

Bayesian Persuasion: Four Papers

Félix Muñoz-García*

1 Introduction

Consider one person, whom we call the Sender, who wants to change the action of another person, the Receiver. The Sender cannot pay the Receiver, threaten her, or in any way alter her payoffs. He can only choose what information she will see before she acts. Examples of this problem are everywhere:

- a prosecutor deciding which evidence to collect before a judge rules;
- a lobbyist funding a study that a politician will read;
- a firm designing the free trial that a consumer will try before buying;
- a school issuing only coarse grades so the labor market places its graduates better;
- a company choosing how much to disclose about its environmental record to a green-minded consumer.

The question we study, following Kamenica and Gentzkow (2011), is whether a Sender can benefit from this kind of persuasion even when the Receiver is a rational Bayesian who understands that the Sender designed the information with his own goal in mind. The answer, perhaps surprisingly, is yes. This handout works through the simplest example in the paper, the prosecutor and the judge, and shows every step needed to arrive at the same number the authors report: a rational judge who knows that 70 percent of defendants are innocent can nonetheless be led to convict 60 percent of them.

We proceed slowly, and each paper builds on the one before it. Section 2 solves the baseline model of Kamenica and Gentzkow (2011): with a free investigation the prosecutor convicts a rational judge $2\mu_0$ of the time, and the answer is the value of the concave closure of his payoff evaluated at the prior. Section 3 makes the investigation costly (Gentzkow and Kamenica 2014)

*Professor. School of Economic Sciences, Washington State University, Pullman, WA. This material was prepared for the EconS 594 course, Industrial Organization.

and recovers the baseline when the cost is zero. Section 4 lets the prosecutor and the judge hold different priors (Alonso and Cámara 2016) and recovers the baseline when they agree. Section 5 sets a defense attorney beside the prosecutor (Gentzkow and Kamenica 2017) and recovers the baseline when one side stays silent. Sections 6 and 7 tour the rest of the literature and point to open questions. One thread runs through all four papers: write the Sender’s payoff as a function of the Receiver’s belief, and the value of the concave closure of that payoff, evaluated at the prior, is the best the Sender can do. Each extension only reshapes that payoff, and each contains the baseline as the value of a single parameter. We keep the prior μ_0 generic throughout, so that every result is a formula in μ_0 , and we set $\mu_0 = 0.3$ only for the tables and figures.

2 Costless Persuasion (Kamenica and Gentzkow, 2011)

2.1 The model

States. There are two states of the world, $\omega \in \{G, I\}$, where G means the defendant is guilty and I means innocent. The prosecutor (Sender) and the judge (Receiver) share a common prior that the defendant is guilty. We call this prior μ_0 and keep it as a free parameter, so that every result below is a formula in μ_0 ; our running numerical illustration sets $\mu_0 = 0.3$. Because there are only two states, any belief is a single number, the probability it places on guilt, and we write $\mu \equiv \Pr(\text{guilty})$, so the prior places

$$\mu_0(G) = \mu_0 \quad \text{and} \quad \mu_0(I) = 1 - \mu_0.$$

We assume $\mu_0 < \frac{1}{2}$: a judge with no information would acquit, which is the case in which persuasion has any work to do. The parameter μ_0 is the one free parameter we carry through the whole handout; each later part adds one more parameter of its own, but the prior stays generic throughout.

Actions. The judge chooses one of two actions, $a \in \{\text{convict}, \text{acquit}\}$.

Payoffs. The judge wants to match the verdict to the truth. She earns 1 from the just action, convicting the guilty or acquitting the innocent, and 0 from the unjust action:

$$u(\text{convict}, G) = 1, \quad u(\text{acquit}, I) = 1, \quad u(\text{convict}, I) = 0, \quad u(\text{acquit}, G) = 0.$$

The prosecutor wants a conviction regardless of the truth. He earns 1 whenever the judge convicts and 0 whenever she acquits:

$$v(\text{convict}, \omega) = 1 \quad \text{and} \quad v(\text{acquit}, \omega) = 0 \quad \text{for both } \omega.$$

Notice that the prosecutor’s payoff depends only on the judge’s *action*, not on the state. This will

simplify the geometric picture we draw later.

The instrument. The prosecutor chooses an *investigation*: a signal π that generates a message for the judge, with the message drawn from state-dependent distributions $\pi(\cdot | G)$ and $\pi(\cdot | I)$. He commits to the investigation before the evidence is in, and reports its outcome honestly; he cannot rewrite the rules once he sees what turns up. The judge observes the design of the investigation and its realized message, updates her belief by Bayes' rule, and then convicts or acquits.

When does the judge convict? At any belief μ , the judge convicts if the expected payoff from convicting is at least that of acquitting:

$$\underbrace{\mu \cdot 1 + (1 - \mu) \cdot 0}_{\text{convict}} \geq \underbrace{\mu \cdot 0 + (1 - \mu) \cdot 1}_{\text{acquit}} \iff \mu \geq 1 - \mu \iff \mu \geq \frac{1}{2}.$$

We record the judge's optimal action as a function of her belief,

$$\bar{a}(\mu) = \begin{cases} \text{convict} & \text{if } \mu \geq \frac{1}{2}, \\ \text{acquit} & \text{if } \mu < \frac{1}{2}, \end{cases}$$

where we assign a tie in the prosecutor's favor, consistent with the Sender-preferred equilibrium of the paper; without this convention the best conviction rate would be a value the prosecutor could approach but never quite reach.

2.2 Two benchmarks

Before designing anything clever, we compute what happens at the two extremes. These give the prosecutor a floor to improve upon.

Example 1 (No disclosure). Suppose the investigation is completely uninformative, so the judge learns nothing and keeps her prior μ_0 . Since $\mu_0 < \frac{1}{2}$, the judge acquits, and the prosecutor's payoff is

$$v_{\text{no info}} = 0.$$

We call *acquit* the *default action*, since it is what the judge does when she has no information beyond the prior. $\square$

Example 2 (Full disclosure). Suppose the investigation reveals the state with certainty. Then the judge's posterior is either $\mu = 1$ (and she convicts) or $\mu = 0$ (and she acquits). She convicts exactly when the defendant is truly guilty, which happens with probability μ_0 . The prosecutor's expected

payoff is

$$v_{\text{full info}} = \mu_0 \cdot 1 + (1 - \mu_0) \cdot 0 = \mu_0 \quad (= 0.3 \text{ at } \mu_0 = 0.3).$$

Full disclosure beats silence, raising the conviction rate from 0 to μ_0 . The question is whether the prosecutor can do better still. $\square$

Intuitively, revealing everything is a crude instrument. It hands the judge a conviction only on the guilty, and it throws away the possibility of convicting on the strength of merely suggestive evidence. We now show that a partial investigation does better than both benchmarks.

2.3 The one idea that makes the problem tractable

The prosecutor has infinitely many investigations to choose among, differing in how many messages they send and how each message correlates with the state. Searching over all of them directly looks hopeless. The insight of Kamenica and Gentzkow is that we do not need to. What the prosecutor ultimately cares about is the judge's *belief* when she acts, because the belief is what determines the verdict. So we can search over the beliefs an investigation can produce, rather than over the investigations themselves.

Each message the judge might receive leaves her with some posterior belief μ . An investigation therefore induces a *distribution of posteriors*: a list of the beliefs that can arise, together with the probability of each. The question becomes: which distributions of posteriors are attainable?

The answer is that a distribution of posteriors is attainable if and only if it is **Bayes plausible**, meaning that the beliefs average back to the prior:

$$\mathbb{E}[\mu] = \mu_0.$$

The requirement is nothing more than the law of iterated expectations. Whatever the judge might come to believe after seeing a message, her beliefs, weighted by how likely each message is, must return to where she started. No investigation can, on average, make the judge more convinced of guilt than the prior warrants; it can only redistribute her conviction across messages.

This is the whole content of Proposition 1 in the paper. It lets us restate the prosecutor's problem in a clean form:

$$\max_{\tau \text{ Bayes plausible}} \mathbb{E}_{\tau}[\hat{v}(\mu)],$$

where τ is the distribution of posteriors and $\hat{v}(\mu)$ denotes the prosecutor's payoff when the judge holds belief μ and acts optimally. In our example, since the prosecutor earns 1 whenever the judge

convicts,

$$\hat{v}(\mu) = v(\bar{a}(\mu)) = \begin{cases} 1 & \text{if } \mu \geq \frac{1}{2}, \\ 0 & \text{if } \mu < \frac{1}{2}. \end{cases}$$

2.4 Solving the prosecutor's problem

We now solve the problem step by step. It is convenient to work with *straightforward* investigations: those whose messages are recommendations, “convict” or “acquit,” which the judge is willing to obey. The paper shows this is without loss of generality, exactly as the revelation principle lets us restrict attention to truthful mechanisms. There are only two messages to design, so the problem reduces to two numbers. Two messages are all the prosecutor ever needs: because the judge takes only one of two actions, a signal with more realizations cannot raise his payoff.

Step 1. Name the choice variables. Let

$$\sigma_G = \pi(g \mid G) \quad \text{and} \quad \sigma_I = \pi(g \mid I)$$

be the probabilities that the investigation sends the message g (“recommend convict”) in each state. The complementary probabilities, $1 - \sigma_G$ and $1 - \sigma_I$, are the chances of the message i (“recommend acquit”).

Step 2. Write the objective. If the judge obeys, the prosecutor wins a conviction exactly when the message is g . That message arrives with total probability

$$\Pr(g) = \mu_0 \sigma_G + (1 - \mu_0) \sigma_I,$$

so the prosecutor maximizes $\mu_0 \sigma_G + (1 - \mu_0) \sigma_I$.

Step 3. Impose obedience on the conviction recommendation. For the judge to actually convict when she sees g , her posterior on guilt given g must be at least $\frac{1}{2}$. By Bayes' rule,

$$\mu_g \equiv \Pr(G \mid g) = \frac{\mu_0 \sigma_G}{\mu_0 \sigma_G + (1 - \mu_0) \sigma_I} \geq \frac{1}{2}.$$

Cross-multiplying, this obedience condition is

$$\mu_0 \sigma_G \geq (1 - \mu_0) \sigma_I.$$

Step 4. Impose obedience on the acquittal recommendation. For the judge to acquit when

she sees i , her posterior on guilt given i must be at most $\frac{1}{2}$:

$$\mu_i \equiv \Pr(G | i) = \frac{\mu_0 (1 - \sigma_G)}{\mu_0 (1 - \sigma_G) + (1 - \mu_0) (1 - \sigma_I)} \leq \frac{1}{2},$$

which simplifies to $\mu_0 (1 - \sigma_G) \leq (1 - \mu_0) (1 - \sigma_I)$.

Step 5. Argue that a guilty defendant is always recommended for conviction. The objective $\mu_0 \sigma_G + (1 - \mu_0) \sigma_I$ rises in σ_G , and raising σ_G also loosens the conviction-obedience condition in Step 3, since it puts more guilt mass behind the message g . There is thus no reason to hold back: the prosecutor sets

$$\sigma_G = 1.$$

Intuitively, evidence pointing at a guilty defendant can only help the case, so the optimal investigation never recommends acquittal for someone who is in fact guilty. This is the example's version of the general rule that when the judge takes the prosecutor's least-preferred action, she does so only when she is certain of the state.

Step 6. Convict as many innocents as obedience allows. With $\sigma_G = 1$, the objective becomes $\mu_0 + (1 - \mu_0) \sigma_I$, which rises in σ_I . The prosecutor would like σ_I as large as possible, but Step 3 caps it at $\mu_0 \geq (1 - \mu_0) \sigma_I$. The constraint binds, so

$$\sigma_I = \frac{\mu_0}{1 - \mu_0}.$$

We check Step 4 is satisfied: $\mu_0(1 - 1) = 0 \leq (1 - \mu_0)(1 - \frac{\mu_0}{1 - \mu_0}) = 1 - 2\mu_0$, which is positive because $\mu_0 < \frac{1}{2}$, so the judge is indeed willing to acquit on the message i .

Step 7. Read off the optimal investigation. Translating (σ_G, σ_I) back into the language of the two messages,

$$\begin{aligned} \pi(g | G) &= 1, & \pi(g | I) &= \frac{\mu_0}{1 - \mu_0}, \\ \pi(i | G) &= 0, & \pi(i | I) &= \frac{1 - 2\mu_0}{1 - \mu_0}. \end{aligned}$$

At the running prior $\mu_0 = 0.3$ this is $(1, \frac{3}{7}, 0, \frac{4}{7})$, precisely the investigation reported in equation (1) of Kamenica and Gentzkow (2011).

Step 8. Compute the conviction rate. The prosecutor's payoff equals the probability of the message g :

$$v^* = \Pr(g) = \mu_0 (1) + (1 - \mu_0) \frac{\mu_0}{1 - \mu_0} = \mu_0 + \mu_0 = 2\mu_0.$$

The optimal investigation therefore convicts with probability $2\mu_0$, exactly twice the full-disclosure

rate μ_0 , whenever $\mu_0 < \frac{1}{2}$. At $\mu_0 = 0.3$ this is 0.6: a judge who knows that 70 percent of defendants are innocent convicts 60 percent of them, while fully aware that the investigation was built to maximize convictions.

Why the judge goes along

The prosecutor and the judge are not pure adversaries: both want to convict the guilty, so some messages serve them at once. And because the prosecutor commits to the investigation in advance and the judge knows its design, she can trust what a message implies rather than second-guess his motive. With that in place, it is worth reading the two posteriors the investigation produces, because they show there is no trick. On the message i , the posterior is

$$\mu_i = \frac{\mu_0 \cdot 0}{\mu_0 \cdot 0 + (1 - \mu_0) \cdot \frac{1-2\mu_0}{1-\mu_0}} = 0,$$

so the judge is *certain* the defendant is innocent and acquits without hesitation. On the message g , the posterior is

$$\mu_g = \frac{\mu_0 \cdot 1}{\mu_0 \cdot 1 + (1 - \mu_0) \cdot \frac{\mu_0}{1-\mu_0}} = \frac{\mu_0}{\mu_0 + \mu_0} = \frac{1}{2},$$

so the judge is exactly indifferent between convicting and acquitting, and, breaking the tie in the prosecutor's favor, she convicts. The prosecutor has engineered the weakest belief consistent with a conviction: just enough to tip the judge over, and not a shade more.

We can also confirm Bayes plausibility directly. The message g arrives with probability $2\mu_0$ and leaves belief $\frac{1}{2}$; the message i arrives with probability $1 - 2\mu_0$ and leaves belief 0. Averaging,

$$2\mu_0 \cdot \frac{1}{2} + (1 - 2\mu_0) \cdot 0 = \mu_0,$$

so the two posteriors do return to the prior, as they must.

2.5 Graphical representation

The problem has a picture that gives the answer at once, and it generalizes far beyond this example. Plot the prosecutor's payoff $\hat{v}(\mu)$ against the judge's belief μ . As computed above, $\hat{v}$ is a step function: it is 0 for every belief below $\frac{1}{2}$, where the judge acquits, and jumps to 1 once the belief reaches $\frac{1}{2}$, where she convicts.

Now recall the prosecutor's problem: choose beliefs that average to the prior μ_0 , so as to maximize the average height of $\hat{v}$. Mixing two beliefs traces the straight line joining the two points

on the graph of $\hat{v}$. The best the prosecutor can do at any prior is therefore given by the **concave closure** of $\hat{v}$, written V : the lowest concave curve that lies weakly above $\hat{v}$ everywhere. Figure 1 plots $\hat{v}$, its concave closure V , and the construction at the prior.

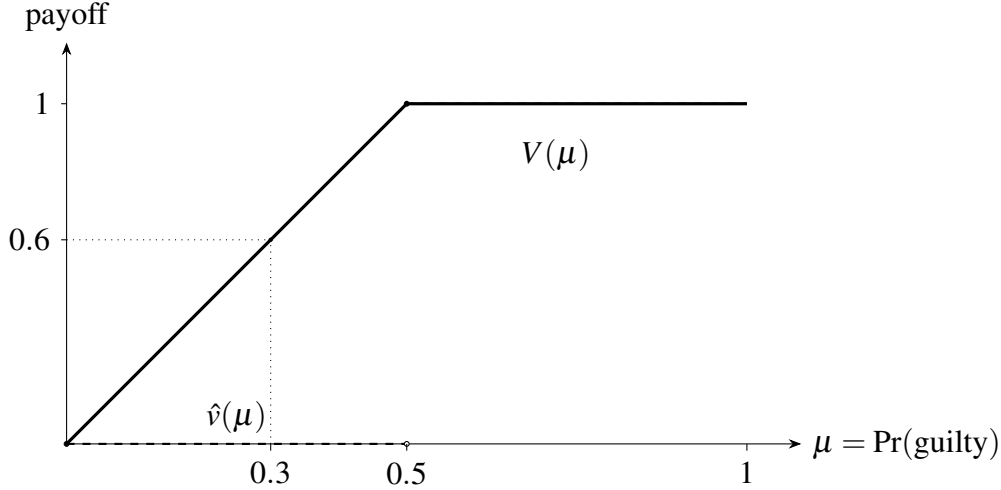


Figure 1: The step payoff $\hat{v}$ (dashed), its concave closure V (solid), and the value $V(0.3) = 0.6$ at the prior.

Reading the figure recovers everything from the algebra above, and faster. The concave closure is the line $V(\mu) = 2\mu$ for $\mu \leq \frac{1}{2}$ and $V(\mu) = 1$ for $\mu > \frac{1}{2}$. Its value at the prior is

$$V(\mu_0) = 2\mu_0 \quad (= 0.6 \text{ at } \mu_0 = 0.3),$$

which is the optimal conviction rate. The two points the closure touches, $\mu = 0$ and $\mu = \frac{1}{2}$, are exactly the two posteriors the optimal investigation induces. And the weights that place their average at μ_0 are $1 - 2\mu_0$ on $\mu = 0$ and $2\mu_0$ on $\mu = \frac{1}{2}$, the probabilities of the two messages. The gap between the curves at the prior,

$$V(\mu_0) - \hat{v}(\mu_0) = 2\mu_0 - 0 = 2\mu_0,$$

is the prosecutor's gain from persuasion. Wherever this gap is positive, the Sender benefits; wherever $\hat{v}$ already equals its own concave closure, no investigation helps.

2.6 Main points from the above example

The example teaches lessons that recur throughout what follows.

Persuasion works when the payoff is not concave. The prosecutor gains precisely because $\hat{v}$ lies below its concave closure at the prior. If the prosecutor's payoff were concave in the judge's belief,

no spreading of beliefs could raise its average, and silence would be optimal. If it were convex, full disclosure would be optimal. The interesting case, and the realistic one, is the step in between.

The Sender loses nothing on the downside. Providing information sometimes moves the judge toward guilt and sometimes toward innocence. The move toward guilt earns a conviction; the move toward innocence costs the prosecutor nothing, because the judge was already acquitting by default. Persuasion benefits the Sender whenever the Receiver does not already take the Sender's preferred action, and her action stays fixed over a range of beliefs around the prior.

The Sender induces the worst belief consistent with each action. When the judge acquits, she is made certain of innocence; when she convicts, she is made just indifferent. The prosecutor never leaves the judge more convinced than the action requires, because any leftover conviction could be spent buying additional convictions elsewhere. This is the content of Propositions 4 and 5 in the paper, and the reader can see both at work in the two posteriors computed above.

2.7 An environmental application: green signaling and greenwashing

The courtroom is one dressing of a structure that recurs, and a second reading will run alongside us through the rest of the handout. Read the Sender as a firm and the Receiver as a green-minded consumer, or a regulator. The state is whether the firm's product is genuinely clean or dirty; the prior μ_0 is now the share of clean products, and the consumer buys, or the regulator approves, only when she is convinced enough that the product is clean, say when her belief that it is clean is at least $\frac{1}{2}$. The firm, like the prosecutor, wants the favorable action whatever the truth, and it chooses what to disclose about its environmental performance: which tests to run, which metrics to report, how to design a sustainability label.

The baseline result then reads as a theory of greenwashing. A firm sure to be exposed as dirty says nothing; a firm with a genuinely clean case reveals it; and in between the firm designs a label that raises a skeptical consumer's belief to exactly the buying threshold and not a shade higher. The optimal green signal is therefore a partial, noisy label that clears the consumer at rate $2\mu_0$, twice the share of products that are actually clean, precisely the "convict 60 percent while only 30 percent are guilty" logic of the courtroom. The three extensions carry over intact: costly testing (Section 3) shrinks how much the firm discloses; a firm more optimistic about its own product than its audience is (Section 4) values the same label more without changing it; and a watchdog or a rival that can also disclose (Section 5) pushes the market toward full revelation of the firm's true type. This environmental reading is the subject of a large literature on certification and greenwashing (Lyon and Maxwell 2011).

2.8 A numerical summary

Table 1 collects the three investigations, no disclosure, full disclosure, and the optimal partial investigation, at the running prior $\mu_0 = 0.3$. Reading down the conviction column shows the prosecutor's problem in one glance: silence convicts no one, full disclosure convicts only the truly guilty, and the optimal partial investigation doubles the full-disclosure rate. The judge's own welfare, in the last column, is highest under full disclosure, where she always matches the verdict to the truth, and falls back to her uninformed level of 0.7 under the optimal investigation, since the prosecutor leaves her exactly indifferent whenever she convicts.

Investigation	Conviction rate	Prosecutor payoff	Judge welfare
No disclosure	0.00	0.00	0.70
Full disclosure	0.30	0.30	1.00
Optimal (partial)	0.60	0.60	0.70

Table 1: The three investigations at the prior $\mu_0 = 0.3$.

How much the prosecutor gains depends on the prior. Since the value of persuasion is $V(\mu_0) = \min(2\mu_0, 1)$ while full disclosure convicts at rate μ_0 , the gain is μ_0 for every prior below $\frac{1}{2}$ and vanishes once $\mu_0 \geq \frac{1}{2}$, where the judge convicts by default. Figure 2 plots the two rates.

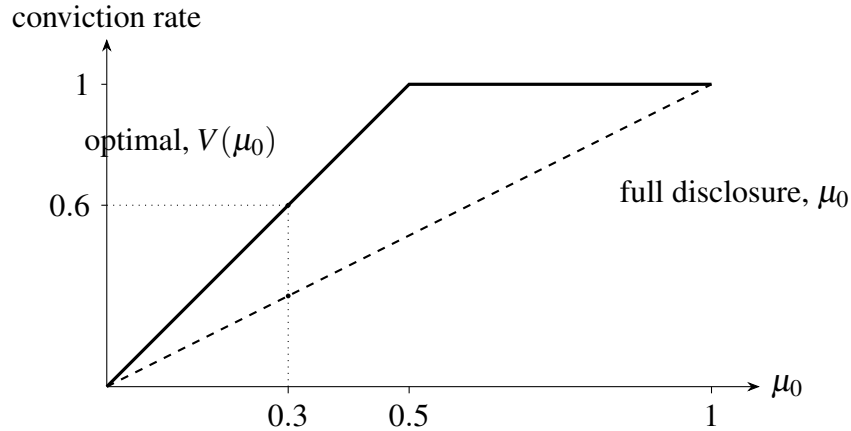


Figure 2: Conviction rate under the optimal investigation, $V(\mu_0) = \min(2\mu_0, 1)$, and under full disclosure, μ_0 .

3 Costly Persuasion (Gentzkow and Kamenica, 2014)

3.1 From free information to costly information

The investigation in Section 2 was free. A prosecutor, though, does not gather evidence at no charge: forensic tests, expert witnesses, depositions, and subpoenas all consume time and money. It is natural to ask whether the elegant concavification of Section 2 survives once the Sender must pay for the information he generates.

At first sight the answer looks like no. The whole method rested on writing the Sender's payoff as a function of the Receiver's posterior alone, $\hat{v}(\mu)$, so that the problem reduced to averaging $\hat{v}$ over Bayes-plausible beliefs. Once signals are costly, this breaks: two investigations can leave the judge at the very same posterior μ while costing the prosecutor different amounts, one being a cheap, coarse investigation and the other an expensive, finely tuned one. The Sender's payoff at a belief then depends on *how* the belief was reached, not on the belief alone, and the concavification picture no longer applies.

Gentzkow and Kamenica (2014), henceforth GK, identify a family of cost functions for which the difficulty disappears. For this family the cost can be absorbed into the value function, producing a *cost-adjusted* payoff that again depends on the posterior alone, so the same picture goes through. And when the cost vanishes, the model returns to Kamenica and Gentzkow (2011), henceforth KG. The costless model of Section 2 is thus the special case of a one-parameter family, recovered when that parameter equals zero.

3.2 The model with costs

The states, actions, payoffs, and prior are exactly those of Section 2: two states $\omega \in \{G, I\}$ with the same generic prior $\mu_0 = \Pr(\text{guilty}) < \frac{1}{2}$ (running value 0.3); the judge convicts if and only if her belief satisfies $\mu \geq \frac{1}{2}$; and the prosecutor earns 1 from a conviction and 0 from an acquittal. The only change is that running an investigation π now costs the prosecutor $c(\pi)$, so his payoff becomes

$$v(a, \omega) - c(\pi).$$

GK's central assumption is that the cost of an investigation is proportional to the *expected reduction in uncertainty* it produces. Let $H(\mu)$ be a measure of the uncertainty in a belief μ : a concave function that is largest when the belief is most spread out and zero when the belief is degenerate. Then the assumption is

$$c(\pi) = k \mathbb{E}[H(\mu_{\text{ref}}) - H(\mu_s)],$$

where μ_{ref} is a fixed reference belief, μ_s the posterior after message s , and $k \geq 0$ a cost parameter. Concavity of H guarantees, by Jensen's inequality, that information reduces uncertainty on average, so the cost is never negative, and it is zero for an uninformative investigation.

Two measures of uncertainty are standard. The first is Shannon entropy, $H(\mu) = -\sum_{\omega} \mu(\omega) \ln \mu(\omega)$, which underlies the rational-inattention literature. The second is residual variance, $H(\mu) = \sum_{\omega} \mu(\omega) (1 - \mu(\omega))$, which in our two-state model is

$$H(\mu) = 2\mu(1 - \mu).$$

We adopt residual variance, and we place the reference belief at the prior, $\mu_{\text{ref}} = \mu_0$, so that the cost measures the uncertainty an investigation removes relative to what the judge knew at the outset. Both are choices GK explicitly allow. This pair keeps the algebra in closed form and lets us see every step; GK's own headline example instead uses entropy with a uniform reference belief, and we reproduce its numbers in the numerical illustration below. With the reference at the prior the cost simplifies to

$$c(\pi) = k(H(\mu_0) - \mathbb{E}_{\tau}[H(\mu)]),$$

where τ is, as before, the distribution of posteriors the investigation induces.

3.3 The unifying result: concavify a cost-adjusted payoff

The Sender's problem is again to choose a Bayes-plausible distribution of posteriors, but now net of cost:

$$\begin{aligned} \max_{\tau \text{ Bayes plausible}} \mathbb{E}_{\tau}[\hat{v}(\mu)] - c(\pi) &= \max_{\tau} \mathbb{E}_{\tau}[\hat{v}(\mu)] - k(H(\mu_0) - \mathbb{E}_{\tau}[H(\mu)]) \\ &= \max_{\tau} \mathbb{E}_{\tau}[\hat{v}(\mu) + kH(\mu)] - kH(\mu_0). \end{aligned}$$

The last line is the key. The cost has been absorbed into a single *cost-adjusted value function*

$$\hat{v}_k(\mu) \equiv \hat{v}(\mu) + kH(\mu),$$

which depends on the posterior alone, so the concavification method of Section 2 applies without change. Writing V_k for the concave closure of $\hat{v}_k$, the Sender's value is

$$V_k(\mu_0) - kH(\mu_0),$$

and the Sender benefits from persuasion if and only if $V_k(\mu_0) > \hat{v}_k(\mu_0)$. This is GK's Proposition 1, specialized to our cost.

The nesting. Set $k = 0$. Then $\hat{v}_k = \hat{v}$, the concave closure V_k coincides with V from Section 2, and the constant $kH(\mu_0)$ vanishes. The Sender's value is $V(\mu_0) = 2\mu_0$, exactly the costless result of Section 2. Kamenica and Gentzkow (2011) is the $k = 0$ member of the family.

Intuitively, adding the concave term $kH(\mu)$ to $\hat{v}$ makes the payoff more concave. Since persuasion feeds on the gap between a function and its concave closure, and a more concave function leaves a smaller gap, information costs shrink the room for persuasion. The subtracted constant $kH(\mu_0)$ is what the prosecutor forgoes simply by moving the judge's belief away from the prior at all.

3.4 The toy model, worked

Step 1. Write the cost-adjusted payoff. With $\hat{v}(\mu)$ the step function of Section 2 and $H(\mu) = 2\mu(1 - \mu)$,

$$\hat{v}_k(\mu) = \begin{cases} 2k\mu(1 - \mu) & \text{if } \mu < \frac{1}{2}, \\ 1 + 2k\mu(1 - \mu) & \text{if } \mu \geq \frac{1}{2}. \end{cases}$$

Each piece is a downward parabola, and the two pieces are separated by an upward jump of size 1 at the judge's indifference belief $\mu = \frac{1}{2}$.

Step 2. The conviction belief stays at the threshold. To win a conviction the prosecutor must leave the judge with a posterior $\mu_h \geq \frac{1}{2}$. Raising μ_h above $\frac{1}{2}$ buys no extra conviction, since the judge already convicts, and it is costly, since H is falling above $\frac{1}{2}$ and a sharper belief removes more uncertainty. The prosecutor therefore leaves the judge exactly indifferent, $\mu_h = \frac{1}{2}$, just as in Section 2.

Step 3. The acquittal belief now involves a trade-off. Let $\mu_l < \frac{1}{2}$ be the belief that leads the judge to acquit. In Section 2 the prosecutor drove μ_l all the way to 0, making the judge certain of innocence. That is no longer clearly best. Lowering μ_l toward 0 still raises the conviction probability, because Bayes plausibility then allows more weight on the conviction message; but it also raises the cost, because a belief nearer certainty carries less uncertainty and the cost rewards uncertainty left in place. The prosecutor must weigh the two.

Step 4. Solve for the acquittal belief. With $\mu_h = \frac{1}{2}$ and weight p on the conviction message, Bayes plausibility reads $p \cdot \frac{1}{2} + (1 - p)\mu_l = \mu_0$, so

$$p = \frac{\mu_0 - \mu_l}{\frac{1}{2} - \mu_l}.$$

The prosecutor's net payoff, gross conviction probability p minus cost, is

$$N(\mu_l) = p - k \left(H(\mu_0) - [p \cdot H(\frac{1}{2}) + (1-p)H(\mu_l)] \right),$$

with $H(\frac{1}{2}) = \frac{1}{2}$ and $H(\mu_0) = 2\mu_0(1 - \mu_0)$. Differentiating and evaluating at $\mu_l = 0$ gives the clean expression

$$\left. \frac{dN}{d\mu_l} \right|_{\mu_l=0} = 2(1 - 2\mu_0) \left(\frac{k}{2} - 1 \right),$$

which at $\mu_0 = 0.3$ is $-0.8 + 0.4k$.

Step 5. Read off the three regimes. Because $1 - 2\mu_0 > 0$, the sign of the derivative is the sign of $\frac{k}{2} - 1$, so the threshold sits at $k = 2$ for every prior $\mu_0 < \frac{1}{2}$. This sorts the problem into three cases.

- If $k \leq 2$, the derivative is not positive, so the prosecutor keeps $\mu_l = 0$. The optimal investigation is *identical* to the costless one of Section 2: the judge is made certain of innocence on the acquittal message and indifferent on the conviction message, and the conviction rate stays at $2\mu_0$. Only the prosecutor's net payoff falls, to

$$N = 2\mu_0 - k\mu_0(1 - 2\mu_0) = \mu_0[2 - k(1 - 2\mu_0)],$$

which at $\mu_0 = 0.3$ is $0.6 - 0.12k$.

- If $2 < k < k^*(\mu_0)$, the derivative at $\mu_l = 0$ is positive, so the prosecutor raises μ_l above 0. He no longer fully clears the innocent, the acquittal message stops implying certain innocence, and the conviction rate falls below $2\mu_0$.
- If $k \geq k^*(\mu_0)$, no investigation is worthwhile, and the prosecutor gathers no evidence.

The knife-edge at $k = 2$ is where a binary verdict first makes it worth economizing on information; the upper cutoff $k^*(\mu_0)$, where persuasion stops altogether, depends on the prior and equals about 12.5 at $\mu_0 = 0.3$ (see the numerical illustration below).

Step 6. Recover the investigation at an interior cost. Return to the running prior $\mu_0 = 0.3$ and take $k = 3$, in the middle regime. Solving $N'(\mu_l) = 0$ yields $\mu_l = 0.0918$, hence $p = 0.510$. The two induced posteriors are $\mu_l = 0.0918$ with weight $1 - p = 0.490$ and $\mu_h = \frac{1}{2}$ with weight $p = 0.510$. Recovering the investigation from $\pi(s | \omega) = \mu_s(\omega) \tau(\mu_s) / \mu_0(\omega)$,

$$\begin{aligned} \pi(g | G) &= \frac{0.5 \times 0.510}{0.3} = 0.850, & \pi(g | I) &= \frac{0.5 \times 0.510}{0.7} = 0.364, \\ \pi(i | G) &= \frac{0.0918 \times 0.490}{0.3} = 0.150, & \pi(i | I) &= \frac{0.9082 \times 0.490}{0.7} = 0.636. \end{aligned}$$

The conviction rate is $\Pr(g) = 0.3(0.850) + 0.7(0.364) = 0.510$, down from 0.6. Notice that $\pi(i | G) = 0.150 > 0$: the prosecutor now recommends acquittal for some guilty defendants. This is the sharpest contrast with Section 2. When information was free, the acquittal message meant certain innocence (Proposition 4 of KG); once information is costly, buying that certainty is too expensive, so the prosecutor accepts an acquittal message that leaves the judge with lingering doubt.

3.5 What cost adds to the picture

Costly persuasion reshapes the baseline in ways worth drawing out.

The costless model is a boundary, not a different model. Everything in Section 3 collapses to Section 2 at $k = 0$. The extension does not overturn Bayesian persuasion; it embeds it in a family and asks what happens as one moves away from the free-information boundary. This is the shape of a good extension: the earlier result sits inside the newer one at a single parameter value, and the new results describe the neighborhood around it.

Costs shrink disclosure, and do so in stages. As k rises, the optimal investigation is first unchanged, then Blackwell less informative, then silent. The prosecutor first absorbs the cost while running the same investigation, then economizes by tolerating a noisier acquittal signal, and finally abandons persuasion. Lower costs always yield a weakly more informative investigation in this example, matching GK's comparative static.

The Receiver captures none of the value of information. A short calculation shows the judge's expected payoff under the optimal investigation is

$$p \cdot \frac{1}{2} + (1 - p)(1 - \mu_l) = 1 - \mu_0 \quad \text{for every } k,$$

using $p \cdot \frac{1}{2} + (1 - p)\mu_l = \mu_0$. This is exactly the judge's payoff with no information at all, $\Pr(\text{innocent}) = 1 - \mu_0$ (namely 0.7 at $\mu_0 = 0.3$). Whatever the cost of investigation, the prosecutor leaves the judge no better off than her uninformed benchmark and keeps the entire value of information for himself.

3.6 A numerical summary

Table 2 traces the optimal investigation as the cost parameter k rises, under the residual-variance cost worked out above. The three regimes are visible in the columns. For $k \leq 2$ the acquittal belief stays at $\mu_l = 0$ and the conviction rate holds at 0.6, exactly the costless investigation; only the prosecutor's net payoff falls, one-for-one at rate 0.12 per unit of cost. For $2 < k < k^*$ the acquittal belief lifts off 0 and the conviction rate slides down. At $k^* \approx 12.5$ the prosecutor stops investigating. The judge's welfare, not shown, equals 0.7 in every row.

k	μ_l	Conviction rate	Prosecutor net	Regime
0	0.000	0.600	0.600	identical to KG
1	0.000	0.600	0.480	signal = KG
2	0.000	0.600	0.360	threshold
3	0.092	0.510	0.260	interior
4	0.146	0.434	0.189	interior
6	0.211	0.307	0.094	interior
10	0.276	0.106	0.011	interior
12.5	0.300	0.000	0.000	no disclosure

Table 2: Optimal investigation as the cost k rises, residual-variance cost with reference belief $\mu_0 = 0.3$.

Figure 3 draws the case $k = 3$ the way Figure 1 drew the costless problem. The cost bends each branch of the payoff into a parabola and lifts the whole picture, but the method is unchanged: the concave closure V_3 of the cost-adjusted payoff $\hat{v}_3$ still touches the graph at the two induced posteriors, now $\mu_l = 0.092$ and $\mu_h = \frac{1}{2}$. Reading the closure at the prior gives $V_3(0.3) = 1.52$, and the prosecutor's payoff is $V_3(0.3) - kH(\mu_0) = 1.52 - 1.26 = 0.26$, matching Table 2.

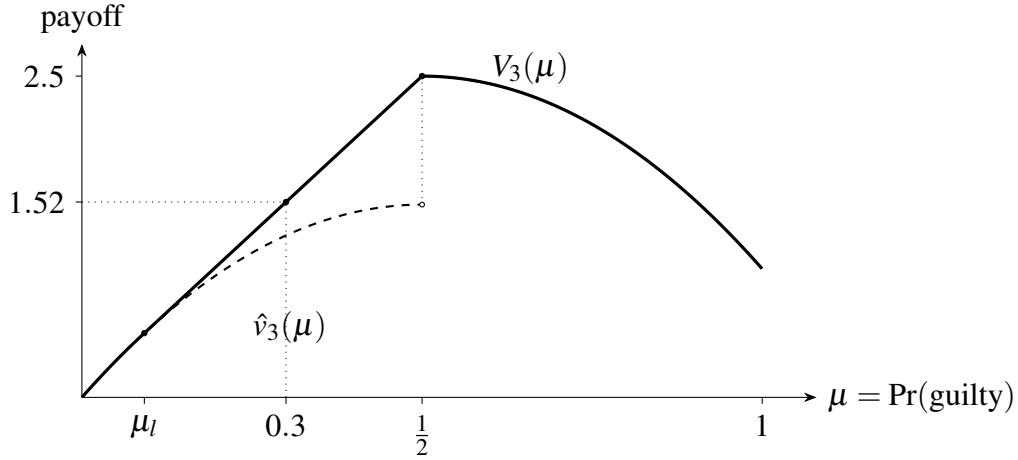


Figure 3: The cost-adjusted payoff $\hat{v}_3$ (dashed) and its concave closure V_3 (solid) at $k = 3$, with induced posteriors $\mu_l = 0.092$ and $\mu_h = \frac{1}{2}$.

Finally, Table 3 reports GK's own version of the example, which uses Shannon entropy as the measure of uncertainty and the uniform belief as the reference, rather than the residual variance and prior reference used above. The pattern is the same, and $k = 0$ again returns the costless investigation of Section 2.

¹The published example reports a conviction rate of 0.43 at $k = 2$. The signal listed there induces a conviction belief $\mu_g = 0.4988$, just short of the judge's indifference point $\frac{1}{2}$, so it narrowly fails her obedience constraint; the obedience-respecting optimum, shown here, convicts at rate 0.457. The qualitative comparative statics are identical.

k	Conviction rate	μ_i	Prosecutor net
0	0.600	0.000	0.600
1	0.543	0.062	0.384
2	0.457	0.131	0.259
3	0.382	0.176	0.177
4	0.315	0.208	0.119
5	0.255	0.232	0.077
7	0.147	0.265	0.025
10	0.009	0.298	0.000

Table 3: GK’s entropy version of the example, reference belief uniform.¹

4 Persuasion under Disagreement (Alonso and Cámara, 2016)

4.1 From a shared prior to disagreement

Sections 2 and 3 both assume that the prosecutor and the judge start from the same prior, μ_0 . Often they do not. A prosecutor who has lived with a case for months may be more convinced of guilt than the judge who sees it for the first time; an advocate is typically more optimistic than the audience he addresses. Alonso and Cámara (2016) allow the Sender and the Receiver to hold different priors, both common knowledge, so that the two agree to disagree.

We keep the judge’s prior at the same generic value μ_0 (running value 0.3), so her behavior is exactly as in Section 2, and we let the prosecutor’s prior be a number $q \in (0, 1)$. When $q = \mu_0$ the two agree and we are back in Section 2. The question is what disagreement changes when $q \neq \mu_0$.

4.2 The model with two priors

The states, actions, and payoffs are those of Section 2. The judge holds the prior μ_0 and convicts if and only if her posterior is at least $\frac{1}{2}$, unchanged. The prosecutor holds a possibly different prior q . He designs a signal; the judge updates from her prior μ_0 ; but the prosecutor evaluates outcomes using his prior q , that is, using his own assessment of how likely each message, and each resulting conviction, is. Both priors are common knowledge.

4.3 The unifying result: reweight the value function

The prosecutor and the judge read the same evidence but weigh the possible verdicts differently, because they disagree about how likely guilt is.

First, the judge’s behavior still depends only on her posterior μ : she convicts if and only if

$\mu \geq \frac{1}{2}$, and the prosecutor's gross payoff at that posterior is the same step function $\hat{v}(\mu)$ as in Section 2.

Second, the prosecutor weights each message not by the judge's probability but by his own. For a message that leaves the judge at posterior μ , the ratio of the prosecutor's probability of that message to the judge's probability of it is

$$\ell(\mu) = \mu \frac{q}{\mu_0} + (1 - \mu) \frac{1 - q}{1 - \mu_0},$$

which depends only on the belief the message induces, not on the signal that produced it. The reason is that, once the judge's posterior μ is known, the message's likelihood in each state relative to its average likelihood is fixed, and the two priors simply reweight those likelihoods.

Putting the two facts together, the prosecutor's expected payoff is $\mathbb{E}_\tau[\ell(\mu) \hat{v}(\mu)]$, where τ is the judge's Bayes-plausible distribution of posteriors. Define the reweighted value $w(\mu) \equiv \ell(\mu) \hat{v}(\mu)$. The prosecutor's problem is

$$\max_{\tau \text{ Bayes plausible}} \mathbb{E}_\tau[w(\mu)] \quad \text{subject to } \mathbb{E}_\tau[\mu] = \mu_0,$$

solved once again by concavifying w and reading it at the judge's prior μ_0 .

The nesting. When $q = \mu_0$, the reweighting is $\ell(\mu) = \mu + (1 - \mu) = 1$, so $w = \hat{v}$ and we recover Section 2 exactly. Heterogeneous priors is the family; the common-prior model of Kamenica and Gentzkow (2011) is its $q = \mu_0$ member.

This reweighting is close kin to the device behind costly persuasion. When the cost in Section 3 is measured from a reference belief other than the prior, Gentzkow and Kamenica (2014) translate between that reference belief and the prior in exactly the way we here translate between the prosecutor's and the judge's priors, and they credit the device to Alonso and Cámara's work on heterogeneous priors. A single belief-map underlies both extensions, which is why the two share a shape.

4.4 The toy model, worked

Step 1. Reweight the payoff. With judge prior μ_0 ,

$$\ell(\mu) = \mu \frac{q}{\mu_0} + (1 - \mu) \frac{1 - q}{1 - \mu_0}, \quad w(\mu) = \begin{cases} 0 & \text{if } \mu < \frac{1}{2}, \\ \ell(\mu) & \text{if } \mu \geq \frac{1}{2}. \end{cases}$$

Since ℓ is positive and rises at most linearly in μ , the reweighted payoff w is still a single upward step, now with a sloped top.

Step 2. The optimal signal is unchanged. Because $\hat{v}$ is a step and ℓ is a positive affine function of μ , the concave closure of w touches the graph at the same two beliefs as in Section 2: $\mu_l = 0$ and $\mu_h = \frac{1}{2}$. Raising μ_h above $\frac{1}{2}$ lowers the ratio $\ell(\mu_h)/\mu_h$, and lowering μ_l to 0 frees weight for the conviction message, exactly as before. So the judge's optimal investigation is the one derived in Section 2, the same two messages and the same conviction rate of $2\mu_0$, for every prosecutor prior q . Intuitively, disagreement about the prior does not change which beliefs are cheapest to induce, because the judge's Bayes plausibility and her indifference point are hers alone, and neither depends on q .

Step 3. What the disagreement changes. It changes how much the prosecutor values the unchanged investigation. His expected payoff is

$$\text{value} = 2\mu_0 \ell\left(\frac{1}{2}\right) = q + \frac{\mu_0(1-q)}{1-\mu_0},$$

which rises in his own conviction q . At $q = \mu_0$ it equals $2\mu_0$, the Section 2 value; at the running $\mu_0 = 0.3$ the value is $(3 + 4q)/7$. A prosecutor more sure of guilt than the judge, $q > \mu_0$, values the same investigation more, because he believes convictions are more likely than the judge does.

Step 4. The prosecutor always benefits. With no investigation the judge keeps her prior $\mu_0 < \frac{1}{2}$ and acquits, so the prosecutor's payoff is 0 whatever his own prior. Since the value $q + \mu_0(1-q)/(1-\mu_0)$ is positive for every q , the prosecutor benefits from persuasion at every prior, even when he is nearly certain of guilt. This is Alonso and Cámara's striking observation: a Sender gains from generating information even when the audience is already inclined his way.

4.5 A numerical summary

Table 4 reports the prosecutor's value as his prior q varies while the judge's prior stays at 0.3. The judge's conviction rate and her welfare do not move; only the prosecutor's valuation does, climbing along the line $(3 + 4q)/7$.

Figure 4 draws the case $q = 0.5$ in the same way as Figures 1 and 3. The reweighted payoff w is again a single step, but its upper branch now slopes upward, because the prosecutor values a firmer conviction belief more than a marginal one. The concave closure still touches the graph at $\mu_l = 0$ and $\mu_h = \frac{1}{2}$, the very posteriors of Section 2, and reading it at the prior gives the prosecutor's value 0.71.

q	$\ell(\frac{1}{2})$	Prosecutor value	Judge conviction	Judge welfare
0.1	0.810	0.486	0.60	0.70
0.2	0.905	0.543	0.60	0.70
0.3	1.000	0.600	0.60	0.70
0.5	1.190	0.714	0.60	0.70
0.7	1.381	0.829	0.60	0.70
0.9	1.571	0.943	0.60	0.70

Table 4: Prosecutor value as the prosecutor’s prior q varies, judge’s prior fixed at $\mu_0 = 0.3$. The row $q = \mu_0 = 0.3$ is Section 2.

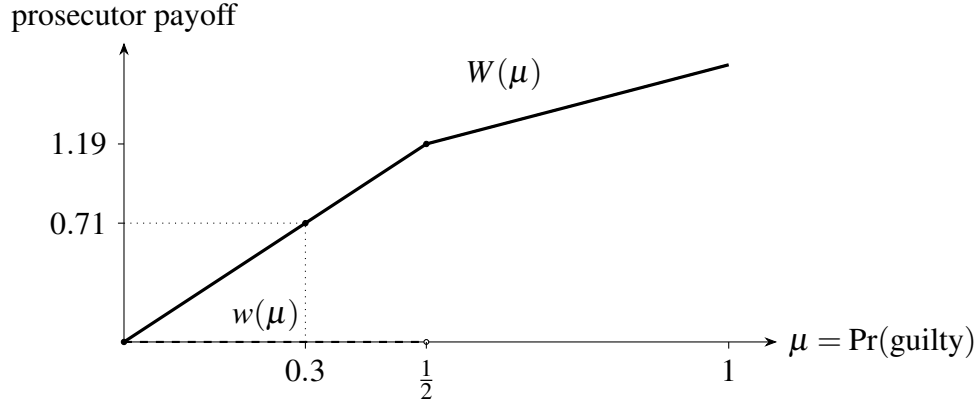


Figure 4: The reweighted payoff w (dashed) and its concave closure W (solid) at $q = 0.5$, with the same induced posteriors $\mu_l = 0$ and $\mu_h = \frac{1}{2}$ as in Figure 1.

5 Competition in Persuasion (Gentzkow and Kamenica, 2017)

5.1 Adding a second persuader

So far a single Sender controlled all the evidence. A courtroom has two. The prosecutor wants a conviction; the defense attorney wants an acquittal; each is free to commission investigations, and the judge sees the results of both. The old faith in the adversarial system holds that setting two interested persuaders against each other protects the audience. Gentzkow and Kamenica (2017) make the idea precise, and in our binary courtroom the answer is stark: two fully opposed persuaders reveal the whole truth.

5.2 The model with two senders

The states, actions, judge, and prior μ_0 are those of Section 2. Now two senders design evidence. The prosecutor earns $\hat{v}_P(\mu) = 1$ when the judge convicts ($\mu \geq \frac{1}{2}$) and 0 otherwise; the defense

attorney earns the mirror image, $\hat{v}_D(\mu) = 1$ when the judge acquits ($\mu < \frac{1}{2}$) and 0 otherwise. Each sender chooses a signal; the judge pools everything she sees into a single posterior and acts; a profile of signals is an equilibrium when neither sender can gain by changing his own signal alone. With only the prosecutor present, this is exactly Section 2.

5.3 The unifying result: both senders sit on their concave closure

The concavification of Section 2 told a lone sender where he would place the judge's belief. With two senders the same object characterizes equilibrium, through a short argument. Suppose an equilibrium leaves the judge, with positive probability, at a belief μ where the prosecutor's payoff is strictly below its concave closure, $\hat{v}_P(\mu) < V_P(\mu)$. The prosecutor could then add information at that belief, splitting it into two whose average is μ , and raise his expected payoff, since the concave closure is the most that splitting can reach. He would deviate, so this is not an equilibrium. The same argument applies to the defense attorney. Hence, in equilibrium, every belief the judge might hold lies where *both* senders sit exactly on their own concave closure.

Write $M_i = \{\mu : \hat{v}_i(\mu) = V_i(\mu)\}$ for the beliefs at which sender i cannot improve by splitting, and $M = M_P \cap M_D$. Gentzkow and Kamenica (2017) show that a Bayes-plausible distribution of posteriors is an equilibrium outcome if and only if its support lies in M .

The nesting. With a single sender, $M = M_P$, and the requirement is simply that the judge's beliefs sit on the prosecutor's concave closure, which is the concavification of Section 2. Kamenica and Gentzkow (2011) is the one-sender case; competition is what a second, opposed sender adds.

5.4 The toy model, worked

Step 1. The prosecutor's coincidence set. The prosecutor's payoff $\hat{v}_P$ is 0 below $\frac{1}{2}$ and 1 above, and its concave closure is $V_P(\mu) = \min(2\mu, 1)$. The two agree wherever $\hat{v}_P$ is already at its ceiling, the interval $[\frac{1}{2}, 1]$, together with the single belief $\mu = 0$ where both equal 0. Hence

$$M_P = \{0\} \cup [\frac{1}{2}, 1].$$

Step 2. The defense attorney's coincidence set. His payoff $\hat{v}_D$ is 1 below $\frac{1}{2}$ and 0 above, with concave closure $V_D(\mu) = \min(1, 2(1 - \mu))$. The two agree on $[0, \frac{1}{2})$, where $\hat{v}_D$ is already at its ceiling, together with $\mu = 1$ where both equal 0. Hence

$$M_D = [0, \frac{1}{2}) \cup \{1\}.$$

Step 3. Intersect. The only beliefs in both sets are the two certainties,

$$M = M_P \cap M_D = \{0, 1\}.$$

At every other belief one of the two senders would add information.

Step 4. The equilibrium. The judge's beliefs must therefore be 0 or 1, so the only Bayes-plausible outcome is full disclosure: the evidence reveals the state, leaving belief 1 with probability μ_0 and belief 0 with probability $1 - \mu_0$. The judge convicts exactly the guilty, at rate μ_0 , half the lone prosecutor's $2\mu_0$.

Intuitively, at any partial belief short of certainty, one side always holds evidence that pulls the judge its way: if the prosecutor tries to stop at the indifference belief, the defense attorney can push her toward innocence, and conversely. The only beliefs no one wishes to disturb are the two at which nothing more can be said. Competition unravels all the way to the truth.

5.5 What competition adds

Competition reveals more, and the judge gains. A lone prosecutor convicts $2\mu_0$ of defendants; two opposed persuaders convict only the guilty, μ_0 , and the judge, now fully informed, reaches her first-best. This is the sharpest form of the general comparative static: competition never reduces the information revealed, and here it forces the maximum.

Full opposition is doing the work. The unraveling is this clean because the senders' interests are exactly reversed, so their coincidence sets meet only at certainty. Gentzkow and Kamenica (2017) show, well beyond the binary case, that a competitive equilibrium is never strictly less informative than what the senders would choose by colluding, whatever their preferences; full disclosure is the extreme our courtroom lands on.

5.6 A numerical summary

Table 5 sets the one-sender and two-sender outcomes side by side at the running prior $\mu_0 = 0.3$. Adding the defense attorney halves the conviction rate and lifts the judge to full information.

Setting	Conviction rate	Prosecutor payoff	Judge welfare	Defense payoff
One sender (Section 2)	0.60	0.60	0.70	—
Two opposed senders	0.30	0.30	1.00	0.70

Table 5: One persuader versus two opposed persuaders, at $\mu_0 = 0.3$. In general the conviction rate falls from $2\mu_0$ to μ_0 and the judge's welfare rises from $1 - \mu_0$ to 1.

Figure 5 draws the two senders' payoffs. The prosecutor's step rises at $\mu = \frac{1}{2}$; the defense attorney's falls there. Each payoff meets its concave closure only on the half where it is already at its ceiling, plus the far certainty, so the two coincidence sets overlap only at $\mu = 0$ and $\mu = 1$: the beliefs of full disclosure.

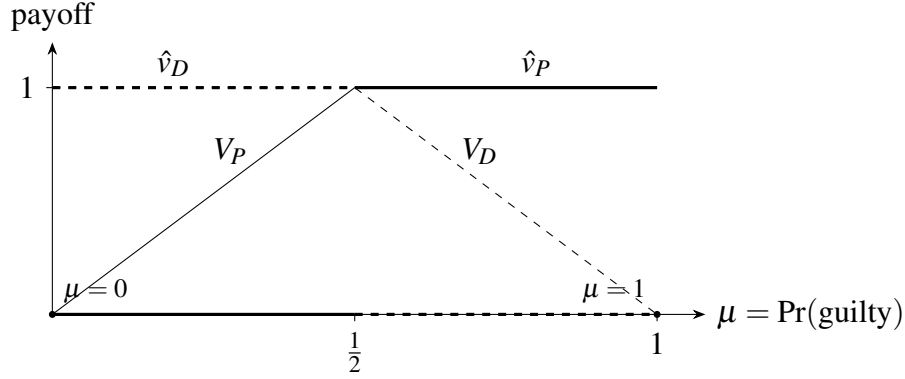


Figure 5: The prosecutor's payoff $\hat{v}_P$ (thick solid) and the defense attorney's $\hat{v}_D$ (thick dashed), with their concave closures V_P, V_D (thin). The closures are met by their own payoffs only at $\mu = 0$ and $\mu = 1$, the full-disclosure beliefs.

6 A Guided Tour of the Broader Literature

The four sections above share one move: write the Sender's payoff as a function of the Receiver's belief, and the best the Sender can do is the value of the concave closure of that payoff at the prior. Costs, disagreement, and competition enter only by reshaping that function or by intersecting two of them. The method works cleanly under three conditions: the state space is small enough to plot the value function, the Sender commits fully to the experiment, and there is a single Receiver whose action depends on her posterior. Each active branch of the literature relaxes one of these conditions. This section takes the branches in turn, keeping the prosecutor and judge in view. For each, we say what it changes relative to our baseline, describe concretely how the model would be modified to capture it, and point to representative papers, so that a student can move from our courtroom to the frontier without changing examples. The literature is large; we refer the reader to Kamenica (2019) and Bergemann and Morris (2019) for the complete bibliography.

6.1 Many receivers

What changes. There is now a panel rather than a single judge, and the Sender may address the panel with one public signal or with private, receiver-specific signals; the receivers may also care

about one another’s verdicts. *How the model changes.* Replace the judge with n jurors and a collective rule, say conviction by majority. If the prosecutor is restricted to a single *public* signal, nothing else changes: reinterpret $\bar{a}(\mu)$ as the panel’s collective verdict when all jurors share the posterior μ , and concavify $\hat{v}$ exactly as in Section 2. If instead he sends *private* signals and jurors care about the collective outcome, a juror’s vote depends not only on her own belief but on what she thinks the others learned, and the belief-based value function no longer captures the problem. One then works in the language of Bayes correlated equilibria (Bergemann and Morris 2013, 2016) and information design (Taneva 2019; Mathevet, Peregó, and Taneva 2020): characterize the outcomes attainable under *some* information structure, a set carved out by each juror’s obedience constraints, and pick the best. Persuading voters is the workhorse case (Alonso and Cámara 2016b; Wang 2015; Bardhi and Guo 2018), and the same tools yield the map of feasible profit–surplus pairs under third-degree price discrimination in Bergemann, Brooks, and Morris (2015).

6.2 Many senders: competition in persuasion

What changes. The single persuader becomes several, with possibly opposed preferences, each free to design evidence that the Receiver pools. *How the model changes.* Give sender i a value function $\hat{v}_i$ and its concave closure V_i , exactly as we did for the prosecutor and the defense attorney; an equilibrium is a Bayes-plausible belief distribution supported only where *every* sender sits on his own concave closure, that is, on $\bigcap_i M_i$. Gentzkow and Kamenica (2017) show, well beyond the binary case, that a competitive equilibrium is never strictly less informative than what the senders would choose by colluding; our full-disclosure result is the extreme this comparative static reaches when the two sides are exactly opposed. The conclusion is not automatic once senders cannot freely correlate their signals: Board and Lu (2018) show that in a search market where sellers can coordinate their disclosures, competition may reveal no more than a monopolist would, while the same sellers reveal everything once coordination is impossible. Whether senders who move sequentially rather than simultaneously reveal more is studied by Li and Norman (2018).

6.3 Limited commitment

What changes. The prosecutor no longer commits fully; with some probability he can quietly alter the message he reports, which is the assumption that had separated persuasion from cheap talk (Crawford and Sobel 1982) and from verifiable disclosure (Grossman 1981; Milgrom 1981). *How the model changes.* Add one parameter, the probability ρ that the prosecutor can secretly replace an unfavorable message with a favorable one. At $\rho = 0$ we are back in Section 2; at $\rho = 1$ the prosecutor is a pure cheap talker, and the judge trusts nothing that is not self-serving to reveal. In

between, the value is governed by an object that combines the *concavification* of full commitment with the *quasi*-concavification that solves the cheap-talk problem for a state-independent Sender (Lipnowski and Ravid 2020; Lipnowski, Ravid, and Shishkin 2022; Min 2017). A striking lesson is that the judge can be better off when the prosecutor is *less* credible, since a prosecutor who cannot be believed cannot mislead. A related relaxation lets the Sender observe something about the state before he designs the experiment, so that his very choice of investigation leaks information (Perez-Richet 2014; Hedlund 2017), and laboratory tests of how closely subjects follow these predictions are beginning to appear (Fréchette, Lizzeri, and Perego 2022).

6.4 A better-informed receiver

What changes. The judge is no longer a blank slate: she holds private information, about her own threshold for conviction or about the state, or she can gather her own evidence after seeing the prosecutor's. *How the model changes.* Give the judge a private type θ , say a privately known conviction threshold drawn from a known distribution, and redefine the prosecutor's value $\hat{v}$ by averaging over θ ; the problem stays concavification-friendly, but the optimal investigation acquires structure, becoming a monotone rule that reveals middling cases and pools extreme ones (Kolotilin, Mylovanov, Zapechelnuyk, and Li 2017; Kolotilin 2018), or an interval rule when the private information is about the state (Guo and Shmaya 2019). Rayo and Segal (2010), contemporaneously with Kamenica and Gentzkow (2011), already give the receiver a private outside option. Alternatively, let the judge pay a cost to run her own test after the prosecutor's message: the mere threat disciplines him into disclosing more, and it is without loss to look at equilibria in which she never actually investigates (Matysková 2018).

6.5 Dynamics

What changes. The problem is no longer static: the state evolves over time, or the receiver's past choices shape her future options, so the Sender chooses not only how much to reveal but when. *How the model changes.* Let the defendant's guilt evolve as a Markov process and let the judge act myopically each period on her current belief; the prosecutor now chooses a disclosure policy over time rather than a single experiment. Ely (2017) solves this case and finds that the optimal policy discloses good news with a deterministic delay. When the receiver's sunk actions change her opportunities, dynamic disclosure can sustain more of the action the Sender wants than any one-shot experiment (Ely and Szydlowski 2020). A separate dynamic branch puts beliefs directly into the utility function, modelling entertainment as the management of suspense and surprise (Ely, Frankel, and Kamenica 2015).

6.6 Large state spaces and solution methods

What changes. The state is a continuum and the Sender’s payoff depends only on the *mean* of the receiver’s posterior, so the value function can still be drawn but its concave closure no longer solves the problem, because not every distribution of posterior means with the right average is attainable.

How the model changes. Let guilt be a continuous severity in $[0, 1]$ rather than a yes-or-no, with the judge’s action depending on expected severity. Two reformulations restore tractability: represent each signal by a convex function of the posterior mean and search over that space (Gentzkow and Kamenica 2016), or cast the design as a linear program, equivalently a price-theoretic problem in which the Sender “buys” posterior means using the prior as his budget, with a simple dual test for optimality (Kolotilin 2018; Dworczak and Martini 2019). These methods say when the optimum is a monotone partition (Ivanov 2015; Mensch 2018) and now carry much of the applied work. When even they fail, computation is the fallback: the general problem is hard, but good approximation algorithms exist (Dughmi and Xu 2016; Dughmi 2017).

6.7 Applications

What changes. Nothing in the mechanism; only the names of the Sender, the Receiver, and the state. *How the model changes.* Rename the players. A regulator designing a bank stress test reveals just enough about solvency to avert a run (Goldstein and Leitner 2018; Inostroza and Pavan 2023); a school choosing a grading policy reveals just enough about ability to place its graduates (Ostrovsky and Schwarz 2010; Boleslavsky and Cotton 2015); a routing app gives drivers noisy advice to unclog roads (Das, Kamenica, and Mirka 2017); a police force randomizes speed traps to deter speeding (Lazear 2006); and, in the environmental reading we return to throughout, a firm designs a sustainability label that raises a green-minded consumer’s belief to the buying threshold and no further, which is precisely greenwashing (Lyon and Maxwell 2011). In each case the prosecutor–judge diagram is a faithful first picture, and the comparative statics of Sections 3–5 carry over: costlier testing means coarser disclosure, a more confident designer values the same disclosure more, and a watchdog forces the truth out.

6.8 A neighboring market: selling information

What changes. Information is no longer given away to move an action but sold as a good, and the designer, a data broker, a credit bureau, a rating agency, is paid in fees rather than in the Receiver’s action (Admati and Pfleiderer 1990; Bergemann and Bonatti 2019). *How the model changes.* Replace the prosecutor’s verdict payoff with a price: instead of valuing a conviction, the Sender posts a menu of experiments and prices and the Receiver buys one. The design question shifts from

what to reveal to what to sell and at what price, and the difference between selling an information structure up front and selling its realization after the fact maps onto the products that brokers and publishers actually trade. Our courtroom becomes a pricing problem, and the two literatures meet in a shared object.

7 Directions for Future Work

The prosecutor and the judge are not only a teaching device; several of the field’s open questions can be posed inside them, which is a useful discipline on speculation. We describe four directions where the model of this handout runs into the research frontier.

Combining contracts with experiments. Incentives and information are usually studied in separate literatures, yet real designers set both at once: a firm chooses a wage schedule and an evaluation system, a court sets both penalties and rules of evidence, a platform sets both prices and what buyers are shown. Inside our model this means giving the prosecutor a second instrument, a transfer to the judge, alongside the investigation, and asking how the optimal experiment changes when the Sender can also move the Receiver’s payoffs directly. Li (2017) takes a first step by adding transfers to the two-action, two-state model, but a general theory of the joint design of a contract and an experiment is missing, and it is the natural home for the question of when disclosure and payment are substitutes and when they are complements.

A theory, not a catalogue, of many receivers. The papers on persuading voters, juries, and bidders each solve a particular case, but there is no organizing statement of when a designer prefers a single public signal to receiver-specific private ones, or of how that preference turns on whether the receivers’ actions are strategic complements or substitutes. The machinery to pose the question exists, in the form of Bayes correlated equilibria (Bergemann and Morris 2016); what is missing is the comparative-statics theory built on top of it. In our setting this is the question of how the optimal investigation changes as the lone judge becomes a voting panel, and as the voting rule moves from unanimity to simple majority.

Designing information and options together, over time. A court controls both the evidence a jury sees and the order in which questions are put to it; a firm controls both an employee’s feedback and the tasks that remain open to her after she acts. Treating the design of information and the design of the receiver’s evolving choice set as one problem, rather than two, is young and promising (Ely and Szydlowski 2020), and the promotion and trial narratives used in this handout are natural laboratories for it: let the prosecutor choose not only what the judge learns but when the verdict must be cast, and the timing of disclosure becomes a second lever.

Solving harder problems, and testing the benchmark. Most models past the smallest exam-

ples resist a closed form. The convex-analysis and linear-programming methods of Section 6 have widened the solvable class, but many applied designs, especially with several interacting receivers, remain out of analytic reach, and the algorithmic methods developed in computer science are only beginning to be imported into economics (Dughmi 2017). Running alongside is an empirical question the theory cannot answer on its own: do real receivers update as Bayesian persuasion assumes? Laboratory work has just begun to test this (Fréchette, Lizzeri, and Perego 2022), and systematic evidence would tell how far to trust the rational benchmark that makes the concavification picture exact.

References

- Admati, A. R., and P. Pfleiderer (1990). Direct and Indirect Sale of Information. *Econometrica*, 58(4), 901–928.
- Alonso, R., and O. Cámara (2016a). Bayesian Persuasion with Heterogeneous Priors. *Journal of Economic Theory*, 165, 672–706.
- Alonso, R., and O. Cámara (2016b). Persuading Voters. *American Economic Review*, 106(11), 3590–3605.
- Aumann, R. J., and M. B. Maschler (1995). *Repeated Games with Incomplete Information*. Cambridge, MA: MIT Press.
- Bardhi, A., and Y. Guo (2018). Modes of Persuasion toward Unanimous Consent. *Theoretical Economics*, 13(3), 1111–1150.
- Bergemann, D., and A. Bonatti (2019). Markets for Information: An Introduction. *Annual Review of Economics*, 11, 85–107.
- Bergemann, D., B. Brooks, and S. Morris (2015). The Limits of Price Discrimination. *American Economic Review*, 105(3), 921–957.
- Bergemann, D., and S. Morris (2013). Robust Predictions in Games with Incomplete Information. *Econometrica*, 81(4), 1251–1308.
- Bergemann, D., and S. Morris (2016). Information Design, Bayesian Persuasion, and Bayes Correlated Equilibrium. *American Economic Review: Papers & Proceedings*, 106(5), 586–591.
- Bergemann, D., and S. Morris (2019). Information Design: A Unified Perspective. *Journal of Economic Literature*, 57(1), 44–95.

- Board, S., and J. Lu (2018). Competitive Information Disclosure in Search Markets. *Journal of Political Economy*, 126(5), 1965–2010.
- Boleslavsky, R., and C. Cotton (2015). Grading Standards and Education Quality. *American Economic Journal: Microeconomics*, 7(2), 248–279.
- Crawford, V. P., and J. Sobel (1982). Strategic Information Transmission. *Econometrica*, 50(6), 1431–1451.
- Das, S., E. Kamenica, and R. Mirka (2017). Reducing Congestion through Information Design. In *55th Annual Allerton Conference on Communication, Control, and Computing*, 1279–1284.
- Dughmi, S. (2017). Algorithmic Information Structure Design: A Survey. *SIGecom Exchanges*, 15(2), 2–24.
- Dughmi, S., and H. Xu (2016). Algorithmic Bayesian Persuasion. *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*, 412–425.
- Dworczak, P., and G. Martini (2019). The Simple Economics of Optimal Persuasion. *Journal of Political Economy*, 127(5), 1993–2048.
- Ely, J. C. (2017). Beeps. *American Economic Review*, 107(1), 31–53.
- Ely, J. C., A. Frankel, and E. Kamenica (2015). Suspense and Surprise. *Journal of Political Economy*, 123(1), 215–260.
- Ely, J. C., and M. Szydlowski (2020). Moving the Goalposts. *Journal of Political Economy*, 128(2), 468–506.
- Fréchette, G. R., A. Lizzeri, and J. Perego (2022). Rules and Commitment in Communication: An Experimental Analysis. *Econometrica*, 90(5), 2283–2318.
- Gentzkow, M., and E. Kamenica (2014). Costly Persuasion. *American Economic Review: Papers & Proceedings*, 104(5), 457–462.
- Gentzkow, M., and E. Kamenica (2016). A Rothschild–Stiglitz Approach to Bayesian Persuasion. *American Economic Review: Papers & Proceedings*, 106(5), 597–601.
- Gentzkow, M., and E. Kamenica (2017). Competition in Persuasion. *Review of Economic Studies*, 84(1), 300–322.
- Goldstein, I., and Y. Leitner (2018). Stress Tests and Information Disclosure. *Journal of Economic Theory*, 177, 34–69.

- Grossman, S. J. (1981). The Informational Role of Warranties and Private Disclosure about Product Quality. *Journal of Law and Economics*, 24(3), 461–483.
- Guo, Y., and E. Shmaya (2019). The Interval Structure of Optimal Disclosure. *Econometrica*, 87(2), 653–675.
- Hedlund, J. (2017). Bayesian Persuasion by a Privately Informed Sender. *Journal of Economic Theory*, 167, 229–268.
- Inostroza, N., and A. Pavan (2023). Adversarial Coordination and Public Information Design. *Northwestern working paper*.
- Ivanov, M. (2015). Optimal Signals in Bayesian Persuasion Mechanisms. *McMaster working paper*.
- Kamenica, E. (2019). Bayesian Persuasion and Information Design. *Annual Review of Economics*, 11, 249–272.
- Kamenica, E., and M. Gentzkow (2011). Bayesian Persuasion. *American Economic Review*, 101(6), 2590–2615.
- Kolotilin, A. (2018). Optimal Information Disclosure: A Linear Programming Approach. *Theoretical Economics*, 13(2), 607–635.
- Kolotilin, A., T. Mylovanov, A. Zapechelnyuk, and M. Li (2017). Persuasion of a Privately Informed Receiver. *Econometrica*, 85(6), 1949–1964.
- Lazear, E. P. (2006). Speeding, Terrorism, and Teaching to the Test. *Quarterly Journal of Economics*, 121(3), 1029–1061.
- Li, F. (2017). A Note on the Effect of Payments on Persuasion. *Economics Letters*, 156, 155–157.
- Li, F., and P. Norman (2018). On Bayesian Persuasion with Multiple Senders. *Economics Letters*, 170, 66–70.
- Lipnowski, E., and D. Ravid (2020). Cheap Talk with Transparent Motives. *Econometrica*, 88(4), 1631–1660.
- Lipnowski, E., D. Ravid, and D. Shishkin (2022). Persuasion via Weak Institutions. *Journal of Political Economy*, 130(10), 2705–2730.
- Lyon, T. P., and J. W. Maxwell (2011). Greenwash: Corporate Environmental Disclosure under Threat of Audit. *Journal of Economics & Management Strategy*, 20(1), 3–41.

- Mathevet, L., J. Perego, and I. Taneva (2020). On Information Design in Games. *Journal of Political Economy*, 128(4), 1370–1404.
- Matysková, L. (2018). Bayesian Persuasion with Costly Information Acquisition. *CERGE-EI working paper*.
- Mensch, J. (2018). Monotone Persuasion. *Games and Economic Behavior*, 122, 220–237.
- Milgrom, P. R. (1981). Good News and Bad News: Representation Theorems and Applications. *Bell Journal of Economics*, 12(2), 380–391.
- Min, D. (2017). Bayesian Persuasion under Partial Commitment. *Bank of Korea working paper*.
- Ostrovsky, M., and M. Schwarz (2010). Information Disclosure and Unraveling in Matching Markets. *American Economic Journal: Microeconomics*, 2(2), 34–63.
- Perez-Richet, E. (2014). Interim Bayesian Persuasion: First Steps. *American Economic Review: Papers & Proceedings*, 104(5), 469–474.
- Rayo, L., and I. Segal (2010). Optimal Information Disclosure. *Journal of Political Economy*, 118(5), 949–987.
- Taneva, I. (2019). Information Design. *American Economic Journal: Microeconomics*, 11(4), 151–185.
- Wang, Y. (2015). Bayesian Persuasion with Multiple Receivers. *Available at SSRN 2625399*.