obscure what would be feasible if more experiments became available.

5.2 More general preferences

The key object of our analysis is Sender's indirect utility function over experiments. We have assumed that Sender's utility depends only on Receiver's action choices, and that Receiver selects his actions by performing Bayesian updating. But our approach allows for much more general preferences, as long as we can compute the indirect utility function V. In order to focus on alternative preference structures, we assume in this section that all experiments are feasible.

First, suppose that experiments are costly. Figure 5 (left) shows Sender's indif-

ference curves if each experiment π entails a mutual information cost

$$c(\pi) = 3[H(p) - \mathbf{E}\, H(q(\pi))].$$

See footnote 8 for the definition of entropy. In this case, we define a new indirect (net) utility function $V' = V - c$. For simplicity, we plot the indifference curves only for experiments that change Receiver's default action. Among experiments that do not change Receiver's default action, a completely uninformative experiment is optimal since it minimizes the cost function c. Sender's net utility is increasing towards the southeast. The optimal experiment is indicated by the orange circle. This experiment satisfies the indifference condition, but it is less costly than the BP solution.

Alternatively, consider a frequentist Receiver. This assumption captures the process of drug approval. Sender is a pharmaceutical company and Receiver is the FDA. The states are ω_0 (ineffective drug) and ω_1 (effective drug), and the actions are a_0 (reject) and a_1 (approve). An experiment (clinical trial) is defined by its significance π_0 and power π_1. In this case, it is more convenient to relabel the signal realizations so that $\pi_0 \leq 1/2$. This way, the set of experiments is represented as the left half of the unit square.[11] The FDA's policy for regulating clinical trials is based only on significance, not power (Isakov et al., 2019). Pharmaceutical companies want their drugs to be approved and understand this FDA policy. Moreover, they have flexibility in designing the experiment, through the sample size or stopping criterion (Schulz and Grimes, 2005).

Figure 5 (right) shows Sender's indirect utility function in this setting. The vertical line is given by $\pi_0 = \alpha$, where α is the FDA's significance threshold. An experiment π with $\pi_0 > \alpha$ will not be considered for approval, so the region to the right of this vertical line is a thick indifference set. Experiments left of this vertical line will be considered for approval. Sender's payoff is the ex ante probability that the null of ineffectiveness is rejected. The indifference curves are parallel lines, with utility increasing to the northeast. The optimal experiment for Sender is $(\alpha, 1)$, indicated by the circle.

Compare Sender's indifference curves for a non-Bayesian FDA in Figure 5 (right) with Sender's indifference curves with a Bayesian receiver in Figure 2 (left). Since

[11] Sender's utility is still defined for the entire unit square, but much of the square is redundant, provided that Receiver's decision rule is invariant to relabeling the signal realizations. It suffices to plot Sender's preferences over a subset S of $[0, 1]$ with the property that $[0, 1] = S \cup (\mathbf{1} - S)$.

Sender is Bayesian in both cases, her preferences are piecewise linear. For all experiments on a fixed line through the origin, a Bayesian *receiver* must take the same action after seeing s_1. This property can fail for a frequentist receiver. In Figure 5, for example, every line through the origin will cross the significance threshold.

The tension between frequentist and Bayesian approaches to drug approval has spawned a heated policy debate. Isakov et al. (2019) argue that different diseases and therapies require different trade-offs between type-I and type-II error. The FDA does have some leeway to incorporate these considerations. Recently, the FDA offered "accelerated approval" to Biogen's Alzheimer's drug Aduhelm, despite weak clinical evidence, because of the severity of Alzheimer's and the absence of alternative treatments (Belluck and Robbins, 2021). Our results can shed light on this debate by formalizing the trade-offs presented by firms' strategic responses to the FDA's drug approval criteria.

References

Aumann, R. J., M. B. Maschler, and R. E. Stearns (1995): *Repeated Games with Incomplete Information*, MIT Press. [2]

Belluck, P. and R. Robbins (2021): "F.D.A. Approves Alzheimer's Drug Despite Fierce Debate Over Whether It Works," *New York Times*, published June 7, 2021; updated July 20, 2021. [20]

Blackwell, D. (1951): "Comparison of Experiments," in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 93–102. [6]

Boleslavsky, R. and C. Cotton (2015): "Grading Standards and Education Quality," *American Economic Journal: Microeconomics*, 7, 248–79. [2]

Haghtalab, N., N. Immorlica, B. Lucier, M. Mobius, and D. Mohan (2021): "Persuading with Anecdotes," NBER Working Paper 28661. [3]

Henry, E. and M. Ottaviani (2019): "Research and the Approval Process: The Organization of Persuasion," *American Economic Review*, 109, 911–955. [3]

Isakov, L., A. W. Lo, and V. Montazerhodjat (2019): "Is the FDA too conservative or too aggressive?: A Bayesian decision analysis of clinical trial design," *Journal of Econometrics*, 211, 117–136. [19, 20]

Kamenica, E. (2019): "Bayesian Persuasion and Information Design," *Annual Review of Economics*, 11, 249–272. [3]

Kamenica, E. and M. Gentzkow (2011): "Bayesian Persuasion," *American Economic Review*, 101, 2590–2615. [2, 3, 5, 7, 12, 14]

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

Schulz, K. F. and D. A. Grimes (2005): "Sample Size Calculations in Randomised Trials: Mandatory and Mystical," *Lancet*, 365, 1348–1353. [2, 19]

A Proof of Theorem 1

First we prove that V is upper semicontinuous. Denote the mixed extension of v to $A \times \Delta(\Omega)$ by the same symbol v, and use analogous notation for u. We have

$$\begin{aligned} V(\pi) &= \sum_{i,j} p_i \pi_{ij} v(\hat{a}(q^j(\pi)), \omega_i) \\ &= \sum_j \bar{q}_j(\pi) \max_{a \in a^*(q^j(\pi))} v(a, q^j(\pi)). \end{aligned}$$

The belief $q^j(\pi)$ is defined only if $\bar{q}_j(\pi) \neq 0$, but the terms with $\bar{q}_j(\pi) = 0$ vanish anyway, so we can formally define the summation over realizations s_j with $\bar{q}_j(\pi) \neq 0$. The functions $(a, q) \mapsto v(a, q)$ and $(a, q) \mapsto u(a, q)$ are both continuous. By Berge's theorem, the correspondence $q \mapsto a^*(q)$ is upper hemicontinuous, and hence the function $q \mapsto \max_{a \in a^*(q)} v(a, q)$ is upper semicontinuous. Fix π such that $\bar{q}_j(\pi) > 0$ for all j. Since the map $\pi \mapsto q^j(\pi)$ is continuous, the entire sum is upper semicontinuous at π. To complete the proof, observe that if $\bar{q}_j(\pi) = 0$ for some j, then we can restrict the summation to those j for which $\bar{q}_j(\pi)$ is positive and repeat the argument. For any sequence π^n converging to π, the contribution of the excluded terms converges to 0 in the limit (since v is bounded).

The set Π_0 of all experiments is compact, so it suffices to check that the feasible set is closed. In the no-garbling regime, this is immediate by assumption. For the garbling regime, it remains to check that $\downarrow \Pi$ is also closed. Consider a sequence of experiments π^n in $\downarrow \Pi$ that converges to some experiment π in Π_0. For each n, write $\pi^n = \bar{\pi}^n G^n$ for some $\bar{\pi}^n$ in Π and some garbling matrix G^n. Since Π and the space of garblings are both compact, we can find a convergent subsequence $(\bar{\pi}^{n_k}, G^{n_k})$ of $(\bar{\pi}^n, G^n)$. Denote the limit by $(\bar{\pi}, G)$. Since matrix multiplication is jointly continuous,

$$\pi = \lim_{k \to \infty} \bar{\pi}^{n_k} G^{n_k} = \hat{\pi} G,$$

Hence π is in $\downarrow \Pi$.