NBER WORKING PAPER SERIES

BAYESIAN NON-PERSUASION

Joshua S. Gans

Working Paper 35312
http://www.nber.org/papers/w35312

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2026

Thanks to Refine.ink, ChatGPT 5.5 and Claude Opus 4.8 for valuable research assistance. Responsibility for all errors remains my own. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Bayesian Non-Persuasion
Joshua S. Gans
NBER Working Paper No. 35312
June 2026
JEL No. D82, D83, K40

ABSTRACT

Rules against persuasion often focus on beliefs: an institution should not manipulate the information on which a decision rests. We show that such rules can miss a distinct source of directional influence. An institution can steer an outcome without changing beliefs by selecting among authorised procedures that translate the same belief into different actions. Jury instructions provide the leading example: a judge may alter no juror's assessment of the facts yet still affect the verdict through the permissible formulation of the law. The gap is sharpest near a decision threshold. There, the scope for belief movement that preserves the default action vanishes, while procedural steering can remain maximal whenever authorised procedures disagree near the boundary, even with genuinely informative communication. In general, exposure to procedural steering is the concavification of a local steering score. The geometry of Bayesian persuasion thus reappears in reverse: it measures not the value of manipulating beliefs, but the directional discretion left open by institutional rules.

Joshua S. Gans
University of Toronto
Rotman School of Management
and NBER
joshua.gans@rotman.utoronto.ca

"He may analyze and dissect the evidence, but he may not either distort it or add to it."

— *Quercia v. United States*, 289 U.S. 466, 470 (1933)

1 Introduction

An institution can tilt a decision towards a favoured party without moving anyone's beliefs. A trial judge who alters no juror's assessment of the facts may still steer the verdict by selecting, among authorised ways of explaining the law, the one whose mapping from belief to action runs in the preferred direction. Rules meant to keep judges neutral typically police evidence and the belief movement it induces. Rules that discipline only that margin leave procedural selection unchecked. We show that the omission can matter most precisely where outcome-neutral belief movement is most tightly constrained while procedural exposure remains complete.

The tension is sharpest in the law of jury instructions. A trial judge's explanation of the law is meant to matter: it tells jurors what must be proved, whose burden it is, and how a factual finding bears on the verdict. A clearer formulation may change a verdict by improving comprehension, and that change is not, in itself, bias—instruction comprehensibility is known to be consequential for adjudication (Charrow and Charrow, 1979; Tiersma, 1993). The judge's influence is, in the words of the Supreme Court, "necessarily and properly of great weight," so that "his lightest word or intimation is received with deference, and may prove controlling" (*Starr v. United States*, 153 U.S. 614, 626 (1894)).[1] The institutional problem is therefore not communication without influence, which would have no instrumental value for the represented decision. It is influence that is directionally deployed.

We make this idea precise as a *procedural channel of persuasion*: holding the evidentiary assessment fixed, a judge can steer outcomes by selecting among authorised procedures that translate that assessment into a verdict.[2] That such selection can move outcomes at a single belief is immediate; it is the setup, not the result. The result is quantitative. We ask how much directional influence survives once information must be Bayes plausible, procedures are confined to an authorised menu, and their non-directional use is pinned

[1] See also *Bollenbach v. United States*, 326 U.S. 607, 612 (1946) ("Particularly in a criminal trial, the judge's last word is apt to be the decisive word."). In federal trials the judge may comment on the evidence but "may not either distort it or add to it" (*Quercia v. United States*, 289 U.S. 466, 470 (1933)); many state systems restrict or prohibit such comment more sharply (*Bute v. Illinois*, 333 U.S. 640, 648 n.3 (1948)).

[2] The same reduced form arises beyond jury instructions. An auditor may select among authorised reporting standards with different warning implications at the same assessment of a firm's viability. A genetic counsellor may select among consultation protocols with different action implications at the same risk assessment.

down by a target-independent comparator. We then ask how that quantity behaves exactly where a neutrality rule is conventionally thought to bite hardest.

The answer is a divergence between two notions of neutrality. Consider a posterior assessment near a decision threshold—the evidentiary standard at which a verdict for one party becomes warranted. Here regulators of evidence are most anxious because a small belief movement can be decisive. The scope for *outcome-neutral* belief movement collapses: as the prior approaches the threshold, the maximum posterior variance consistent with leaving the default action optimal converges to zero. Yet over the same region the *procedural* channel remains at full strength: an authorised alternative procedure continues to flip the verdict at its comparator-relative maximum. When the comparator applies the default procedure, procedural selection determines the verdict with probability one. Where belief-based persuasion runs out of room, procedural steering does not.

This divergence is not an artefact of silence. One might suspect that the procedural channel survives only because outcome-neutral communication can degenerate into saying nothing. It does not. A non-degenerate, genuinely informative experiment can keep both posteriors inside the pivotal region and preserve the entire procedural advantage even as its variance vanishes. Informative communication and complete procedural exposure are compatible.

The persistence result is not automatic. In the general model, steering vanishes near an action boundary when authorised disagreement stays away from that boundary. Persistent exposure requires boundary contact. The binary threshold model is the leading case in which it occurs.

Behind this result is a general principle that inverts the central tool of information design. Kamenica and Gentzkow (2011) show that the most a sender can achieve by Bayes-plausible choice of information is the concavification of a payoff function over beliefs.[3] We show that the most an institution is *exposed to*, when it authorises a menu of belief-to-action procedures, is the concavification of a *local steering score*—a measure of how far procedural selection can shift the target action at a belief, relative to the comparator. The familiar geometry reappears, but its role is reversed: it measures the directional flexibility a neutrality rule has failed to close off, not gains a persuader can capture.

For jury instructions, the reduced form must be read carefully. A belief is a juror's assessment of a fact required for liability or conviction. The authorised procedures are formulations or comment practices that correctly state the governing law but may differ in how jurors translate that assessment into a verdict. A misstated burden or a distorted

[3]Judicial decision-making is among the motivating examples of the persuasion literature; here the same example motivates its mirror image.

record is not an extreme procedure in the model but an unauthorised one. The concern is narrower: regulating only the movement of beliefs leaves unchecked the directional selection of procedures that remain within the authorised menu.

Our exercise is diagnostic rather than a complete design of institutions. We take the authorised menu and the comparator as given and locate where a neutrality standard must operate. Target independence makes the comparator independent of the favoured action, but not, on its own, normatively neutral. Choosing menus, comparator classes, and directional budgets is a distinct problem, developed in companion work (Gans, 2026). Isolating the diagnostic step clarifies the point a neutrality rule must confront first: consequential communication is not biased merely because it changes a decision, but authorised procedural discretion can be biased when it is selected for its direction.

Related literature. The contribution is closest to, and reverses the direction of, the information-design literature. Bayesian persuasion selects a Bayes-plausible distribution of beliefs to advance a sender's objective (Kamenica and Gentzkow, 2011; Kamenica, 2019). Work with a separate designer studies how constraints on information shape induced outcomes (Taneva, 2019; Mathevet et al., 2020), and rules of persuasion govern how communication translates into action (Glazer and Rubinstein, 2004). By contrast, we fix the authorised procedure menu and the comparator and isolate a different margin—the target-directed flexibility left by selecting among belief-to-action mappings at the *same* posterior. Our fixed-decision benchmark is closely related to the payoff–belief geometry of de Lara and Gossner (2020); here its role is to explain why unchanged outcomes cannot define useful neutral communication.

Plan. Section 2 sets out the fixed-decision benchmark and shows that no-impact neutrality has no decision value. Section 3 defines procedural steering and derives its concavification representation. Section 4 compares belief movement and procedural steering near an action boundary. Section 5 establishes the pivotal-wedge result and its informative-signal version. Section 6 draws out the implications for neutrality rules.

2 Outcome Neutrality in a Fixed Decision Problem

We first consider a setting in which communication changes beliefs but does not change the rule by which the receiver selects an action. This is the natural benchmark for the requirement that communication leave a decision unchanged.

There is a finite state space Ω, a finite action set A, and a receiver with prior belief $\mu_0 \in \Delta(\Omega)$. At belief μ, the expected payoff from action a is

$$U(a, \mu) = \sum_{\omega \in \Omega} \mu(\omega) u(a, \omega),$$

and the value of choosing optimally is

$$v(\mu) = \max_{a \in A} U(a, \mu).$$

Let a_0 be a default action that is optimal at the prior: $a_0 \in \arg\max_{a \in A} U(a, \mu_0)$. Its action cell is

$$N_0 = \{\mu \in \Delta(\Omega) : a_0 \in \arg\max_{a \in A} U(a, \mu)\}.$$

Thus N_0 is the set of beliefs at which the receiver may continue to use the default action. Since expected payoffs are linear in beliefs, N_0 is a closed convex polytope.

A signal is represented by a Borel probability distribution τ over posterior beliefs. It is feasible when it satisfies Bayes plausibility:

$$\int_{\Delta(\Omega)} \mu \, d\tau(\mu) = \mu_0.$$

Throughout, we identify information structures with Bayes-plausible posterior distributions. This is without loss in the finite-state setting: every signal induces a Bayes-plausible distribution over posteriors, and the splitting lemma implies that every Bayes-plausible distribution over posteriors is implementable by some signal (Kamenica and Gentzkow, 2011).

The instrumental decision value of the signal is

$$V(\tau) = \int_{\Delta(\Omega)} v(\mu) \, d\tau(\mu) - v(\mu_0).$$

This notion of value is intentionally narrow. It measures improvement in the represented action problem, rather than psychological, expressive, or legal value of receiving an explanation.

Definition 1 (Outcome-neutral information)**.** *A feasible posterior distribution τ is* outcome-neutral *for the default action if $\tau(N_0) = 1$. If ties are resolved in favour of a_0, the receiver takes the default action after every posterior realisation of an outcome-neutral signal.*

The definition gives the literal no-impact interpretation its most favourable form: the

default need only remain optimal, with unchanged behaviour following when the institution resolves ties towards it. The next result records the difficulty with imposing this requirement on useful communication.

Proposition 1 (No-impact neutrality has no decision value)**.** *For any feasible posterior distribution τ,*

$$V(\tau) \geq 0, \qquad V(\tau) = 0 \quad \Longleftrightarrow \quad \tau(N_0) = 1.$$

Consequently, every signal with positive instrumental decision value violates outcome neutrality with positive probability.

Proof. Because a_0 is optimal at μ_0, we have $v(\mu_0) = U(a_0, \mu_0)$. Bayes plausibility and linearity of $U(a_0, \cdot)$ give

$$\int U(a_0, \mu)\, d\tau(\mu) = U(a_0, \mu_0).$$

It follows that

$$V(\tau) = \int_{\Delta(\Omega)} \underbrace{\big[v(\mu) - U(a_0, \mu)\big]}_{\text{gain from departing from the default}} \qquad d\tau(\mu).$$

The integrand is non-negative and equals zero exactly when a_0 is optimal, that is, when $\mu \in N_0$. The integral therefore vanishes if and only if $\tau(N_0) = 1$. □

Proposition 1 separates impact from bias. Information can keep an action unchanged, but then it cannot improve that action choice in the fixed problem. If a judge's clarification is required to improve a represented decision, the communication must sometimes be behaviourally consequential. The appropriate neutrality question is then not whether the action changed, but whether the route by which it changed was selected to favour an outcome.

Outcome neutrality may nevertheless accommodate belief movement that has no instrumental decision value. In the binary analysis below, we measure this possibility by posterior variance. That measure will provide a direct comparison with procedural steering near the default action boundary.

2.1 A binary threshold benchmark

It is useful to carry one example through the argument. Let $\Omega = \{0, 1\}$, and write $p = \Pr(\omega = 1)$ for the receiver's posterior belief. The receiver chooses between action 0 and action 1. We normalise the expected payoff from action 0 to zero and write the payoff from action 1 as

$$U(0, p) = 0, \qquad U(1, p) = p - h, \qquad h \in (0, 1).$$

Thus, action 1 is optimal precisely when the posterior exceeds the threshold h. At the threshold, we resolve the tie in favour of action 0. One may read action 1 as a verdict for a specified party and h as the evidentiary standard that must be met before that verdict is selected.

Suppose the prior satisfies $p_0 < h$. Without communication the receiver takes action 0. A posterior experiment is outcome-neutral if it never moves the belief above h, or equivalently if its support lies in $[0, h]$. It can nevertheless be statistically informative. To make the remaining scope for such communication transparent, measure belief movement by posterior variance,

$$K(\tau) = \mathrm{Var}_\tau(p).$$

Proposition 2 (Binary non-persuasion)**.** *Suppose $p_0 < h$. The largest posterior variance generated by any outcome-neutral information structure is*

$$\max_{\tau:\, \mathbb{E}_\tau[p]=p_0,\ \tau([0,h])=1} K(\tau) = \underbrace{p_0}_{\text{room below}} \underbrace{(h - p_0)}_{\text{gap to the action threshold}}.$$

In particular, this capacity converges to zero as $p_0 \uparrow h$.

Proof. For any posterior $p \in [0, h]$, we have $p^2 \leq hp$. Bayes plausibility gives $\mathbb{E}_\tau[p] = p_0$, and hence

$$\mathrm{Var}_\tau(p) = \mathbb{E}_\tau[p^2] - p_0^2 \leq hp_0 - p_0^2 = p_0(h - p_0).$$

The bound is attained by placing probability p_0/h on $p = h$ and the remaining probability on $p = 0$. Both posterior beliefs preserve action 0 under the stated tie-breaking rule. □

In Bayesian persuasion, a sender who favours action 1 seeks to place posterior probability above the threshold h. In this literal version of Bayesian non-persuasion, those posterior beliefs are excluded: beliefs may move, but only within the region where action 0 remains optimal. Proposition 2 shows that this remaining room collapses as the prior approaches the decision threshold. The procedural analysis below asks whether the collapse also removes the ability to steer once the institution authorises more than one rule mapping beliefs into actions.

3 Procedural Communication

We now allow communication to specify, or select among, authorised decision procedures. This extension is natural for jury instructions: the instruction does not purport to tell jurors

which factual belief to hold, but it does tell them how their factual assessment maps into a verdict. The same reduced-form object can represent a reporting standard or an advisory protocol.

Let $\mathcal{R}$ be a finite set of institutionally authorised procedures. A procedure $\rho \in \mathcal{R}$ is a measurable map

$$\alpha_\rho : \Delta(\Omega) \longrightarrow A,$$

where $\alpha_\rho(\mu)$ is the action selected after posterior belief μ. The procedure includes any applicable tie-breaking convention. This formulation does not require procedures to be payoff-optimal rules for a single receiver utility function: they are alternative rules that the institution has permitted for the communication problem being considered.

The timing/no-signalling convention is part of the object being measured. An experiment first induces a posterior μ. The selector observes μ and the authorised menu, chooses a procedure, and the receiver treats μ as the complete summary of state-relevant evidence. The selection rule is common knowledge and carries no state information beyond μ. If the receiver updates from the procedure choice itself, there is an additional signalling channel; the capacity below isolates only the residual map from a common posterior into an action.

An institution that permits several procedures must specify their non-directional use. Let $B \subseteq A$ be the set of actions towards which steering is assessed, and let

$$\overline{q} : \Delta(\Omega) \longrightarrow \Delta(\mathcal{R})$$

be a measurable authorised comparator rule, so each coordinate $\overline{q}_\rho$ is Borel measurable. At posterior μ, the comparator selects procedure ρ with probability $\overline{q}_\rho(\mu)$. Intuitively, it is the rule used absent an objective to favour a target action. It may depend on the posterior because an authorised instruction or reporting rule may properly depend on the evidentiary posture. The comparator is a primitive benchmark, not neutral by definition: its normative force must come from the governing legal or professional rule.

For target action $b \in B$, the comparator produces b at posterior μ with probability

$$\overline{m}_b(\mu) = \sum_{\rho \in \mathcal{R}} \overline{q}_\rho(\mu) \mathbf{1}\{\alpha_\rho(\mu) = b\}.$$

A sender wishing to increase the probability of b selects, at that same posterior, whichever authorised procedure is most favourable to b. The sender's local protocol-selection advant-

age is therefore

$$r_b(\mu \mid \overline{q}) = \underbrace{\max_{\rho \in \mathcal{R}} \mathbf{1}\{\alpha_\rho(\mu) = b\}}_{\text{best authorised procedure for target } b} - \underbrace{\overline{m}_b(\mu)}_{\text{target-independent comparator}}. \tag{1}$$

This score is not a welfare loss and is not the probability of action b. It measures how much discretion over an authorised procedure can increase that probability, conditional on the same belief. Because $\mathcal{R}$ is finite, the pointwise maximum in (1) is implementable by a measurable selector: fix an ordering of $\mathcal{R}$, choose the first procedure that produces b when at least one does, and otherwise choose the first procedure in the ordering. The relevant selection sets are measurable because each α_ρ is measurable and A is finite.

3.1 Protocol selection and evidence selection

The conditioning in (1) is important. Procedure selection is not the only route through which a sender can favour an action. A sender might instead choose evidence that makes the comparator itself more likely to select the target action. To distinguish these routes, let $\overline{\tau}$ be the comparator's posterior experiment and let τ be a candidate experiment. For a measurable procedure-selection rule $q : \Delta(\Omega) \to \Delta(\mathcal{R})$, define

$$P_b(q; \tau) = \int_{\Delta(\Omega)} \sum_{\rho \in \mathcal{R}} q_\rho(\mu) \mathbf{1}\{\alpha_\rho(\mu) = b\} \, d\tau(\mu).$$

Relative to the full comparator environment $(\overline{q}, \overline{\tau})$, total directional advantage decomposes as

$$\begin{aligned} P_b(q; \tau) - P_b(\overline{q}; \overline{\tau}) = {} & \underbrace{P_b(q; \tau) - P_b(\overline{q}; \tau)}_{\text{protocol-selection component}} \\ & + \underbrace{P_b(\overline{q}; \tau) - P_b(\overline{q}; \overline{\tau})}_{\text{evidence-selection component}}. \end{aligned} \tag{2}$$

The two components differ in what is held fixed:

	Held fixed	Changed
Protocol selection	realised posterior distribution	procedure-selection rule
Evidence selection	comparator procedure rule	posterior distribution

The paper maximises the first component over Bayes-plausible experiments. This exposes vulnerability created by procedural discretion while holding realised evidence common between candidate and comparator. A rule against total directional advantage would also have to regulate evidence selection.

For a posterior domain $X \subseteq \Delta(\Omega)$, write

$$T_X(\mu_0) = \left\{ \tau : \int_X \mu \, d\tau(\mu) = \mu_0,\ \tau(X) = 1 \right\}.$$

Two domains are important. Taking $X = \Delta(\Omega)$ allows unrestricted information. Taking $X = N_0$ permits only information that is outcome-neutral under the fixed default decision problem. In either case, the procedural steering capacity towards b is the supremum

$$S^*_{b,X}(\mu_0 \mid \overline{q}) = \sup_{\tau \in T_X(\mu_0)} \int_X \underbrace{r_b(\mu \mid \overline{q})}_{\text{local steering score}} \underbrace{d\tau(\mu)}_{\text{belief allocation}}. \tag{3}$$

The capacity is thus a product of two forces. The steering score records how much an authorised procedure can favour the target at a given belief, while the belief allocation records how much posterior mass an information structure can place there; capacity is large only when an experiment can concentrate beliefs at posteriors where procedures disagree. It measures the greatest protocol-selection advantage obtainable when information can be chosen subject to Bayes plausibility and the specified posterior domain. In particular, S^*_{b,N_0} asks how far procedure selection can steer outcomes even while beliefs remain outcome-neutral under the default rule. The supremum formulation is deliberate: measurability alone need not guarantee an optimal posterior distribution.

Assumption 1 (Regular steering score)**.** *The posterior domain X is non-empty, compact, and convex, contains μ_0, and $r_b(\cdot \mid \overline{q})$ is upper semicontinuous on X for every $b \in B$.*

The regularity condition ensures that the supremum in (3) is attained. It is substantive: arbitrary measurable procedures and arbitrary measurable comparator rules need not generate an upper semicontinuous steering score. In applications it must be verified directly. In the threshold model below, the tie-breaking conventions make the procedure-disagreement wedge closed and the score is a constant multiple of that closed set's indicator, hence upper semicontinuous. Without such regularity, the capacity remains well defined by (3), but a maximising posterior distribution need not exist.

Let $\operatorname{cav}_X f$ denote the pointwise smallest concave majorant of f on X, equivalently the infimum of concave functions $\phi : X \to \mathbb{R}$ satisfying $\phi \geq f$. Let $\dim(X)$ denote the affine dimension of X. The substantive step came earlier: authorised procedures can disagree at a common posterior. The next theorem applies familiar information-design geometry to measure the resulting exposure.

Theorem 1 (Concavifying procedural vulnerability)**.** *Under Assumption 1, for every target action $b \in B$,*

$$S^*_{b,X}(\mu_0 \mid \overline{q}) = \operatorname{cav}_X\big[r_b(\cdot \mid \overline{q})\big](\mu_0).$$

A maximising posterior distribution exists and can be chosen with support on at most $\dim(X)+1$ *posteriors. In particular, with $X = \Delta(\Omega)$, no more than $|\Omega|$ posteriors are required.*

Proof. For $\mu \in X$, let $T_X(\mu)$ denote the set of probability measures on X with barycentre μ. The set $T_X(\mu_0)$ is compact in the weak topology because X is compact and the barycentre constraint is closed. Since $r_b(\cdot \mid \overline{q})$ is bounded and upper semicontinuous, the map $\tau \mapsto \int_X r_b(\mu \mid \overline{q})\, d\tau(\mu)$ is upper semicontinuous on $T_X(\mu_0)$. A maximiser therefore exists.

Define, for each $\mu \in X$,

$$\Gamma_b(\mu) = \max_{\tau \in T_X(\mu)} \int_X r_b(\mu' \mid \overline{q})\, d\tau(\mu').$$

If $\tau_1 \in T_X(\mu_1)$, $\tau_2 \in T_X(\mu_2)$, and $\lambda \in [0,1]$, then $\lambda\tau_1 + (1-\lambda)\tau_2$ is feasible at $\lambda\mu_1 + (1-\lambda)\mu_2$. Hence Γ_b is concave. A degenerate posterior distribution at μ is feasible at μ, so $\Gamma_b(\mu) \geq r_b(\mu \mid \overline{q})$. Thus Γ_b is a concave majorant of $r_b(\cdot \mid \overline{q})$, and $\Gamma_b \geq \operatorname{cav}_X[r_b(\cdot \mid \overline{q})]$.

Conversely, let ϕ be any concave majorant of $r_b(\cdot \mid \overline{q})$. For every posterior distribution feasible at μ_0, Jensen's inequality gives

$$\int_X r_b(\mu \mid \overline{q})\, d\tau(\mu) \leq \int_X \phi(\mu)\, d\tau(\mu) \leq \phi(\mu_0).$$

Taking the supremum over τ, and then the infimum over concave majorants, gives $\Gamma_b(\mu_0) \leq \operatorname{cav}_X[r_b(\cdot \mid \overline{q})](\mu_0)$. The equality follows.

It remains to justify the support bound. The set of maximisers is a non-empty compact face of $T_X(\mu_0)$, because the objective is affine in τ. By the Krein–Milman theorem,[4] choose an extreme point τ^* of this face; it is also an extreme point of $T_X(\mu_0)$. Let $d = \dim(X)$. If $\operatorname{supp}(\tau^*)$ contained $d+2$ distinct points, we could choose pairwise disjoint Borel sets $E_1, \ldots, E_{d+2}$ with positive τ^*-mass. Let

$$x_i = \frac{1}{\tau^*(E_i)} \int_{E_i} \mu\, d\tau^*(\mu) \qquad (i = 1, \ldots, d+2).$$

The points x_i lie in the d-dimensional affine hull of X, so they are affinely dependent.

[4]The Krein–Milman theorem states that a non-empty compact convex subset of a locally convex Hausdorff topological vector space coincides with the closed convex hull of its extreme points; in particular, such a set has at least one extreme point. We apply it within the space of finite signed Borel measures on X under the weak topology, in which the face of maximisers is non-empty, compact, and convex.

There are coefficients c_i, not all zero, such that $\sum_i c_i = 0$ and $\sum_i c_i x_i = 0$. Define the signed measure

$$\sigma = \sum_{i=1}^{d+2} \frac{c_i}{\tau^*(E_i)} \, \tau^*|_{E_i}.$$

Then $\sigma(X) = 0$ and $\int_X \mu \, d\sigma(\mu) = 0$. For sufficiently small $\varepsilon > 0$, the measures $\tau^* + \varepsilon\sigma$ and $\tau^* - \varepsilon\sigma$ are distinct probability measures on X with barycentre μ_0, and

$$\tau^* = \frac{1}{2}(\tau^* + \varepsilon\sigma) + \frac{1}{2}(\tau^* - \varepsilon\sigma),$$

contradicting the extremality of τ^*. Hence $\operatorname{supp}(\tau^*)$ has at most $d + 1$ points. When $X = \Delta(\Omega)$, $d = |\Omega| - 1$, so at most $|\Omega|$ posteriors are required. □

In ordinary persuasion, concavification allocates posterior beliefs to improve a sender's payoff. Here it allocates posterior beliefs to expose a protocol channel: beliefs at which authorised procedures disagree receive value because a procedure can then be chosen directionally. The formula does not say that any such selection is legitimate. It identifies the exposure that the menu and comparator leave available.

The restricted version, with $X = N_0$, is the point of contact with the fixed-decision benchmark. Proposition 1 says that every posterior experiment supported on N_0 has zero value for changing the default action. Theorem 1 says that the very same experiment may nevertheless have high value for selecting a procedure that favours a target action. We next compare these objects near a general decision boundary, before deriving an exact binary formula.

4 Steering at a Boundary

The central comparison can be stated without restricting the state space to two states. Return to the fixed decision problem in Section 2. For a competitor action $a \neq a_0$, define its payoff gap relative to the default as

$$g_a(\mu) = U(a_0, \mu) - U(a, \mu).$$

The default cell satisfies $g_a(\mu) \geq 0$, with equality on the face where the competitor ties the default. Write

$$d_a(\mu_0) = g_a(\mu_0), \qquad g_a^{\max} = \max_{\mu \in N_0} g_a(\mu).$$

The first number is the receiver's payoff-gap distance from the boundary; the second is the greatest room in that direction while remaining in the default cell. The natural measure of outcome-neutral movement towards this boundary is

$$K_a^{\perp}(\tau) = \int_{N_0} \big(g_a(\mu) - g_a(\mu_0)\big)^2 \, d\tau(\mu).$$

It is posterior variance in the payoff-gap direction, rather than in directions tangent to the boundary. Accordingly, the higher-dimensional comparison below concerns boundary-normal belief dispersion, not all posterior movement inside the default cell.

For the procedural comparison, suppose the comparator always applies a default procedure ρ_0 satisfying $\alpha_{\rho_0}(\mu) = a_0$ for every $\mu \in N_0$. For a target action $b \neq a_0$, define the default-neutral procedure-disagreement set

$$G_b = \left\{\mu \in N_0 : \alpha_{\rho_0}(\mu) \neq b \text{ and } \alpha_\rho(\mu) = b \text{ for some } \rho \in \mathcal{R}\right\}.$$

On G_b, an authorised procedural departure changes the action towards the target while beliefs remain in the default action cell. With a deterministic comparator, the local steering score on N_0 is the indicator of G_b. The action a indexes the boundary being approached, while b indexes the steering target; in the binary application they coincide. The key qualification is simple: steering persists near a boundary only if authorised procedures disagree arbitrarily close to it.

Proposition 3 (Boundary contact)**.** *Suppose G_b is closed, $b \neq a_0$, and the comparator always applies a procedure ρ_0 satisfying $\alpha_{\rho_0}(\mu) = a_0$ on N_0. Fix $a \neq a_0$.*

1. *For every outcome-neutral posterior distribution τ,*

$$K_a^{\perp}(\tau) \leq d_a(\mu_0)\big(g_a^{\max} - d_a(\mu_0)\big).$$

2. *If there is $\gamma > 0$ such that $g_a(\mu) \geq \gamma$ for every $\mu \in G_b$, then*

$$S^*_{b,N_0}(\mu_0 \mid \overline{q}) \leq \min\left\{1, \frac{d_a(\mu_0)}{\gamma}\right\}.$$

   Under this separation condition, protocol-selection capacity goes to zero whenever the prior approaches the boundary in the g_a-direction.

3. *If a sequence $\mu_0^n \in G_b$ satisfies $d_a(\mu_0^n) \to 0$, then*

$$S^*_{b,N_0}(\mu_0^n \mid \overline{q}) = 1 \quad \textit{for every } n,$$

while the maximum of $K_a^{\perp}$ over outcome-neutral posterior distributions at μ_0^n converges to zero.

Proof. For part 1, if μ is distributed according to an outcome-neutral τ, the random variable $Y = g_a(\mu)$ belongs to $[0, g_a^{\max}]$. Since g_a is linear and τ is Bayes plausible, $\mathbb{E}[Y] = d_a(\mu_0)$. The inequality $Y^2 \leq g_a^{\max} Y$ therefore gives

$$K_a^{\perp}(\tau) = \mathrm{Var}(Y) \leq d_a(\mu_0)\big(g_a^{\max} - d_a(\mu_0)\big).$$

For part 2, the deterministic comparator implies that the protocol-selection advantage of an outcome-neutral experiment is $\tau(G_b)$. If $g_a(\mu) \geq \gamma$ on G_b, then

$$d_a(\mu_0) = \int_{N_0} g_a(\mu)\, d\tau(\mu) \geq \int_{G_b} g_a(\mu)\, d\tau(\mu) \geq \gamma\tau(G_b).$$

Maximising over outcome-neutral τ gives the claimed bound.

For part 3, at any prior $\mu_0^n \in G_b$, the degenerate posterior distribution at μ_0^n remains outcome-neutral and assigns probability one to G_b. Its steering score is one, which is the largest possible value. Part 1 implies convergence of the maximum normal belief variance because $0 \leq d_a(\mu_0^n) \leq g_a^{\max}$ and $d_a(\mu_0^n) \to 0$. □

Proposition 3 formalises the qualification. Proximity to a boundary is not enough. If procedure-disagreement beliefs remain a positive distance away, both boundary-normal information and steering vanish as the prior reaches the boundary. Persistent steering requires disagreement to touch the boundary. In the binary model, contact occurs throughout an interval of priors.

5 The Pivotal Wedge

We now add procedural discretion to the binary benchmark of Section 2.1. The receiver still faces two actions and the same posterior p. The benchmark threshold h is implemented by a high-threshold procedure H, while the institution also authorises a low-threshold procedure L. The two procedures select actions according to

$$\alpha_L(p) = \begin{cases} 0, & p < \ell, \\ 1, & p \geq \ell, \end{cases} \qquad \alpha_H(p) = \begin{cases} 0, & p \leq h, \\ 1, & p > h, \end{cases} \qquad 0 < \ell < h < 1. \tag{4}$$

We take H to implement the optimal action rule in the fixed benchmark problem, including its default-favouring tie-break at h. It is the *demanding* procedure: it requires a high

posterior before selecting action 1. Procedure L is an authorised alternative whose use is evaluated relative to that default rule, not a second payoff-optimal rule for the same fixed decision problem. It is the *permissive* procedure: it selects action 1 at weaker posteriors. The procedures agree below ℓ and above h. Between their thresholds they disagree: L selects action 1 and H selects action 0. We call the closed interval

$$W = [\ell, h]$$

the *pivotal wedge*. The tie-breaking conventions in (4) close the wedge and satisfy the regularity condition used in Theorem 1. In applications, L is an effective map induced by a valid protocol, not a licence to lower h. One legal example is a choice between a conventional burden instruction and an approved plain-language formulation that states the same elements and burden. The evidentiary assessment and legal standard are unchanged, but the formulation can alter how jurors apply that assessment to the verdict (Charrow and Charrow, 1979; Tiersma, 1993).

Let the target-independent comparator select L with probability $\chi \in [0, 1]$ and H with probability $1 - \chi$, independently of the target action. On the pivotal wedge, selecting L rather than relying on the comparator increases the probability of action 1 by $1 - \chi$. Outside the wedge, the procedures agree, and procedure selection provides no advantage towards action 1. Thus the local steering score is

$$r_1(p \mid \bar{q}) = (1 - \chi)\mathbf{1}\{p \in W\}. \tag{5}$$

For any feasible posterior distribution τ, its protocol-selection advantage towards action 1 is consequently

$$(1 - \chi)\tau(W).$$

The procedural question is therefore a geometric one: how much Bayes-plausible posterior probability can be placed inside the wedge?

We compare that procedural exposure with the amount of outcome-neutral information. Treat H as the default procedure, and suppose the prior satisfies $p_0 < h$, so the default action is 0. Information is outcome-neutral under H exactly when its posterior support lies in $[0, h]$. As in Section 2.1, we measure belief movement by the posterior variance $K(\tau) = \mathrm{Var}_\tau(p)$. The following proposition contains the paper's main contrast.

Proposition 4 (Pivotal-wedge contrast)**.** *Suppose H is the default procedure and $p_0 < h$.*

1. *The binary non-persuasion capacity is*

$$\max_{\tau:\,\mathbb{E}_\tau[p]=p_0,\ \tau([0,h])=1} K(\tau) = \underbrace{p_0}_{\text{room below gap}} \quad \underbrace{(h-p_0)}_{\text{to the default threshold}}.$$

2. *The maximum outcome-neutral protocol-selection advantage towards action 1 is*

$$S^*_{1,[0,h]}(p_0 \mid \bar{q}) = (1-\chi)\begin{cases} \dfrac{p_0}{\ell}, & p_0 < \ell, \\ 1, & \ell \le p_0 < h. \end{cases}$$

3. *Along any sequence $p_0^n \in [\ell, h)$ with $p_0^n \uparrow h$,*

$$\max_{\tau:\,\mathbb{E}_\tau[p]=p_0^n,\ \tau([0,h])=1} K(\tau) \longrightarrow 0, \qquad S^*_{1,[0,h]}(p_0^n \mid \bar{q}) = 1-\chi.$$

In particular, if the comparator applies H, so $\chi = 0$, procedural steering remains maximal as outcome-neutral belief dispersion vanishes.

Proof. For part 1, any outcome-neutral posterior p lies in $[0, h]$, so $p^2 \le hp$. Bayes plausibility gives $\mathbb{E}_\tau[p] = p_0$, and therefore

$$K(\tau) = \mathrm{Var}_\tau(p) = \mathbb{E}_\tau[p^2] - p_0^2 \le hp_0 - p_0^2 = p_0(h-p_0).$$

The bound is attained by placing probability p_0/h on $p = h$ and the remaining probability on $p = 0$, which is outcome-neutral under H.

For part 2, equation (5) shows that maximising protocol-selection advantage is equivalent to maximising $\tau(W)$ among posterior distributions supported on $[0, h]$. If $p_0 < \ell$, every posterior in W is at least ℓ, so Bayes plausibility implies

$$p_0 = \mathbb{E}_\tau[p] \ge \ell\tau(W).$$

Thus $\tau(W) \le p_0/\ell$. This bound is attained by placing probability p_0/ℓ on $p = \ell$ and the remaining probability on $p = 0$, which is outcome-neutral under H. If $\ell \le p_0 < h$, the degenerate distribution at p_0 is supported both in W and in $[0, h]$, so it assigns probability one to the wedge. Multiplying the maximal wedge probability by $1 - \chi$ yields the formula.

Part 3 follows directly from parts 1 and 2. When $p_0^n \uparrow h$, $p_0^n(h - p_0^n) \to 0$; when each p_0^n is in the wedge, the maximal wedge probability is one. □

The formula in part 2 is the concavification of $(1-\chi)\mathbf{1}_{[\ell,h]}$ on the outcome-neutral domain

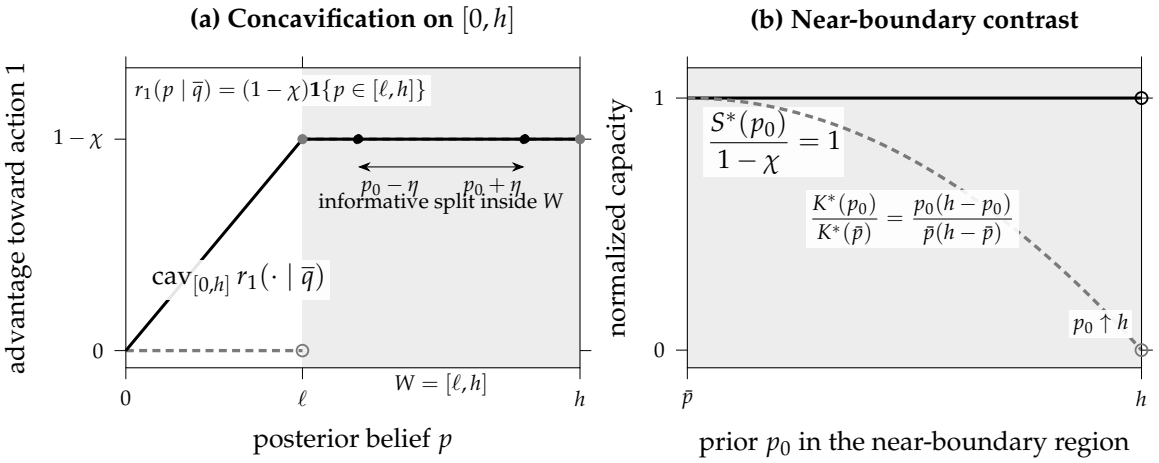

Figure 1: **Concavification and persistent procedural steering in the pivotal wedge.** Panel (a) plots the local score $r_1(p \mid \bar{q}) = (1-\chi)\mathbf{1}\{p \in [\ell,h]\}$ and its concavification on the outcome-neutral domain $X=[0,h]$. The marked interior split illustrates the informative outcome-neutral experiment in Corollary 1. Panel (b) uses $S^*(p_0) \equiv S^*_{1,[0,h]}(p_0 \mid \bar{q})$ and compares normalized capacities near the boundary: $S^*(p_0)/(1-\chi)=1$ throughout the displayed part of the wedge, while $K^*(p_0)=p_0(h-p_0)$ converges to zero as $p_0 \uparrow h$. It is drawn over $p_0 \in [\bar{p},h)$, where $\bar{p}=\max\{\ell,h/2\}$, so the displayed decline of $K^*(p_0)$ follows from the maintained assumptions without imposing $\ell \geq h/2$. The illustrative parameters satisfy $0<\ell<h<1$ and $0 \leq \chi \leq 1$.

$[0,h]$: below ℓ, Bayes plausibility limits the probability that can be assigned to the wedge, while inside the wedge a degenerate posterior already attains the local score. Figure 1 displays this envelope and the accompanying near-boundary contrast: outcome-neutral belief dispersion vanishes as $p_0 \uparrow h$, although protocol-selection advantage remains at its comparator-relative maximum throughout the wedge.

Proposition 4 distinguishes two notions of pivotality. Under the default procedure H, beliefs become increasingly pivotal as p_0 approaches h: there is progressively less room for belief dispersion that does not cross the threshold. This is the belief-pivotality force. For the authorised procedure menu, however, posteriors throughout W are already procedurally pivotal: selecting L instead of H changes the action without requiring any belief movement. This is the procedure-disagreement force.

The comparator weight χ calibrates the contrast. When $\chi=0$, the comparator always applies demanding procedure H, so every departure to permissive procedure L is available as a steering move. When $\chi=1$, the comparator already applies L, so procedure selection cannot further raise the probability of action 1. Intermediate χ interpolate between these cases.

The degenerate posterior used in Proposition 4 is not essential:

Corollary 1 (Informative steering in the pivotal wedge)**.** *Suppose H is the default procedure and $p_0 \in (\ell, h)$. For every*

$$0 < \eta < \min\{p_0 - \ell, h - p_0\},$$

the two-posterior experiment

$$\tau_\eta = \tfrac{1}{2}\delta_{p_0-\eta} + \tfrac{1}{2}\delta_{p_0+\eta}$$

is informative and outcome-neutral under H, and its protocol-selection advantage towards action 1 is $1 - \chi$. Consequently, for every sequence $p_0^n \in (\ell, h)$ with $p_0^n \uparrow h$, there is a sequence of informative outcome-neutral experiments with protocol-selection advantage $1 - \chi$ for every n, even though their posterior variances converge to zero.

Proof. Both posteriors of τ_η lie in $W = [\ell, h]$, their mean is p_0, and $K(\tau_\eta) = \eta^2 > 0$. Thus the experiment is Bayes plausible, informative, and outcome-neutral under H, while equation (5) gives advantage $1 - \chi$. For the final claim, take

$$\eta_n = \tfrac{1}{2}\min\{p_0^n - \ell, h - p_0^n\} > 0.$$

Then $K(\tau_{\eta_n}) = \eta_n^2 \to 0$ because $\eta_n \leq (h - p_0^n)/2$. □

Corollary 1 closes the obvious loophole: steering is not an artefact of saying nothing. Under informative communication, the procedure remains fully decisive relative to the comparator while outcome-neutral belief dispersion vanishes.

For jury instructions, p may be interpreted as the jury's posterior assessment of a fact required for liability or conviction. Procedures H and L are not alternative substantive burdens of proof. They are reduced-form effective mappings induced by authorised formulations or comment practices that correctly state the governing law but may differ in how jurors translate a factual assessment into a verdict. The wedge contains assessments at which that permitted procedural choice affects the target verdict. Close to the default threshold, limiting movement in jurors' factual beliefs therefore need not limit the effect of selecting the formulation.

The other two settings of the introduction admit the same reading. For the auditor, p is an assessment of a firm's viability and H and L are authorised reporting procedures whose warning implications differ on the wedge. For the genetic counsellor, p is a patient's risk assessment and the procedures are consultation protocols with distinct action implications at common assessments. In each setting, regulating only belief movement leaves the procedural channel untouched.

6 Implications for Neutrality Rules

The formal results begin only after the institution has decided which procedures are lawful or professionally permissible. A court must exclude an instruction that misstates the elements, shifts a burden, or otherwise fails to state the governing law; a professional body must similarly define the reporting or counselling protocols within its mandate. Neutrality then asks a different question: how should permissible discretion be used when authorised procedures can yield different actions at the same posterior belief?

The fixed-decision benchmark rules out treating all impact as bias. In the model, communication that leaves the default action optimal at every posterior has no instrumental value for the represented decision. Applied to a jury, a no-impact rule would condemn clarification merely because correcting misunderstanding changes a verdict. The same mistake would treat an appropriately consequential audit report or clinical consultation as non-neutral merely because it matters.

The procedural model instead specifies a target-independent comparator for selecting among permissible instructions, reporting standards, or advisory protocols. Steering is the increase in a target action generated by departing from that comparator while holding posterior evidence fixed. Neutrality does not require an authorised procedure to be inconsequential. It requires that outcome-relevant procedural discretion not be selected to favour a target.

Judicial comment on the evidence illustrates the validity screen and the remaining discretion. In federal trials a judge may summarise and comment on the evidence, provided the comment does not distort or add to the record and the judge makes clear that factual determinations remain for the jury; many state systems more sharply restrict or prohibit judicial comment on factual matters or the weight of the evidence.[5] The menu $\mathcal{R}$ contains only practices allowed by the governing regime: a distortion of the evidence or an instruction that changes the legal burden is not an extreme procedure in the model but an unauthorised one. Within that menu, comments or formulations enter the model only as ways of translating a fixed factual assessment into a verdict. If jurors instead treat the procedure choice as revealing the judge's private view of the facts, the case belongs to the evidence-selection or signalling component, not to the protocol capacity.

The binary result sharpens the institutional implication. Near a decision boundary, an institution might be especially concerned with the presentation of evidence because small

[5] *Quercia v. United States*, 289 U.S. 466, 469–70 (1933) (describing the federal trial judge's authority to comment on evidence and its limits); *Bute v. Illinois*, 333 U.S. 640, 648 n.3 (1948) (noting the greater freedom permitted in federal courts than in many state courts); Wash. Const. art. IV, §16; Tex. Code Crim. Proc. art. 38.05.

belief movements can be decisive. Proposition 4 and Corollary 1 say that this concern is incomplete when authorised procedures disagree near the boundary: informative outcome-neutral communication can have vanishing dispersion while the protocol channel remains fully exposed. The institution must inspect the procedure menu and its comparator, rather than infer neutrality from limited informational movement.

Applications require discipline about the reduced form. In the jury setting, it represents verdict mappings of permissible practices, not a licence to alter the legal standard. In auditing, it isolates the mapping from evidence assessment into a reporting action, not feedback from a warning to a firm's survival. In genetic counselling, it captures protocols translating a fixed risk assessment into action, not selected information that changes risk beliefs or preferences. These limits identify the component to which the procedural channel applies.

The distinction also clarifies how information design is being used. In Bayesian persuasion, a sender chooses information to induce preferred actions. In the present problem, the institution uses the same Bayes-plausibility geometry to measure how exposed an authorised procedure menu is to directional selection. The exercise is not to recommend steering, but to locate where a neutrality standard must operate. If authorised procedures disagree at posteriors that a neutral information structure can reach, a rule governing evidence alone is not a complete rule against bias.

7 Conclusion

Bayesian persuasion asks how Bayes-plausible information can be selected to favour an action. Bayesian non-persuasion identifies a procedural channel through which an institution that permits communication may remain exposed to target-directed influence. The two problems share a geometry but give it opposite institutional roles.

Communication intended to improve a decision cannot generally be required to leave that decision unchanged. In a fixed Bayesian decision problem, outcome-neutral information has no instrumental decision value. Once an institution authorises procedures translating beliefs into actions, outcome neutrality and procedural neutrality diverge: under the stated compactness and upper-semicontinuity condition, the maximum directional use of procedural discretion is the concavification of a local steering score.

The divergence is sharp near an action boundary. In the two-threshold model, as a prior approaches the default threshold, the maximum variance of outcome-neutral beliefs converges to zero, while selection of an authorised alternative remains at its comparator-relative maximum; when the comparator applies the default procedure, it continues to

determine the action with probability one. This conclusion survives a requirement of genuinely informative communication: non-degenerate outcome-neutral experiments preserve maximal comparator-relative procedural exposure throughout the interior of the pivotal wedge.

The analysis deliberately takes the authorised menu and comparator as given. Target independence alone does not make a comparator normatively neutral; it only makes the benchmark independent of the favoured action. The exposure measured here is therefore conditional on an externally justified comparator. Choosing menus, comparator classes, and directional budgets is a separate design problem developed in follow-on work (Gans, 2026). Keeping that problem separate here isolates the point a neutrality rule must first confront: consequential communication is not biased merely because it changes decisions, but authorised procedural discretion can be biased when it is selected for its direction.

For judges and others such as auditors and counsellors, neutrality should therefore be framed as a restriction on target-directed selection of information and procedures relative to an authorised benchmark. The central question is not whether communication changes a decision. It is whether the institution has disciplined the channel through which that change occurs.

References

Charrow, Robert P. and Veda R. Charrow, "Making Legal Language Understandable: A Psycholinguistic Study of Jury Instructions," *Columbia Law Review*, 1979, *79* (7), 1306–1374.

de Lara, Michel and Olivier Gossner, "Payoffs-Beliefs Duality and the Value of Information," *SIAM Journal on Optimization*, 2020, *30* (1), 464–489.

Gans, Joshua S., "Neutrality by Design: Regulating Procedural Influence," 2026. Unpublished manuscript.

Glazer, Jacob and Ariel Rubinstein, "On Optimal Rules of Persuasion," *Econometrica*, 2004, *72* (6), 1715–1736.

Kamenica, Emir, "Bayesian Persuasion and Information Design," *Annual Review of Economics*, 2019, *11*, 249–272.

_ **and Matthew Gentzkow**, "Bayesian Persuasion," *American Economic Review*, 2011, *101* (6), 2590–2615.

Mathevet, Laurent, Jacopo Perego, and Ina Taneva, "On Information Design in Games," *Journal of Political Economy*, 2020, *128* (4), 1370–1404.

Taneva, Ina, "Information Design," *American Economic Journal: Microeconomics*, 2019, *11* (4), 151–185.

Tiersma, Peter Meijes, "Reforming the Language of Jury Instructions," *Hofstra Law Review*, 1993, *22* (1), 37–78.