PERSUASION WITHOUT PREJUDICE

Andriy Zapechelnyuk

Abstract. Evidence (a Blackwell experiment) is prejudiced in a party's favor if, in every state, it shifts the decision toward that party relative to no evidence — e.g., an arbitration exhibit that favors the plaintiff whether or not he is right, or an online rating system that raises sales irrespective of the product's quality. We define unprejudiced evidence and solve for optimal persuasion under this constraint. Optimal evidence under standard Bayesian persuasion is often prejudiced. Because weak evidence tends to be prejudiced, the constraint makes the persuader disclose more, not less: optimal unprejudiced evidence concedes the case outright or else increases the beliefs at least as far as unconstrained persuasion would.

JEL Classification: D82, D83, K41.

Keywords: Bayesian persuasion, information design, prejudice, robustness, disclosure

September 9, 2026

Zapechelnyuk: School of Economics, University of Edinburgh, 31 Buccleuch Place, Edinburgh, EH8 9JT, UK. *E-mail:* azapech@gmail.com.

1. Introduction

Institutions that decide on the basis of evidence rarely admit all of it. Courts exclude relevant evidence whose probative value is judged to be outweighed by the danger of unfair prejudice, and they bar whole categories of evidence, such as a defendant's prior convictions, on the same ground. Drug regulators require trial protocols to be registered before the trial is run, so that a sponsor cannot design a study whose results favor approval however the drug performs. Consumer-protection law restricts the rating systems that online platforms may run, and a rating that raises sales whatever the quality of the product is the kind of design such rules target. From a purely statistical standpoint this restriction is puzzling. A decision maker who is rational, and who understands where evidence comes from, cannot be hurt by seeing more of it: any information weakly raises her expected payoff. This paper tackles this puzzle by hypothesising that evidence is rejected not because it harms the decision maker's expected payoff, but because it systematically pushes the decision in the same direction whichever way the truth happens to lie.

Call evidence *prejudiced* in favor of a party if, conditional on each state of the world separately, it is expected to shift the decision in that party's favor, relative to the decision that would be taken with no evidence at all. Prejudice in this sense is neither an error by the decision maker nor a statistical bias in the evidence: beliefs are updated correctly, and the decision is right on average. It is the failure of a guarantee. The party against whom prejudiced evidence is offered can say, truthfully and before knowing anything about the facts: whatever the truth is, I am worse off for this having been shown. That is a complaint an institution may wish to honor even when the fact-finder is impeccably Bayesian.

Two stories, outlined in Section 2, illustrate what is at stake. In the first story, a plaintiff and a defendant bring a dispute to an arbitrator who must divide a pie between them. One of the two is right. The arbitrator would like to award everything to whoever is right, but she is risk averse and splits the pie when she is unsure. The plaintiff chooses what evidence to put before her. The defendant can neither verify the truth nor audit the plaintiff's motives; what he can ask for is a procedural guarantee — that the exhibit not be one that tilts the award toward the plaintiff whether or not the plaintiff is right. An arbitrator applying such a standard admits evidence that meets this guarantee and rules on her prior otherwise. The second story is about a manufacturer who sells a product of uncertain quality through an online platform that operates a quality screening system, such as a rating built from consumer reviews. The platform wants to maximize sales; consumers differ in their outside options, so the fraction who buy is an increasing function of their belief about quality. A screening system that raises sales whether the product is good or bad is seen as a marketing instrument rather than an informational one, and a regulator concerned with consumer protection may want it prohibited. In both stories, the prohibition is a constraint on the design of information, and the interesting question is what a sophisticated persuader does when he faces it.

In this paper we formalize what it means for evidence to be prejudiced, and we address two questions. First, how does the requirement of no prejudice constrain a persuader's ability to manipulate decisions through the control of information? Second, how, and when, does the resulting design differ from what standard Bayesian persuasion prescribes?

The model. The setup is deliberately simple. The state is either high or low, and the parties share a prior belief. An agent takes a decision on the unit interval, and her decision is an increasing function of her belief that the state is high. A principal benefits from a higher decision, and chooses a piece of evidence: any statistical experiment about the state. Evidence is prejudiced if the expected decision conditional on state 0 and the expected decision conditional on state 1 both exceed the decision the agent would take with no evidence; otherwise it is unprejudiced, and the principal is restricted to unprejudiced evidence. Two features of the definition deserve emphasis. First, the criterion is applied state by state, and each state is compared with the same no-evidence benchmark. It therefore does not weigh states by the prior, and in this sense it is a prior-free guarantee rather than an expected-utility calculation. Second, prejudice is a joint property of the evidence and of the decision rule, not a property of the evidence alone. Indeed, if the decision responded linearly to beliefs, no evidence could ever be prejudiced: in the unfavorable state, beliefs drift downward on average, so a linear decision must drift downward too. The scope for prejudice comes from the curvature of the decision rule. Prejudice, in other words, is manufactured out of the nonlinearity in how decisions respond to information, and this is why it can arise even though the Bayesian fact-finder makes no mistake.

Which evidence is prejudiced. Section 3 studies the constraint itself, before any optimization, and yields two messages.

The first is that whether prejudice is possible at all is settled by the shape of the decision rule: no evidence is prejudiced at any prior precisely when a simple concavity condition holds. In the arbitration story the condition holds exactly when the arbitrator is sufficiently risk averse, so that she under-reacts to news; a cautious arbitrator cannot be tilted in both states, and the admissibility rule is then vacuous. In the platform story the condition says that the distribution of consumer types has a sufficiently thick lower tail.

The second message concerns which pieces of evidence are unprejudiced irrespective of the prior and of the decision rule, and the answer is sharp: exactly the *smoking gun* — the type of evidence that pushes the decision in the principal's favor only through messages that prove the favorable state beyond doubt. Full disclosure is a special case. Anything that pushes the decision in the principal's favor on merely suggestive news is prejudiced for some prior or some decision rule. For institutional design this is a strong and somewhat uncomfortable statement: a blanket, context-free admissibility rule can license nothing beyond conclusive proof, so any workable standard must be applied in context, with reference to the prior and to how the fact-finder actually decides.

Persuasion under the constraint. Section 4 solves the persuader's problem. The headline result is that the constraint does not make the persuader more reticent; it makes him more forthcoming. Whenever the no-prejudice requirement binds, every optimal unprejudiced evidence is strictly more informative, in the Blackwell sense, than the evidence the same persuader would have chosen without the constraint. The reason is that weak evidence is what tends to be prejudiced. Unconstrained Bayesian persuasion works by pooling: it bundles part of the unfavorable state together with the favorable state so as to nudge the agent just past the point at which her decision moves, and that is precisely the design that helps the persuader in both states at once. To satisfy the constraint, the persuader must weaken the pooling.

The optimal design turns out to be one-sided, and is the mirror image of the smoking gun: every message either proves the unfavorable state or increases the belief that the favorable state is true at least to the level that unconstrained persuasion would have targeted. Three distinct messages always suffice. When the constraint binds, the expected decision in the unfavorable state ends up exactly where it would have been with no evidence at all, and the whole of the persuader's gain comes from the favorable state.

Take-aways. The paper shows that unfair prejudice — a notion that legal systems have long applied by intuition — has a clean decision-theoretic meaning for a fully rational fact-finder, and that imposing it has predictable, and in part surprising, consequences for the evidence that gets produced. A no-prejudice standard is an accuracy requirement, not a gag rule. Institutions that resist such standards for fear of suppressing informative disclosure have the sign of the effect wrong. Prohibiting evidence that helps one side in every state does not reduce the amount of information reaching the decision maker; it raises the evidentiary bar and shifts the composition of what is presented toward sharper and more conclusive evidence.

Related literature. The paper contributes to three strands of literature.

Information design with additional constraints. Since Kamenica and Gentzkow (2011), a large literature studies the design of information subject to Bayes' plausibility alone. A growing body of work adds further constraints, which typically arise because the designer must respect incentives elsewhere in the game. Under limited commitment, the designer anticipates that she will be unable to follow through on some messages, so the design must not rely on messages she would not in fact send (e.g., Lipnowski et al., 2022; Min, 2021; Nguyen and Tan, 2021). In the presence of moral hazard or career concerns, the information structure shapes an agent's effort, so the design must remain incentive compatible for him (e.g., Boleslavsky and Cotton, 2015; Zubrickas, 2015; Zhang and Zhou, 2016; Boleslavsky and Kim, 2018; Zapechelnyuk, 2020; Hörner and Lambert, 2021). In sequential persuasion, a first mover designs anticipating the design of those who follow (e.g., Li and Norman, 2021; Mylovanov and Zapechelnyuk, 2024; Wu, 2023). Related constraints arise from voluntary participation (Rosar, 2017) and from limited communication capacity (Le Treust and Tomala, 2019). General tools for such problems are developed by Kolotilin (2018), Dworczak and Martini (2019), and Doval and Skreta (2024); the last shows that with n states and k constraints the optimal design needs at most $n + k$ posteriors, which in our binary-state problem with a single constraint gives the three posteriors of Theorem 2. Our constraint is new, and it differs from those above in its origin: it is imposed by an outside institution as a fairness or admissibility standard, rather than derived from the incentives of a player in the game. Substantively, our contribution is to characterize when such a constraint binds, and to establish that it leads to more, not less, disclosure.

Robustness and ambiguity in information design. In our model, we use a constraint evaluated state by state against a fixed benchmark, rather than weighing states by the prior. In this it is related to work that seeks guarantees rather than expected-utility improvements. Dworczak and Pavan (2022) study persuasion when the designer is uncertain about the receiver's model and evaluates designs by their worst case. Hu and Weng (2021) and Kosterina (2022) study persuasion when the receiver's beliefs are unknown to the sender. Schlag (2026) axiomatizes a *non-guessing* criterion: information should be used so as to improve performance in the true state relative to a benchmark action, rather than on average under a prior.

Evidence in courts, arbitration, and dispute resolution. A long literature in law and economics studies how evidence is produced and evaluated in adjudication. Milgrom and Roberts (1986) show how an unaided decision maker can extract information from interested parties who can withhold it but not fabricate, and Shin (1998) uses that logic to compare adversarial and inquisitorial procedures in arbitration, finding an informational advantage for the adversarial one. Dewatripont and Tirole (1999) show that partisan advocates can be an efficient way to organize the search for evidence, while Daughety and Reinganum (2000) study the bias in the evidentiary record that emerges in equilibrium when both sides select what to present. Closest to our theme are the papers on exclusionary rules: Schrag and Scotchmer (1994) and Sanchirico (2001) tackle the question of why the legal system generally bars using evidence of a person's character in criminal trials, and Lester et al. (2012) study how excluding evidence affects the parties' incentives to acquire it in the first place. Recent work examines statistical and expert evidence in litigation (Che and Severinov, 2017; Bull and Watson, 2019), and the design of standards of proof (Kaplow, 2011). Relative to this literature, we make two contributions. First, we give the legal notion of unfair prejudice a decision-theoretic definition that does not rest on any error by the fact-finder: even a fully rational, correctly updating arbitrator can be tilted toward one party in every state of the world, and it is this, rather than any cognitive failure, that an exclusionary rule may be understood to target. Second, we characterize how the supply of evidence responds to such a rule, and find that it raises rather than suppresses the informativeness of what is presented — an implication that bears on the long-standing debate over whether restrictive rules of evidence impoverish the fact-finder's access to information.

2. A Model and Two Stories

2.1. **Preliminaries.** A principal wishes to influence an agent's decision, by disclosing evidence (a Blackwell experiment) about an uncertain state $\theta \in \{0, 1\}$. The parties share a common prior $\mu \in (0, 1)$ about the state.[1,2]

The agent makes a decision in $A = [0, 1]$. For every posterior belief $x \in [0, 1]$ of the agent, the decision is given by $a^*(x)$. Assume that $a^*(0) = 0$, $a^*(1) = 1$, and $a^*(\cdot)$ is continuous and increasing. In words, the agent's decision matches the state when she is certain what the state is, and it is monotone in her beliefs. In addition, for tractability, assume that

$$a^*(\cdot) \text{ is either strictly } S\text{-shaped or strictly inverted } S\text{-shaped.} \tag{A$_1$}$$

Specifically, $a^*(\cdot)$ is strictly S-shaped if it is strictly convex up to a threshold and strictly concave above that threshold.[3] The threshold can be degenerate, so globally convex and concave cases are included. The definition of inverted S-shaped is symmetric.

Economic interpretations are different for S-shaped and inverted S-shaped decision functions. When a^* is S-shaped, the agent overreacts to changes in intermediate beliefs, with the reaction slowing down as the beliefs become more extreme. The limit case is a step function that dictates $a = 0$ or $a = 1$, depending whether x is below or above a cutoff. In contrast, when a^* is inverted S-shaped, the situation is the opposite: agent underreacts in the middle. The limit case is a constant decision for all interior beliefs.

We take a decision function a^* as given. Two stories motivating how a^* can emerge are provided below, in Sections 2.3 and 2.4.

[1]Throughout the paper, beliefs are expressed as probabilities that $\theta = 1$.
[2]The model and results straightforwardly extend to the case where the principal has a subjective prior $\mu_P \neq \mu$. This extension is discussed in Section 5.1.
[3]The results can be straightforwardly extended to allow a^* to include linear segments at the expense of somewhat more involved notation.

The principal wishes to maximize the expected decision, so the principal's payoff under a posterior x is $a^*(x)$.

The principal can present evidence to persuade the agent. A piece of evidence is a Blackwell experiment $E = (\mathcal{M}, p_0, p_1)$, where $\mathcal{M} = \{m_1, ..., m_M\}$ is a finite set of messages with $M \geq 2$, and $p_\theta(\cdot)$ is a conditional probability distribution over messages in state $\theta \in \{0, 1\}$. W.l.o.g., assume that $p_0(m) + p_1(m) > 0$ for all $m \in \mathcal{M}$, so each message is sent with positive probability. Let $\mathcal{E}$ be the set of all Blackwell experiments that satisfy the above.

2.2. **Prejudiced evidence.** We say evidence is *prejudiced* (in favor of the principal) if, regardless of the true state, it makes the agent choose a higher action relative to what she would have done without evidence. To define this formally, we introduce some notation.

Fix a piece of evidence $E = (\mathcal{M}, p_0, p_1)$ and a prior μ. Let x_m be the agent's posterior belief after observing message m:

$$x_m = \frac{\mu p_1(m)}{\mu p_1(m) + (1-\mu)p_0(m)}, \quad m \in \mathcal{M}. \tag{1}$$

Let $\bar{a}_\theta(E; \mu)$ be the expected decision induced by the evidence in state θ:

$$\bar{a}_\theta(E; \mu) = \sum_{m \in \mathcal{M}} p_\theta(m) a^*(x_m), \quad \theta \in \{0, 1\}. \tag{2}$$

We say that evidence is prejudiced if, conditional on each state $\theta \in \{0, 1\}$, the agent's decision after observing the evidence is expected to shift in favor of the principal.

Definition 1. Say that evidence $E = (\mathcal{M}, p_0, p_1) \in \mathcal{E}$ is *prejudiced* at prior μ if

$$\bar{a}_\theta(E; \mu) > a^*(\mu), \quad \theta \in \{0, 1\}. \tag{3}$$

Otherwise, say that E is *unprejudiced* at prior μ.

The principal's expected payoff after presenting evidence E is given by

$$V(E; \mu) = (1-\mu)\bar{a}_0(E; \mu) + \mu \bar{a}_1(E; \mu).$$

Assume that the principal is not allowed to use prejudiced evidence. Thus, the principal's problem is

$$\max_{E \in \mathcal{E}} V(E; \mu) \;\text{ s.t. } E \text{ is unprejudiced.} \tag{P}$$

2.3. **Story I. Dispute resolution.** Two parties, call them a plaintiff and a defendant, turn to an arbitrator to resolve a dispute over a monetary pie. One of the parties is right, the other is wrong. State $\theta \in \{0, 1\}$ reflects which party is right: it is 0 when the defendant is right and it is 1 when the plaintiff is right. The arbitrator chooses a split of the pie $a \in A = [0, 1]$ between the parties, where a denotes the plaintiff's share. The plaintiff's and defendant's payoffs are equal to the shares of the pie they get, so the plaintiff's payoff is a and the defendant's payoff is $1 - a$. The arbitrator would like to allocate the whole pie to the party that is right. However, the arbitrator is risk averse and would rather split the pie in some way if she is not confident enough about which party is right. Specifically, let the arbitrator have a CRRA utility function with a risk parameter η:

$$u(a) = \frac{a^{1-\eta} - 1}{1 - \eta}, \quad \eta > 0, \;\; \eta \neq 1.$$

The arbitrator's payoff from action a is $u(a)$ in state $\theta = 1$ and $u(1 - a)$ in state $\theta = 0$. The arbitrator's optimal choice $a^*(x)$ under a belief $x \in [0, 1]$ maximizes $xu(a) + (1 - x)u(1 - a)$. Under the CRRA assumption, we have

$$a^*(x) = \frac{x^{1/\eta}}{x^{1/\eta} + (1 - x)^{1/\eta}}. \tag{4}$$

Note that a^* is strictly S-shaped for $0 < \eta < 1$ and strictly inverted S-shaped for $\eta > 1$, so a^* satisfies (A_1). The boundary case, CRRA utility with $\eta = 1$, yields $a^*(x) = x$. This does not satisfy (A_1), but this is an uninteresting case, since no information disclosure can benefit the plaintiff, and the no-prejudice constraint is vacuous.

The arbitration game proceeds as follows. The plaintiff chooses a piece of evidence $E \in \mathcal{E}$. If the evidence is unprejudiced, then the arbitrator admits it, observes a realized message m, and takes an optimal action under the posterior belief induced by m. Alternatively, if the evidence is prejudiced, then the arbitrator does not admit it, and takes an optimal action under her prior.

2.4. **Story II. Product sales via a platform.** A manufacturer sells a product via an online platform at a fixed price. The product can be either low quality or high quality, denoted by $\theta = 0$ and $\theta = 1$, respectively. The product value (net of the price) is normalized to be equal to θ. There is a continuum of heterogeneous consumers indexed by $t \in [0, 1]$, distributed according to a CDF, denoted by $a^*(t)$, where consumer type t is the value of the outside option. Thus, given a belief $x \in [0, 1]$, a consumer of type t buys the product if and only if $x \geq t$, and the total sales quantity is $a^*(x)$.

The online platform has a quality screening system, e.g., an explicit quality certification or a rating system based on consumer reviews. Let us interpret it as evidence, or Blackwell experiment, about the product quality. The platform receives a percentage of revenue from sales, so its objective is to maximize sales, $a^*(\cdot)$. Assumption (A_1) means that the density is either single-peaked (so the distribution of consumers is concentrated) or single-dipped (so the distribution of consumers is polarized).

The operation of the platform is subject to government regulation that does not permit using prejudiced quality screening. Thus, the platform's problem is to choose an unprejudiced quality screening system that maximizes the sales.

3. Properties of Prejudiced Evidence

3.1. **When all evidence is unprejudiced.** First, we characterize the assumptions about $a^*(\cdot)$ such that all the evidence is unprejudiced, so the no-prejudice constraint does not matter.

Proposition 1. *If*

$$(1 - x)a^*(x) \text{ is weakly concave,} \tag{5}$$

then no evidence in $\mathcal{E}$ is ever prejudiced. Otherwise, there exists evidence in $\mathcal{E}$ that is prejudiced at an interval of priors.

The proofs of this and all other formal results are relegated to Appendix B.

In Story I, $a^*(\cdot)$ given by (4) satisfies condition (5) if and only if $\eta \geq 1$, that is, the arbitrator is sufficiently risk averse. The defining feature of the CRRA utility for $\eta \geq 1$ is that the disutility from the wrong decision is unbounded, $u(0^+) = -\infty$. Intuitively, the arbitrator's cautiousness leads her to under-react to changes in beliefs, so that the average reaction to posteriors in state zero can never swing so much as to result in the expected action greater than that under the prior.

In Story II, the density of the consumers' types is the derivative of $a^*(\cdot)$. Condition (5) can be understood as saying that the density has a fat left tail. Intuitively, there are too many consumers at the left side of the spectrum, so that the average mass of persuaded consumers in state zero can never be greater than that under the prior.

3.2. **Smoking guns.** Next, we provide a sufficient condition for a piece of evidence to never be prejudiced, no matter what the prior is.

Say that a piece of evidence is *smoking gun evidence* if the decision is increased only after messages that prove beyond doubt that the state is $\theta = 1$.

Definition 2. Say that evidence $E = (\mathcal{M}, p_0, p_1) \in \mathcal{E}$ is a *smoking gun evidence* if, for every $m \in \mathcal{M}$, either $p_0(m) = 0$ or $p_0(m) \geq p_1(m)$.

A message m where $p_0(m) = 0$ proves that the state is $\theta = 1$, so its posterior is $x_m = 1$. A message m where $p_0(m) \geq p_1(m)$ generates a posterior that is weakly unfavorable to the principal, so $x_m \leq \mu$. Note that full disclosure and, vacuously, no disclosure are special cases of smoking gun evidence.

Proposition 2. *If $E \in \mathcal{E}$ is a smoking gun evidence, then it is unprejudiced at all priors. Otherwise, there exists $a^*(\cdot)$ such that E is prejudiced at an interval of priors.*

To see why this is true, observe that in this type of evidence messages with $p_0(m) < p_1(m)$, which lead to posteriors $x_m > \mu$, do not occur in state $\theta = 0$ at all. Thus, the average decision in state 0 cannot exceed $a^*(\mu)$, so this evidence is unprejudiced. The converse is proved by a construction of $a^*(\cdot)$ and an appropriate choice of μ such that (3) holds.

3.3. **Non-monotonicity in informativeness.** Only smoking gun evidence is always unprejudiced, while anything else can be prejudiced. The property of being unprejudiced is therefore not monotone with respect to the Blackwell order: garbling an unprejudiced experiment can make it prejudiced, and refining one can too. This is because a change in the Blackwell order of informativeness in general imposes little discipline on what happens conditional on each state.

3.4. **Non-monotonicity under combination of evidence.** Two conditionally independent pieces of evidence, each unprejudiced at μ, can combine into a prejudiced experiment when presented jointly. And vice versa, two conditionally independent prejudiced pieces can combine into an unprejudiced experiment. The reason is that each piece is tested against μ, but the combined experiment can be seen as the second experiment applied at each posterior induced by the first experiment. Thus, the combined experiment may fail to be unprejudiced if the second experiment is prejudiced under some posteriors of the first experiment. This is illustrated by an example in Appendix A.

Furthermore, when considering two correlated pieces, each of which is unprejudiced, neither direction of correlation between the individual pieces provides any guarantee about whether the joint evidence will be unprejudiced. For example, holding both marginals fixed, a combination can be unprejudiced under perfect conditional positive correlation, prejudiced under conditional independence, and unprejudiced again under maximal conditional negative correlation. What matters is not the sign of the correlation but how the dependence differs across states, since that is what determines the likelihood ratios of the joint messages.

4. Persuasion by Unprejudiced Evidence

4.1. **Reformulating the problem.** Every evidence $E = (\mathcal{M}, p_0, p_1) \in \mathcal{E}$ can be represented as a discrete distribution over posterior beliefs (e.g. Kamenica and Gentzkow, 2011). For this purpose, for each message m, denote by q_m the unconditional probability that this message is sent:

$$q_m = \mu p_1(m) + (1-\mu)p_0(m), \quad m \in \mathcal{M}. \tag{6}$$

Recall that x_m is the posterior induced by m. Let

$$H(x) = \Pr[\text{posterior} \leq x] = \begin{cases} 0, & \text{if } x \in [0, x_1), \\ \sum_{i=1}^{m} q_i, & \text{if } x \in [x_m, x_{m+1}),\ m = 1, ..., M-1, \\ 1, & \text{if } x \in [x_M, 1]. \end{cases}$$

In what follows, we will often identify evidence E with the distribution of the posteriors H it induces.

Let $\mathcal{H}$ be the set of all distributions on $[0,1]$ with finite support, and denote by H a representative element of $\mathcal{H}$. Denote

$$\psi(x) = (1-x)a^*(x).$$

We show that the principal's problem (P) is a standard Bayesian persuasion problem as in Kamenica and Gentzkow (2011), but with an extra constraint stemming from the requirement of no prejudice.

Lemma 1. *Evidence $E \in \mathcal{E}$ solves problem (P) if and only if the distribution of the posterior beliefs induced by E solves*

$$\max_{H\in\mathcal{H}} \mathbb{E}_H[a^*(x)] \quad s.t. \; (I) \;\; \mathbb{E}_H[x] = \mu, \quad (II) \;\; \mathbb{E}_H[\psi(x)] \le \psi(\mu). \tag{P$'$}$$

Here, constraint (I) is Bayes' plausibility and constraint (II) corresponds to no prejudice in state 0, that is, $\bar{a}_0(E;\mu) \le a^*(\mu)$. It turns out that no prejudice in state 1 is redundant, and thus it does not appear in (P′). Intuitively, this is because state 1 favors the principal to begin with, so when E is prejudiced in state 0, it must also be prejudiced in state 1, and thus the constraint in state 1 is vacuous.

4.2. **Bayesian persuasion benchmark.** Consider problem (P′) but without constraint (II). This is a standard Bayesian persuasion problem solved by concavification. Let

$$x^* \in \arg\max_{x\in[0,1]} \frac{a^*(x)}{x} \quad \text{and} \quad x^{**} \in \arg\min_{x\in[0,1]} \frac{1-a^*(x)}{1-x}.$$

When a^* is strictly S-shaped, x^* is the unique maximizer of $a^*(x)/x$. Geometrically, x^* is a tangency point to a^* of the ray originating from $(0,0)$, as is illustrated in Figure 1 (Section 4.4). Thus, x^* defines the concave closure of a^*, which is the straight line for $x \in [0,x^*]$, and it is a^* itself for $x \ge x^*$. Similarly, when a^* is strictly inverted S-shaped, x^{**} is the unique minimizer of $(1-a^*(x))/(1-x)$ on $[0,1]$. Mirroring the S-shaped case, here the concave closure of a^* is a^* itself for $x \le x^{**}$, and it is the straight line for $x \in [x^{**},1]$. Note that $x^* = 0$ and $x^{**} = 1$ when a^* is concave.

Proposition 3. *Consider problem (P′) without constraint (II).*

(a) If a^* is strictly S-shaped and $\mu < x^*$, then the optimal evidence is unique and has two messages with posterior beliefs 0 and x^*, each occurring with positive probabilities.

(b) If a^* is strictly inverted S-shaped and $\mu > x^{**}$, then the optimal evidence is unique and has two messages with posterior beliefs x^{**} and 1, each occurring with positive probabilities.

(c) Otherwise, no-disclosure evidence is optimal.

We omit the proof of this proposition, as it is standard in Bayesian persuasion literature, e.g., Proposition 3 in Kolotilin (2018) applied to the binary state case.

In case (c), the principal cannot benefit from information disclosure at all, so the no-prejudice constraint is vacuous.

In case (b), the principal may benefit from information disclosure. However, notice that the evidence whose posterior beliefs are in $\{x^{**}, 1\}$ is the *smoking gun* evidence, as defined in Section 3.2. Consequently, by Proposition 2, optimal evidence is unprejudiced, so the no-prejudice constraint does not bind.

Case (a) is, thus, the only interesting case. Here, the principal may benefit from information disclosure and constraint (II) may bind. Notice that the evidence whose posterior beliefs are in $\{0, x^*\}$ is the symmetric counterpart of *smoking gun* evidence: the agent's decision is increased (relative to no disclosure) only after messages that prove beyond doubt that state is $\theta = 0$.

4.3. **No prejudice improves disclosure.** The central question is how the no-prejudice constraint affects optimal persuasion by the principal when it binds. Since no-disclosure evidence and full-disclosure evidence are both unprejudiced, one may wonder if the constraint pushes the principal to disclose less or more. Here we answer this question.

We compare informativeness of evidence by Blackwell order. Namely, say that E is *strictly more informative* than E' if the associated distributions of posteriors, H and H', are such that H is a mean-preserving spread of H' and $H \neq H'$.

Let V^* be the optimal value of persuasion without prejudice, i.e., problem (P′). Let V_{BP} be the optimal value of standard Bayesian persuasion, i.e., problem (P′) without constraint (II).

Theorem 1. *If $V^* < V_{BP}$, then every optimal evidence under persuasion without prejudice is strictly more informative than the optimal evidence under standard Bayesian persuasion.*

Note that the statement of Theorem 1 is vacuous in cases (b) and (c) of Proposition 3, since in these cases we have $V^* = V_{BP}$. The only relevant case is (a), where a^* is S-shaped and $\mu < x^*$. To prove Theorem 1, we need the result about the structure of solutions of (P′), which we present below.

4.4. **One-sided evidence.** Let us characterize the solutions of problem (P′). We provide a necessary and sufficient condition for the no-prejudice constraint to be binding, and show that the optimal evidence is one-sided. Namely, it is the symmetric counterpart of *smoking gun* evidence: either it certifies that state is $\theta = 0$ or it shifts the posterior above x^*.

Recall the function $\psi(x) = (1-x)a^*(x)$ introduced in Section 4.1 to describe the no-prejudice constraint (II) in (P′).

Theorem 2. *Suppose that a^* is strictly S-shaped. Then, $V^* < V_{BP}$ if and only if*

$$x^* > \mu \quad \textit{and} \quad \frac{\psi(x^*)}{x^*} > \frac{\psi(\mu)}{\mu}. \tag{7}$$

When (7) holds, every solution H^* of (P′) has the following properties: its support is contained in $\{0\} \cup [x^*, 1]$, and the largest posterior is strictly greater than x^*. In addition, at least one solution has at most three posteriors, $\{0, x', x''\}$, where x' and x'' are in $[x^*, 1]$.

The theorem says that the optimal evidence is one-sided: each message either proves beyond doubt that the state is $\theta = 0$, or raises the agent's belief to x^* or above. Moreover, the largest posterior is strictly above x^*. This observation immediately proves Theorem 1. Since V_{BP} is attained by the evidence H_{BP} with support $\{0, x^*\}$ and every solution H^* of (P′) is supported on $\{0\} \cup [x^*, 1]$ with some posterior strictly above x^*, H^* is a mean-preserving spread of H_{BP}.

The theorem implies that when (7) holds and the no-prejudice constraint is binding, we can find an optimal solution with support on at most three points. Finding a solution thus simplifies to solving the problem

$$\max_{(x',x'')\in[x^*,1]^2,\,(q',q'')\in[0,1]^2} q'a^*(x') + q''a^*(x''), \tag{P''}$$

subject to

$$q'x' + q''x'' = \mu \qquad \text{(Bayes' plausibility)}$$

$$q'\psi(x') + q''\psi(x'') = \psi(\mu) \quad \text{(No prejudice)}$$

Note that $q' + q'' \leq \mu/x^*$ (due to $x', x'' \geq x^* > \mu$ and Bayes' plausibility). The remaining probability, $1 - q' - q''$, is assigned to the posterior 0, which does not appear in (P″) since $a^*(0) = 0$ and $\psi(0) = 0$.

Example. Let a^* be given by (4) with $\eta = 1/2$, that is, $a^*(x) = x^2/(x^2 + (1-x)^2)$. In Story I, this is the case of $u(x) \sim \sqrt{x}$. Then $x^* = 1/\sqrt{2}$, so the standard Bayesian persuasion solution has support on $\{0, 1/\sqrt{2}\}$. Condition (7) holds if and only if $\mu < 1 - 1/\sqrt{2}$. When it holds, the solution of (P″) has support $\{0, 1-\mu\}$, that is, $x' = x'' = 1 - \mu$, and

$$V^* = \frac{\mu(1-\mu)}{\mu^2 + (1-\mu)^2} \;<\; V_{BP} = \frac{1+\sqrt{2}}{2}\,\mu.$$

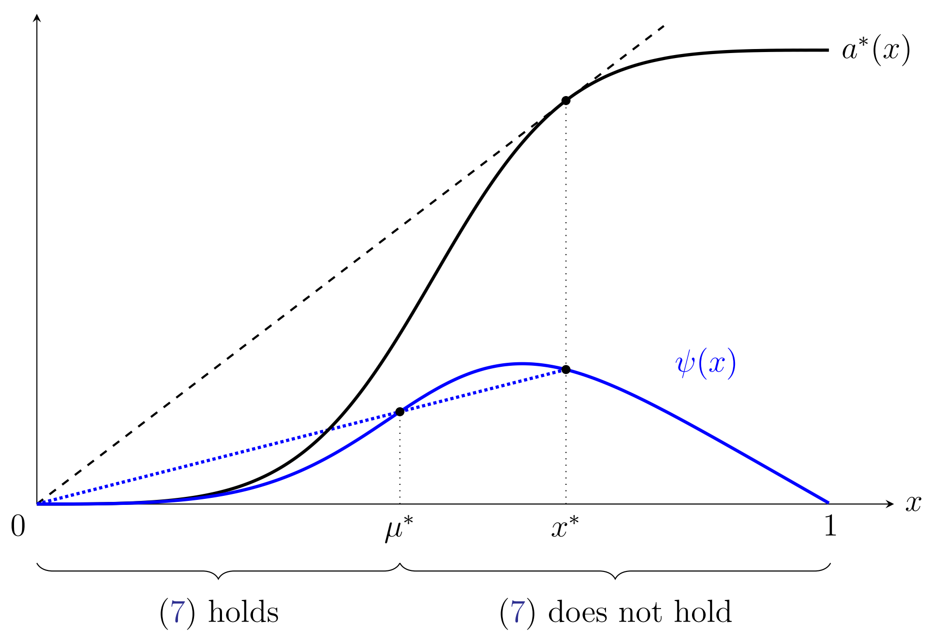

Figure 1. No-prejudice constraint binds for $\mu < \mu^*$ and does not bind for $\mu \geq \mu^*$.

Let us comment on the proof of Theorem 2. The proof of the statement that $V^* < V_{BP}$ if and only if (7) holds is simple. If $\mu \geq x^*$, then, by Proposition 3(c), the principal cannot do better than disclosing no evidence, so $V^* = V_{BP}$. Suppose that $\mu < x^*$. By Proposition 3(a), the optimal evidence under standard Bayesian persuasion has support $\{0, x^*\}$. With this support, the no-prejudice condition (II) reduces to

$$\left(1 - \frac{\mu}{x^*}\right) \psi(0) + \frac{\mu}{x^*} \psi(x^*) \leq \psi(\mu),$$

which is the same as $\psi(x^*)/x^* \leq \psi(\mu)/\mu$ since $\psi(0) = 0$. If it is satisfied, then the optimal evidence under standard Bayesian persuasion is unprejudiced, so $V^* = V_{BP}$. Otherwise, as that evidence is the unique solution of the problem without constraint (II), the value V_{BP} cannot be achieved in (P′), so $V^* < V_{BP}$. Figure 1 illustrates how constraint (7) applies. In (7), condition $x^* > \mu$ determines for which μ the evidence with support $\{0, x^*\}$ exists, and condition $\psi(x^*)/x^* > \psi(\mu)/\mu$ determines for which μ this evidence is prejudiced. The latter condition is illustrated by the comparison of $\psi(\mu)$, depicted as the solid blue curve, and the function $x \cdot \psi(x^*)/x^*$, depicted as the dotted blue ray from the origin to point $(x^*, \psi(x^*))$. As shown in the figure, $\psi(\mu)$ is below the dotted ray for $\mu < \mu^*$, which is where constraint (II) binds.

The proof of the statement about the support is more involved. Since problem (P′) is a linear program with three constraints (probabilities adding up to 1, Bayes' plausibility, and no prejudice), a standard linear programming argument implies that a solution with a support of at most three points exists. To locate the support of an arbitrary solution we use duality: the support of every solution is contained in the set of maximizers of the Lagrangian $a^*(x) - \lambda x - \nu \psi(x)$ for suitable multipliers $\lambda \in \mathbb{R}$ and $\nu \geq 0$. Writing this function as $x\big[(1 - \nu(1 - x))a^*(x)/x - \lambda\big]$ and using the fact that $a^*(x)/x$ is uniquely maximized at $x = x^*$ (this is where the S-shape of a^* enters), we show that no point of $(0, x^*)$ can be a maximizer of the Lagrangian. Bayes' plausibility then forces the posterior 0 to be in the support, because $\mu < x^*$.

5. Concluding Remarks

To conclude, we discuss four possible extensions or generalizations whose detailed analysis is left for future research.

5.1. **Asymmetric information and subjective priors.** In reality, it is hard to imagine a situation where the principal and the agent are symmetrically (un)informed about the state. An alternative is the informed principal model, where the principal is better informed and can signal his information through his choices (Perez-Richet, 2014; Hedlund, 2017; Koessler and Skreta, 2023; Zapechelnyuk, 2023). This model presents many challenges, in particular, a multiplicity of equilibria, and their fragility to changes in off-path beliefs and choices.

Instead, one can consider a more tractable model of subjective priors, as in Alonso and Câmara (2016). Therein, the parties may have different priors. One justification is that the parties simply agree to disagree: they are not sophisticated enough to build and apply a general model of the world à la Harsanyi to internalize the difference in their beliefs. Another justification is that the principal is better informed, but the agent, due to an institutional constraint or lack of sophistication, cannot make inference from the principal's choices. For example, an arbitrator may be institutionally constrained to rely only on substantive evidence and not on musing about what the plaintiff's choice of evidence may mean about who is right.

Suppose that the principal's prior is μ_P and the agent's is μ. The key difference is that the principal's objective in (P′) becomes

$$\max_{H\in\mathcal{H}} \ \mathbb{E}_H\left[a^*(x)\left(\frac{\mu_P}{\mu}x+\frac{1-\mu_P}{1-\mu}(1-x)\right)\right].$$

The constraints in (P′) remain unaffected, since Bayes' plausibility applies to the agent's inference, and the no-prejudice constraint applies conditional on state 0. The rest of the analysis extends straightforwardly.

5.2. **Two-sided constraint.** We call evidence prejudiced if it shifts the decision in the principal's favor in each state. However, in the arbitration story, the arbitrator may wish to reject not only evidence that shifts the decision in favor of the plaintiff, but also evidence that shifts the decision in favor of the defendant, irrespective of which party is right. Our no-prejudice condition can be straightforwardly extended to include its symmetric counterpart. Note that this would not have any impact on the optimal persuasion analysis, because the plaintiff would never wanted to present such evidence in the first place.

5.3. **Continuous state.** The definition of no-prejudice easily extends to a continuum of states, provided that the principal's preferences are monotone, i.e., the principal prefers higher decisions in all states. Say, the state is in $[0, 1]$. Then, evidence is prejudiced if it strictly shifts the agent's decision in the principal's favor in all states, except, perhaps, on a set of zero measure under the prior. In the mean-measurable case (where the posterior mean state is a sufficient statistic for both parties' payoffs), the results can be generalized at the expense of the simplicity of exposition, while the general case is challenging and may require methodology that is not covered in this paper.

5.4. **Sequential disclosure.** Suppose that the principal is unable to design evidence arbitrarily; instead he has an exogenously given set of pieces of evidence and is tasked to decide which of them to present and in what order. One may ask how the problem and the results would change if the no-prejudice constraint applies in every interim stage. That is, after one piece of evidence is presented, the next piece may or may not be prejudiced conditional on what is revealed by the first piece. While this framework is quite different, we would like to point out that at each stage the methodology of this paper applies as follows: each realized posterior of one piece becomes the prior for the evaluation whether the next piece is prejudiced.

Appendix A. Non-monotonicity under combination of evidence: Example

Consider the evidence $E = (\mathcal{M}, p_0, p_1)$ with $\mathcal{M} = \{L, H\}$, $p_0(H) = 1/10$ and $p_1(H) = 4/5$, and compare it with two conditionally independent pieces of E presented simultaneously, denoted by $E \oplus E$. The messages of $E \oplus E$ are LL, LH, HL, and HH. Messages LH and HL induce the same posterior, so let us merge them and call this message HL.

E unprejudiced $\not\Longrightarrow$ $E \oplus E$ unprejudiced. Let a^* be given by (4) with $\eta = 1/5$, and let $\mu = 2/5$, so that $a^*(\mu) \approx 0.116$. The single evidence E yields the posteriors $x_L \approx 0.129$ and $x_H \approx 0.842$, and it is unprejudiced, as condition (3) is violated in state 0:

$$(1 - p_0(H))a^*(x_L) + p_0(H)a^*(x_H) \approx 0.1 < 0.116 \approx a^*(\mu).$$

The double evidence $E \oplus E$ yields the posteriors $x_{LL} \approx 0.032$, $x_{HL} \approx 0.542$, and $x_{HH} \approx 0.977$, and it is prejudiced, as condition (3) is satisfied:

$$(1-p_0(H))^2 a^*(x_{LL}) + 2(1-p_0(H))p_0(H)a^*(x_{HL}) + (p_0(H))^2 a^*(x_{HH}) \approx 0.136,$$
$$(1-p_1(H))^2 a^*(x_{LL}) + 2(1-p_1(H))p_1(H)a^*(x_{HL}) + (p_1(H))^2 a^*(x_{HH}) \approx 0.864,$$

both are greater than $a^*(\mu) \approx 0.116$.

E prejudiced $\not\Rightarrow$ *$E \oplus E$ prejudiced.* Let now a^* be given by (4) with $\eta = 1/2$, and let $\mu = 1/5$, so that $a^*(\mu) \approx 0.059$. The same evidence E now yields the posteriors $x_L \approx 0.053$ and $x_H = 2/3$, and it is prejudiced, as condition (3) is satisfied:

$$(1-p_0(H))a^*(x_L) + p_0(H)a^*(x_H) \approx 0.083,$$
$$(1-p_1(H))a^*(x_L) + p_1(H)a^*(x_H) \approx 0.641,$$

both are greater than $a^*(\mu) \approx 0.059$. The double evidence $E \oplus E$ now yields the posteriors $x_{LL} \approx 0.012$, $x_{HL} \approx 0.308$ and $x_{HH} \approx 0.941$, and it is unprejudiced, as condition (3) is violated in state 0:

$$(1-p_0(H))^2 a^*(x_{LL}) + 2(1-p_0(H))p_0(H)a^*(x_{HL}) + (p_0(H))^2 a^*(x_{HH}) \approx 0.040,$$

which is smaller than $a^*(\mu) \approx 0.059$.

Appendix B. Proofs.

B.1. Preliminary lemmas. Let Δ denote the set of Borel probability measures on $[0,1]$ and $\Delta_f \subset \Delta$ the subset of measures with finite support, so that $\mathcal{H} = \Delta_f$. We identify a distribution of posteriors H with the corresponding measure and write $\mathrm{supp} H$ for its support, $H(\{x\})$ for the mass at x, and $\mathbb{E}_H[f] = \int_0^1 f \,\mathrm{d}H$. Recall the function $\psi(x) = (1-x)a^*(x)$ introduced in Section 4.1, and denote

$$\rho(x) = \frac{a^*(x)}{x}.$$

In this notation x^* is the maximizer of ρ.

Assumption (A_1) is used in the following form: there is a threshold $\kappa \in [0,1]$ such that a^* is strictly convex on $[0,\kappa]$ and strictly concave on $[\kappa,1]$ (S-shaped), or strictly concave on $[0,\kappa]$ and strictly convex on $[\kappa,1]$ (inverted S-shaped); degenerate thresholds ($\kappa = 0$ or $\kappa = 1$) are permitted, in which case a^* is concave.

Throughout the appendix, $\mu \in (0,1)$ is fixed unless stated otherwise.

The first lemma collects the elementary identities that translate statements about experiments into statements about distributions of posteriors.

Lemma 2. *Let $E = (\mathcal{M}, p_0, p_1) \in \mathcal{E}$ and let $H \in \mathcal{H}$ be the distribution of posteriors it induces at the prior μ, i.e. H puts mass q_m on x_m for $m \in \mathcal{M}$, with q_m and x_m given by (6) and (1). Then*

(i) $\mathbb{E}_H[x] = \mu$ (Bayes' plausibility);

(ii) $(1-\mu)\,\bar{a}_0(E;\mu) = \mathbb{E}_H[\psi(x)]$ *and* $\mu\,\bar{a}_1(E;\mu) = \mathbb{E}_H[xa^*(x)]$;

(iii) $V(E;\mu) = \mathbb{E}_H[a^*(x)]$;

(iv) $\bar{a}_1(E;\mu) - \bar{a}_0(E;\mu) = \dfrac{Cov_H\big(x, a^*(x)\big)}{\mu(1-\mu)} \;\geq\; 0$;

(v) *E is prejudiced at μ* $\iff$ $\bar{a}_0(E;\mu) > a^*(\mu)$ $\iff$ $\mathbb{E}_H[\psi(x)] > \psi(\mu)$;

(vi) *every $H \in \mathcal{H}$ with $\mathbb{E}_H[x] = \mu$ is induced by some $E \in \mathcal{E}$.*

Proof. By (1) and (6),

$$q_m x_m = \mu p_1(m) \quad \text{and} \quad q_m(1-x_m) = (1-\mu)p_0(m), \qquad m \in \mathcal{M}. \tag{8}$$

(i) Summing the first identity in (8) over m gives

$$\mathbb{E}_H[x] = \sum_m q_m x_m = \mu \sum_m p_1(m) = \mu.$$

(ii) By the second identity in (8),

$$(1-\mu)\bar{a}_0(E;\mu) = \sum_m (1-\mu)p_0(m)a^*(x_m) = \sum_m q_m(1-x_m)a^*(x_m) = \mathbb{E}_H[\psi(x)],$$

and symmetrically $\mu\bar{a}_1(E;\mu) = \sum_m q_m x_m a^*(x_m) = \mathbb{E}_H[xa^*(x)]$.

(iii) Adding the two identities in (ii),

$$V(E;\mu) = (1-\mu)\bar{a}_0(E;\mu) + \mu\bar{a}_1(E;\mu) = \mathbb{E}_H[(1-x)a^*(x) + xa^*(x)] = \mathbb{E}_H[a^*(x)].$$

(iv) Subtracting the two identities in (8) after dividing by μ and $1-\mu$ respectively,

$$p_1(m) - p_0(m) = q_m \frac{(1-\mu)x_m - \mu(1-x_m)}{\mu(1-\mu)} = \frac{q_m(x_m-\mu)}{\mu(1-\mu)},$$

whence

$$\bar{a}_1(E;\mu) - \bar{a}_0(E;\mu) = \sum_m \big(p_1(m) - p_0(m)\big)a^*(x_m) = \frac{\mathbb{E}_H\big[(x-\mu)a^*(x)\big]}{\mu(1-\mu)},$$

which equals $\mathrm{Cov}_H(x, a^*(x))/(\mu(1-\mu))$ by (i). The covariance is nonnegative because a^* is increasing: if x and x' are independent draws from H, then $2\mathrm{Cov}_H(x, a^*(x)) = \mathbb{E}\big[(x-x')(a^*(x) - a^*(x'))\big] \geq 0$, since the integrand is nonnegative pointwise.

(v) By (iv), $\bar{a}_1(E;\mu) \geq \bar{a}_0(E;\mu)$. Hence (3) holds for both states if and only if it holds for $\theta = 0$, i.e. if and only if $\bar{a}_0(E;\mu) > a^*(\mu)$. By (ii) the latter is equivalent to $\mathbb{E}_H[\psi(x)] > (1-\mu)a^*(\mu) = \psi(\mu)$.

(vi) Immediate by Kamenica and Gentzkow (2011, Proposition 1). ☐

The second lemma gives existence of a solution to (P′), the associated multipliers, and the reduction to at most three posteriors. It is stated for the relaxed problem in which H ranges over all of Δ. The two problems have the same value, since some solution of the relaxed problem has finite support and hence lies in $\mathcal{H}$.

Lemma 3. *Let*

$$V^* = \sup\Big\{\mathbb{E}_H[a^*(x)] \;:\; H \in \Delta,\ \mathbb{E}_H[x] = \mu,\ \mathbb{E}_H[\psi(x)] \leq \psi(\mu)\Big\}.$$

Then:

(i) *the supremum is attained;*

(ii) *there exist $\lambda \in \mathbb{R}$ and $\nu \geq 0$ such that, with $L(x) = a^*(x) - \lambda x - \nu\psi(x)$, every solution H^* satisfies supp$H^* \subseteq \arg\max_{[0,1]} L$ and $\nu\big(\psi(\mu) - \mathbb{E}_{H^*}[\psi(x)]\big) = 0$;*

(iii) *some solution has at most three points in its support;*

(iv) *if $V^* < V_{BP}$, then any ν as in (ii) is strictly positive and every solution satisfies $\mathbb{E}_{H^*}[\psi(x)] = \psi(\mu)$.*

Proof. Let $\Gamma(x) = \big(x, \psi(x), a^*(x)\big)$ for $x \in [0,1]$ and let $K = \mathrm{conv}\,\Gamma([0,1]) \subset \mathbb{R}^3$. Since Γ is continuous, $\Gamma([0,1])$ is compact, and K is compact and convex. We claim that K is exactly the set of achievable moment vectors,

$$K = \Big\{\big(\mathbb{E}_H[x], \mathbb{E}_H[\psi], \mathbb{E}_H[a^*]\big) : H \in \Delta\Big\} =: K'.$$

Indeed, $K \subseteq K'$ because every convex combination of points of $\Gamma([0,1])$ is the moment vector of the corresponding finite mixture of Dirac measures; and $K' \subseteq K$ because if $m \notin K$ then, K being compact and convex, there is $\alpha \in \mathbb{R}^3$ with $\alpha \cdot m > \max_{x\in[0,1]} \alpha \cdot \Gamma(x) \geq \mathbb{E}_H[\alpha \cdot \Gamma(x)]$ for every H, so m is not achievable.

(i) The feasible set $K \cap \{m_1 = \mu\} \cap \{m_2 \leq \psi(\mu)\}$ is compact, convex and nonempty (it contains $\Gamma(\mu)$), so m_3 attains a maximum V^* on it; any H whose moment vector is a maximizer is a solution.

(ii) The sets K and $T = \{(\mu, b, v) : b \leq \psi(\mu),\ v > V^*\}$ are convex, nonempty and disjoint, so there are $\alpha = (\alpha_1, \alpha_2, \alpha_3) \neq 0$ and $c \in \mathbb{R}$ with $\alpha \cdot k \leq c$ for all $k \in K$ and $\alpha \cdot t \geq c$ for all $t \in T$. Letting $v \to \infty$ in T gives $\alpha_3 \geq 0$; letting $b \to -\infty$ gives $\alpha_2 \leq 0$.

Note that

$$a^*(x) > 0, \qquad x \in (0,1]. \tag{9}$$

Otherwise, since a^* is increasing and $a^*(0) = 0$, a^* would vanish on $[0,x]$, and hence be affine on a nondegenerate interval, contradicting (A_1).

We claim $\alpha_3 > 0$. Suppose $\alpha_3 = 0$. The full-disclosure distribution (mass $1-\mu$ at 0 and μ at 1) has moment vector $(\mu, 0, \mu) \in K$, since $\psi(0) = \psi(1) = 0$; hence $\alpha_1\mu \leq c$. Taking $b = \psi(\mu)$ in the closure of T gives $\alpha_1\mu + \alpha_2\psi(\mu) \geq c \geq \alpha_1\mu$, so $\alpha_2\psi(\mu) \geq 0$; as $\psi(\mu) = (1-\mu)a^*(\mu) > 0$ by (9), this forces $\alpha_2 \geq 0$, hence $\alpha_2 = 0$. Then $\alpha_1 m_1 \leq c \leq \alpha_1\mu$ for every achievable m_1, i.e. for every $m_1 \in [0,1]$, which forces $\alpha_1 = 0$ because $\mu \in (0,1)$. Thus $\alpha = 0$, a contradiction.

Normalize $\alpha_3 = 1$ and set $\lambda = -\alpha_1$ and $\nu = -\alpha_2 \geq 0$. Applying $\alpha \cdot k \leq c$ to $k = \Gamma(x)$ gives $L(x) \leq c$ for all x, so $\max_{[0,1]} L \leq c$. Let H^* be any solution and $b^* = \mathbb{E}_{H^*}[\psi] \leq \psi(\mu)$. Applying $\alpha \cdot t \geq c$ to $t = (\mu, \psi(\mu), V^*)$ in the closure of T and using $\nu \geq 0$,

$$\mathbb{E}_{H^*}[L] = V^* - \lambda\mu - \nu b^* \ \geq\ V^* - \lambda\mu - \nu\psi(\mu) \ \geq\ c \ \geq\ \max_{[0,1]} L \ \geq\ \mathbb{E}_{H^*}[L].$$

Hence all inequalities are equalities. The last equality means $L = \max L$ H^*-almost surely; as L is continuous, $\{L = \max L\}$ is closed, so it contains $\operatorname{supp} H^*$. The first equality gives $\nu(\psi(\mu) - b^*) = 0$.

(iii) This follows from Doval and Skreta (2024, Corollary 3.1).

(iv) If $\nu = 0$, then for every H with $\mathbb{E}_H[x] = \mu$,

$$\mathbb{E}_H[a^*] = \mathbb{E}_H[L] + \lambda\mu \leq \max_{[0,1]} L + \lambda\mu = \mathbb{E}_{H^*}[L] + \lambda\mu = V^*,$$

so $V^* \geq V_{BP}$, contradicting $V^* < V_{BP}$. Hence $\nu > 0$ and, by the complementary slackness in (ii), $\mathbb{E}_{H^*}[\psi] = \psi(\mu)$ for every solution H^*. □

The final preliminary result is the key structural step: no posterior strictly between 0 and x^* can ever be used. It is used in the proofs of Theorems 1 and 2.

Lemma 4. *Let a^* be strictly S-shaped and not concave, and let $\lambda \in \mathbb{R}$ and $\nu \geq 0$. Then $L(x) = a^*(x) - \lambda x - \nu\psi(x)$ has no maximizer over $[0, 1]$ in the open interval $(0, x^*)$.*

Consequently, if $\mu < x^*$, then every solution H^* of (P′) satisfies supp$H^* \subseteq \{0\} \cup [x^*, 1]$ and $H^*(\{0\}) \geq 1 - \mu/x^* > 0$.

Proof. Write $w(x) = 1 - \nu(1 - x)$, so that $a^*(x) - \nu\psi(x) = a^*(x)\big[1 - \nu(1 - x)\big] = a^*(x)w(x)$, and set $\Phi(x) = \rho(x)w(x)$ for $x \in (0, 1]$. Then

$$L(0) = 0, \qquad L(1) = 1 - \lambda, \qquad L(x) = x\big(\Phi(x) - \lambda\big) \text{ for } x \in (0, 1].$$

Note that w is nondecreasing. Suppose, for a contradiction, that some $y \in (0, x^*)$ maximizes L.

Case 1: $w(y) \leq 0$. Then $a^*(y)w(y) \leq 0$, so $L(y) \leq -\lambda y$. If $\lambda > 0$, then $L(y) \leq -\lambda y < 0 = L(0)$. If $\lambda \leq 0$, then, using $0 < y < 1$, $L(y) \leq -\lambda y \leq -\lambda < 1 - \lambda = L(1)$. Either way y is not a maximizer, a contradiction.

Case 2: $w(y) > 0$. First, note that x^* is the unique maximizer of $\rho = a^*(x)/x$. To see this, observe that ρ is strictly increasing on $(0, \kappa]$, because strict convexity of a^* with $a^*(0) = 0$ gives $a^*(x) < xa^*(y)/y$ for $0 < x < y \leq \kappa$; hence, any maximizer lies in $[\kappa, 1]$. If $x_2 > x_1 \geq \kappa$ both maximized ρ with common value $\bar{\rho}$, strict concavity of a^* on $[\kappa, 1]$ at the midpoint $x_M = (x_1 + x_2)/2$ would give $\rho(x_M) = a^*(x_M)/x_M > \frac{1}{2}(a^*(x_1) + a^*(x_2))/x_M = \bar{\rho}$, contradicting maximality.

Next, since w is nondecreasing and $x^* > y$, we have $w(x^*) \geq w(y) > 0$; and since x^* is the unique maximizer of ρ on $(0, 1]$ and $y \neq x^*$, we have $\rho(x^*) > \rho(y) \geq 0$. Therefore

$$\Phi(x^*) = \rho(x^*)w(x^*) \;\geq\; \rho(x^*)w(y) \;>\; \rho(y)w(y) = \Phi(y).$$

If $\Phi(y) < \lambda$, then $L(y) = y(\Phi(y) - \lambda) < 0 = L(0)$, a contradiction. If $\Phi(y) \geq \lambda$, then, using $x^* > y > 0$,

$$L(x^*) = x^*\big(\Phi(x^*) - \lambda\big) > x^*\big(\Phi(y) - \lambda\big) \geq y\big(\Phi(y) - \lambda\big) = L(y),$$

again a contradiction.

For the consequence, let H^* solve (P′) and let λ, ν be as in Lemma 3(ii). Then $\operatorname{supp} H^* \subseteq \arg\max L \subseteq \{0\} \cup [x^*, 1]$. If $0 \notin \operatorname{supp} H^*$, then $\operatorname{supp} H^* \subseteq [x^*, 1]$ and $\mu = \mathbb{E}_{H^*}[x] \geq x^* > \mu$, a contradiction; so $0 \in \operatorname{supp} H^*$. Finally, $\mu = \mathbb{E}_{H^*}[x] = \int_{[x^*,1]} x \, \mathrm{d}H^* \geq x^*\big(1 - H^*(\{0\})\big)$, which gives $H^*(\{0\}) \geq 1 - \mu/x^* > 0$. □

B.2. Proof of Proposition 1. *Sufficiency.* Let ψ be weakly concave on $[0, 1]$, and let $E \in \mathcal{E}$ and $\mu \in (0, 1)$ be arbitrary, with induced posterior distribution H. By Lemma 2(i) and Jensen's inequality, $\mathbb{E}_H[\psi(x)] \leq \psi\big(\mathbb{E}_H[x]\big) = \psi(\mu)$, so E is unprejudiced at μ by Lemma 2(v). As E and μ were arbitrary, no evidence is prejudiced at any prior.

Necessity. Suppose ψ is not weakly concave. Then there are $0 \leq u < v \leq 1$ and $\lambda \in (0, 1)$ such that, with $\mu_0 = \lambda u + (1 - \lambda)v \in (u, v)$,

$$\lambda\psi(u) + (1 - \lambda)\psi(v) > \psi(\mu_0); \tag{10}$$

note that $\mu_0 \in (0, 1)$. Define $E = (\mathcal{M}, p_0, p_1)$ with $\mathcal{M} = \{m', m''\}$ and

$$p_1(m') = \frac{\lambda u}{\mu_0}, \qquad p_1(m'') = \frac{(1 - \lambda)v}{\mu_0},$$
$$p_0(m') = \frac{\lambda(1 - u)}{1 - \mu_0}, \qquad p_0(m'') = \frac{(1 - \lambda)(1 - v)}{1 - \mu_0}.$$

Both p_0 and p_1 are probability distributions, and $p_0(m) + p_1(m) > 0$ for $m \in \mathcal{M}$ because $\lambda \in (0, 1)$; hence $E \in \mathcal{E}$. At the prior μ_0, formulas (6) and (1) give $q_{m'} = \lambda u + \lambda(1 - u) = \lambda$ and $x_{m'} = \lambda u/\lambda = u$, and similarly $q_{m''} = 1 - \lambda$ and $x_{m''} = v$. Thus the induced distribution H_0 of posteriors puts masses λ and $1 - \lambda$ on u and v, and (10) says $\mathbb{E}_{H_0}[\psi(x)] > \psi(\mu_0)$. By Lemma 2(v), E is prejudiced at μ_0.

It remains to show that E is prejudiced on an interval of priors. Consider

$$D(\mu) = \bar{a}_0(E; \mu) - a^*(\mu) = \sum_{m \in \mathcal{M}} p_0(m)\, a^*\big(x_m(\mu)\big) - a^*(\mu), \qquad \mu \in (0, 1),$$

where $x_m(\mu) = \mu p_1(m)/\big(\mu p_1(m) + (1-\mu)p_0(m)\big)$. For each m the denominator is strictly positive for $\mu \in (0,1)$ because $p_0(m) + p_1(m) > 0$, so $x_m(\cdot)$ is continuous on $(0,1)$; since a^* is continuous, D is continuous on $(0,1)$. By Lemma 2(v), E is prejudiced at μ if and only if $D(\mu) > 0$, and we have shown $D(\mu_0) > 0$. By continuity there is $\varepsilon > 0$ such that $D > 0$ on $(\mu_0 - \varepsilon, \mu_0 + \varepsilon) \cap (0,1)$, so E is prejudiced at every prior in that interval.

B.3. Proof of Proposition 2. *Sufficiency.* Let E be a smoking gun evidence and fix $\mu \in (0,1)$. Partition $\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_0$, where $\mathcal{M}_1 = \{m : p_0(m) = 0\}$ and $\mathcal{M}_0 = \mathcal{M} \setminus \mathcal{M}_1$. That is, messages in $\mathcal{M}_1$ prove that $\theta = 1$. For $m \in \mathcal{M}_0$ we have $p_0(m) > 0$ and, by Definition 2, $p_0(m) \geq p_1(m)$; by (1) this is equivalent to $x_m \leq \mu$, hence $a^*(x_m) \leq a^*(\mu)$. Since $\sum_{m\in\mathcal{M}_1} p_0(m) = 0$, we have $\sum_{m\in\mathcal{M}_0} p_0(m) = 1$ and therefore

$$\bar{a}_0(E;\mu) = \sum_{m\in\mathcal{M}_0} p_0(m) a^*(x_m) \leq a^*(\mu).$$

By Lemma 2(v), E is unprejudiced at μ. As μ was arbitrary, E is unprejudiced at all priors.

Necessity. Let E not be a smoking gun evidence: there is $\hat{m} \in \mathcal{M}$ with $p_0(\hat{m}) > 0$ and $p_0(\hat{m}) < p_1(\hat{m})$. Fix any $\mu_0 \in (0,1)$ and let $\hat{x} = x_{\hat{m}}(\mu_0)$; since $p_1(\hat{m}) > p_0(\hat{m})$, (1) gives $\hat{x} > \mu_0$. Choose $c \in (\mu_0, \hat{x})$ and, for $n > \max\{1, c/(1-c)\}$, define

$$a_n(x) = \begin{cases} \dfrac{1}{2}\Big(\dfrac{x}{c}\Big)^n, & x \in [0,c], \\ 1 - \dfrac{1}{2}\Big(\dfrac{1-x}{1-c}\Big)^{n(1-c)/c}, & x \in [c,1]. \end{cases} \tag{11}$$

Then a_n satisfies all maintained assumptions: it is continuous ($a_n(c^-) = a_n(c^+) = 1/2$), increasing, $a_n(0) = 0$, $a_n(1) = 1$, strictly convex on $[0,c]$ (as $n > 1$) and strictly concave on $[c,1]$ (as $n(1-c)/c > 1$); in particular it is strictly S-shaped, so (A_1) holds. Moreover, as $n \to \infty$,

$$a_n(\mu_0) = \tfrac{1}{2}(\mu_0/c)^n \longrightarrow 0, \qquad a_n(\hat{x}) = 1 - \tfrac{1}{2}\big((1-\hat{x})/(1-c)\big)^{n(1-c)/c} \longrightarrow 1,$$

because $\mu_0 < c < \hat{x}$. Since $\bar{a}_0(E;\mu_0) \geq p_0(\hat{m}) a_n(\hat{x})$ and $p_0(\hat{m}) > 0$, there is n large enough that $\bar{a}_0(E;\mu_0) > a_n(\mu_0)$; fix such an n and set $a^* = a_n$. By Lemma 2(v), E is prejudiced at μ_0, and, exactly as in the last paragraph of the proof of Proposition 1, the map $\mu \mapsto \bar{a}_0(E;\mu) - a^*(\mu)$ is continuous on $(0,1)$, so E is prejudiced at every prior in an interval around μ_0.

B.4. Proof of Lemma 1. Let $\Lambda : \mathcal{E} \to \mathcal{H}$ map an experiment to the distribution of posteriors it induces at μ. By Lemma 2(i) and (vi), Λ maps $\mathcal{E}$ onto the set of distributions in $\mathcal{H}$ satisfying constraint (I). By Lemma 2(v), E is unprejudiced at μ if and only if $\Lambda(E)$ satisfies constraint (II). Hence E is feasible in (P) if and only if $\Lambda(E)$ is feasible in (P′), and the two feasible sets correspond to each other under Λ. By Lemma 2(iii), the objectives agree, $V(E;\mu) = \mathbb{E}_{\Lambda(E)}[a^*(x)]$. Consequently the two problems have the same value, and E attains it in (P) if and only if $\Lambda(E)$ attains it in (P′).

B.5. Proof of Theorem 1. Assume $V^* < V_{BP}$.

Step 1: we are in case (a) of Proposition 3. In case (c) the no-disclosure evidence is optimal in the relaxed problem, and it is feasible in (P′) because $\mathbb{E}_{\delta_\mu}[\psi] = \psi(\mu)$; hence $V^* = V_{BP}$, a contradiction. In case (b), by Proposition 3(b) the optimal evidence has posteriors $x^{**} < \mu$ and 1; the message inducing the posterior 1 has $p_0(m) = 0$ and the message inducing the posterior $x^{**} < \mu$ has $p_1(m) < p_0(m)$ by (1), so this evidence is a smoking gun and hence unprejudiced by Proposition 2; it is therefore feasible in (P′) and again $V^* = V_{BP}$, a contradiction. Therefore a^* is strictly S-shaped and $\mu < x^*$, and by Proposition 3(a) the standard Bayesian persuasion problem has the unique solution H_{BP} with

$$\mathrm{supp} H_{BP} = \{0, x^*\}, \qquad H_{BP}(\{x^*\}) = \frac{\mu}{x^*}, \qquad H_{BP}(\{0\}) = 1 - \frac{\mu}{x^*}.$$

Step 2: Every solution H^* of (P′) satisfies $H^* \neq H_{BP}$ and H^* is a mean-preserving spread of H_{BP}. Let H^* be an arbitrary solution of (P′). First, H_{BP} is infeasible in (P′), for otherwise $V^* = V_{BP}$. In particular, $H^* \neq H_{BP}$. Second, by Lemma 4, $\mathrm{supp} H^* \subseteq \{0\} \cup [x^*, 1]$. Finally, recall that H^* is a mean-preserving spread of H_{BP} if and only if $\mathbb{E}_{H^*}[f] \geq \mathbb{E}_{H_{BP}}[f]$ for every convex $f : [0,1] \to \mathbb{R}$. Let f be convex and let ℓ be the affine function that agrees with f at 0 and at x^*. Convexity of f implies $f \geq \ell$ outside $[0, x^*]$ and, in particular, $f \geq \ell$ on $\{0\} \cup [x^*, 1]$. Since $\mathbb{E}_{H^*}[x] = \mathbb{E}_{H_{BP}}[x] = \mu$ and ℓ is affine,

$$\mathbb{E}_{H^*}[f] \;\geq\; \mathbb{E}_{H^*}[\ell] \;=\; \ell(\mu) \;=\; \mathbb{E}_{H_{BP}}[\ell] \;=\; \mathbb{E}_{H_{BP}}[f],$$

where the last equality holds because $f = \ell$ on $\mathrm{supp} H_{BP} = \{0, x^*\}$. We thus have shown that H^* is a mean-preserving spread of H_{BP} and $H^* \neq H_{BP}$.

B.6. Proof of Theorem 2. *Part 1: $V^* < V_{BP}$ if and only if (7) holds.* Note first that $V^* \le V_{BP}$ always, as (P′) has the smaller feasible set.

Suppose a^* is concave (so that $x^* = 0$ by convention) or $x^* \le \mu$. By Proposition 3(c), $V_{BP} = a^*(\mu)$, and the no-disclosure evidence is feasible in (P′) because $\mathbb{E}_{\delta_\mu}[\psi] = \psi(\mu)$; hence $V^* = V_{BP}$, and (7) fails.

Suppose $x^* > \mu$. By Proposition 3(a), H_{BP} with support $\{0, x^*\}$ is the unique solution of the problem without constraint (II), and, using $\psi(0) = 0$,

$$\mathbb{E}_{H_{BP}}[\psi(x)] = \frac{\mu}{x^*}\psi(x^*).$$

Thus H_{BP} is feasible in (P′) if and only if $\mu\psi(x^*)/x^* \le \psi(\mu)$, i.e. if and only if $\psi(x^*)/x^* \le \psi(\mu)/\mu$. If this holds, then $V^* = V_{BP}$. If it fails, then, by Lemma 3(i), the maximum in (P′) is attained at some feasible H^*, which necessarily differs from H_{BP}; since H_{BP} is the unique maximizer of $\mathbb{E}_H[a^*]$ subject to $\mathbb{E}_H[x] = \mu$ alone, we get $V^* = \mathbb{E}_{H^*}[a^*] < V_{BP}$. Combining the two paragraphs, $V^* < V_{BP}$ if and only if (7) holds.

Part 2: structure of the solutions. Assume (7). By Lemma 3(iv), the no-prejudice constraint binds at every solution. By Lemma 4, every solution H^* satisfies $\text{supp}H^* \subseteq \{0\} \cup [x^*, 1]$ and $H^*(\{0\}) \ge 1 - \mu/x^* > 0$; in particular, the posterior 0 belongs to the support.

Every solution also satisfies $\max \text{supp}H^* > x^*$: otherwise $\text{supp}H^* \subseteq \{0, x^*\}$ and $\mathbb{E}_{H^*}[x] = \mu$ would give $H^* = H_{BP}$, which is infeasible under (7) by Part 1.

Finally, by Lemma 3(iii) there is a solution H° with at most three support points; by the above, $\text{supp}H^\circ = \{0, x', x''\}$ for some $x^* \le x' \le x'' \le 1$ (with $x' = x''$ if the support has only two points).

References

Alonso, R. and O. Câmara (2016): "Bayesian persuasion with heterogeneous priors," *Journal of Economic Theory*, 165, 672–706.

Boleslavsky, R. and C. Cotton (2015): "Grading standards and education quality," *American Economic Journal: Microeconomics*, 7, 248–279.

Boleslavsky, R. and K. Kim (2018): "Bayesian persuasion and moral hazard," Working paper, available at SSRN 2913669.

Bull, J. and J. Watson (2019): "Statistical evidence and the problem of robust litigation," *RAND Journal of Economics*, 50, 974–1003.

Che, Y.-K. and S. Severinov (2017): "Disclosure and legal advice," *American Economic Journal: Microeconomics*, 9, 188–225.

Daughety, A. F. and J. F. Reinganum (2000): "On the economics of trials: Adversarial process, evidence, and equilibrium bias," *Journal of Law, Economics, and Organization*, 16, 365–394.

Dewatripont, M. and J. Tirole (1999): "Advocates," *Journal of Political Economy*, 107, 1–39.

Doval, L. and V. Skreta (2024): "Constrained information design," *Mathematics of Operations Research*, 49, 78–106.

Dworczak, P. and G. Martini (2019): "The simple economics of optimal persuasion," *Journal of Political Economy*, 127, 1993–2048.

Dworczak, P. and A. Pavan (2022): "Preparing for the worst but hoping for the best: Robust (Bayesian) persuasion," *Econometrica*, 90, 2017–2051.

Hedlund, J. (2017): "Bayesian persuasion by a privately informed sender," *Journal of Economic Theory*, 167, 229–268.

Hörner, J. and N. S. Lambert (2021): "Motivational ratings," *Review of Economic Studies*, 88, 1892–1935.

Hu, J. and X. Weng (2021): "Robust persuasion of a privately informed receiver," *Economic Theory*, 72, 909–953.

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

Kaplow, L. (2011): "On the optimal burden of proof," *Journal of Political Economy*, 119, 1104–1140.

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

Kolotilin, A. (2018): "Optimal information disclosure: A linear programming approach," *Theoretical Economics*, 13, 607–635.

Kosterina, S. (2022): "Persuasion with unknown beliefs," *Theoretical Economics*, 17, 1075–1107.

Le Treust, M. and T. Tomala (2019): "Persuasion with limited communication capacity," *Journal of Economic Theory*, 184, 104940.

Lester, B., N. Persico, and L. Visschers (2012): "Information acquisition and the exclusion of evidence in trials," *Journal of Law, Economics, and Organization*, 28, 163–182.

Li, F. and P. Norman (2021): "Sequential persuasion," *Theoretical Economics*, 16, 639–675.

Lipnowski, E., D. Ravid, and D. Shishkin (2022): "Persuasion via weak institutions," *Journal of Political Economy*, 130, 2705–2730.

Milgrom, P. and J. Roberts (1986): "Relying on the information of interested parties," *RAND Journal of Economics*, 17, 18–32.

Min, D. (2021): "Bayesian persuasion under partial commitment," *Economic Theory*, 72, 743–764.

Mylovanov, T. and A. Zapechelnyuk (2024): "Constructive versus toxic argumentation in debates," *American Economic Journal: Microeconomics*, 16, 262–292.

Nguyen, A. and T. Y. Tan (2021): "Bayesian persuasion with costly messages," *Journal of Economic Theory*, 193, 105212.

Perez-Richet, E. (2014): "Interim Bayesian persuasion: First steps," *American Economic Review*, 104, 469–474.

Rosar, F. (2017): "Test design under voluntary participation," *Games and Economic Behavior*, 104, 632–655.

Sanchirico, C. W. (2001): "Character evidence and the object of trial," *Columbia Law Review*, 101, 1227–1311.

Schlag, K. H. (2026): "A 'non-guessing' theory for how to process information," Working paper, University of Vienna. SSRN 7020658.

Schrag, J. and S. Scotchmer (1994): "Crime and prejudice: The use of character evidence in criminal trials," *Journal of Law, Economics, and Organization*, 10, 319–342.

Shin, H. S. (1998): "Adversarial and inquisitorial procedures in arbitration," *RAND Journal of Economics*, 29, 378–405.

Wu, W. (2023): "Sequential Bayesian persuasion," *Journal of Economic Theory*, 214, 105763.

Zapechelnyuk, A. (2020): "Optimal quality certification," *American Economic Review: Insights*, 2, 161–176.

——— (2023): "On the equivalence of information design by uninformed and informed principals," *Economic Theory*, 76, 1051–1067.

Zhang, J. and J. Zhou (2016): "Information disclosure in contests: A Bayesian persuasion approach," *Economic Journal*, 126, 2197–2217.

Zubrickas, R. (2015): "Optimal grading," *International Economic Review*, 56, 751–776.