STAFF
REPORTS

NO. 1194
MAY 2026

Bayesian Persuasion and Cryptography

Pablo D. Azar

FEDERAL RESERVE BANK of NEW YORK

Bayesian Persuasion and Cryptography

Pablo D. Azar

Federal Reserve Bank of New York Staff Reports, no. 1194
May 2026
https://doi.org/10.59576/sr.1194

Abstract

Bayesian Persuasion assumes that a sender can commit ex ante to an information structure and then release the realized signal ex post. This paper asks when that commitment technology can itself be implemented. After observing the state, a sender who also observes the realized signal can suppress unfavorable draws even if every disclosed signal is verifiably correct. We define Receiver-Private Certified Bayesian Persuasion, a benchmark in which the receiver obtains the signal and a certificate of correct generation while the sender does not learn the realized branch of the experiment. The main theorem shows that this benchmark is equivalent in cryptographic power to secure two-party computation. Thus cryptography is not merely an implementation device for persuasion; when the sender must be prevented from changing the signal sent to the receiver, hiding the signal from the sender is necessary. In stress-test applications, the primitive removes ex post discretion over which realized disclosure reaches depositors.

JEL classification: D82, D83, G28
Key words: Bayesian Persuasion, stress testing, central bank communications

---

Azar: Federal Reserve Bank of New York (email: pablo.azar@ny.frb.org).

1. Introduction

How should a regulator disclose information about banks when disclosure itself changes the probability of distress? Stress tests are meant to discipline weak banks and reassure markets, but the same information can coordinate withdrawals when investors fear that others will run. This is precisely the tradeoff studied in Bayesian Persuasion (Kamenica and Gentzkow, 2011): a sender designs an information structure to shape a receiver's beliefs and actions. In the stress-test setting, the sender is a regulator, the receiver is a depositor or market participant, and the hidden state is a bank's balance sheet. Full opacity blunts market discipline; full transparency can create inefficient runs. The relevant design problem is therefore neither concealment nor full disclosure, but a randomized stress-test rule that separates states enough to discipline banks while pooling them enough to avoid self-fulfilling panic.

Randomized disclosure is useful only if the regulator can bind itself before learning what it will reveal. In Bayesian Persuasion, the sender chooses a signaling rule before the state is realized, then observes the state, and then releases a signal drawn from the committed rule. That commitment assumption is the source of the model's force. Without it, the environment moves toward cheap talk (Crawford and Sobel, 1982): after observing the state, the sender would like to send whichever message is best ex post, and the receiver would understand this incentive. In the stress-test setting, a public announcement of a testing methodology is therefore not enough. Depositors must also believe that the realized disclosure was actually generated by the announced randomized procedure.

A public commitment to such a rule is not renegotiation-proof. After observing the state, the regulator may want to replace the prescribed draw with a more favorable signal. Even if each disclosed signal can be certified as drawn from the announced rule, the regulator may still observe the realized draw and suppress disclosure after an unfavorable realization. Conditional on disclosure, the receiver would then see a selected distribution rather than the distribution promised ex ante. The problem is therefore not only to verify that a released signal is valid, but to prevent signal-contingent control over whether it is released. The benchmark implicitly asks for a mediator that sees the state, draws the signal according to the committed rule, gives the realized signal to the receiver, and does not return ex post discretion over which draw gets released.

This paper addresses the problem by introducing *Receiver-Private Certified Bayesian Persuasion*. A certified persuasion mechanism implements the committed experiment rather than merely announcing it: after the state is realized, it draws the signal prescribed by the announced rule and gives the receiver both the signal and a certificate that the draw was generated correctly. The key restriction is receiver privacy. The sender does not learn the realized signal or the certificate before delivery, and therefore cannot condition disclosure on whether the random draw is favorable. Certified Bayesian Persuasion is thus the

commitment benchmark isolated above: it makes the receiver able to verify the realized disclosure while removing the ex post channel through which the sender can grind over randomized disclosures.

We implement Certified Bayesian Persuasion using secure two-party computation. Secure computation lets two parties jointly evaluate a committed rule while revealing only the outputs specified by that rule. In our setting, it lets the mechanism authenticate the state, sample the prescribed stress-test signal, and give the receiver a verifiable disclosure without revealing the realized draw to the sender. This is not a speculative technology. Secure computation can be built from standard cryptographic assumptions, including the existence of public-key cryptography, that already underlie modern telecommunications and digital commerce. Furthermore, the choice of secure two-party computation as a cryptographic primitive is not ad-hoc, as it has been shown to be a foundational primitive from which a large number of modern cryptographic protocols can be derived. (Goldreich et al., 1987; Kilian, 1988).

The use of cryptography is not a convenient implementation detail; it is necessary for the benchmark itself. Our main theorem shows that any protocol implementing Receiver-Private Certified Bayesian Persuasion implies secure two-party computation. The economic commitment problem is therefore not merely analogous to a cryptographic problem: in economic settings where the sender must be prevented from changing the signal sent to the receiver, cryptographic hiding of the signal from the sender itself becomes necessary.

The equivalence reverses the usual economic role of computational hardness: here, hardness is not a constraint on implementation but a condition for improving welfare. In algorithmic game theory and mechanism design, computational limits typically restrict what institutions can achieve: equilibria may be hard to compute (Daskalakis et al., 2009), selfish behavior can create welfare losses measured by the price of anarchy (Roughgarden, 2005), and implementable mechanisms may require a tradeoff between incentives, optimality, and simplicity (Nisan and Ronen, 2001; Hartline and Roughgarden, 2009). Certified Bayesian Persuasion has the opposite logic. The institution improves welfare precisely because bounded parties cannot undo the privacy that keeps the sender from learning the hidden branch of the randomized disclosure rule.

1.1. Practicality

The legal and technical background for certified stress-test disclosure is becoming more concrete. The GENIUS Act, signed on July 18, 2025, created a federal framework for payment stablecoins backed by liquid assets such as dollars and short-term Treasuries, while tokenized Treasuries, tokenized money market funds, and tokenized deposits have made regulated financial claims increasingly representable as auditable digital objects. This does not mean that bank supervision is about to migrate wholesale onto public blockchains. It

means that the implementation problem can be separated into two familiar parts: audits and supervisory processes authenticate the state, while cryptography enforces what can be learned from that authenticated state.

In that environment, certified stress-test disclosure has a natural implementation path. A bank commits cryptographically to balance-sheet inputs; a regulator or auditor authenticates those inputs; and a secure-computation protocol evaluates the committed disclosure rule, draws the prescribed signal, and releases to depositors only the signal and its certificate. Systems for shielded state, selective disclosure, and publicly verifiable private computation already provide pieces of this architecture (Sasson et al., 2014; Bowe et al., 2020).[1] The protocol does not replace the supervisory process that determines whether the committed state is true. It solves the distinct commitment problem: once the state has been authenticated, the regulator cannot alter the randomized disclosure that depositors are meant to see. Certified Bayesian Persuasion is therefore not a speculative vision of future financial infrastructure, but a formalization of an implementation path whose legal and technical ingredients are already visible.

1.2. Main Contributions

The paper makes three contributions. First, it gives an implementation theory for the commitment technology in Bayesian Persuasion. Recent work has studied substitutes for commitment, including reputation, repeated interaction, and restrictions to verifiable messages (Best and Quigley, 2024; Mathevet et al., 2024; Titova and Zhang, 2025). Those approaches are important, but they replace direct commitment with nearby institutions. This paper instead asks when the benchmark itself can be reproduced. The answer is Receiver-Private Certified Bayesian Persuasion: the economically relevant benchmark is anti-grinding.

Second, the paper identifies a setting in which computational limits improve rather than degrade outcomes. In much of the literature, the benchmark is a world of unbounded rationality and computation, while computational limits make equilibrium harder to reach or institutions harder to optimize. Here the ordering is reversed. The sender can be prevented from grinding over realized signals only because the relevant branch information can be hidden computationally. In this setting, computational hardness is not a friction. It is what makes the commitment benchmark attainable.

Third, the paper characterizes the exact cryptographic content of that benchmark. Receiver-Private Certified Bayesian Persuasion is equivalent in cryptographic power to secure two-party computation. This places the implementation problem at the level of private computation rather than sender-visible certification, and it yields a modular implementation

[1] Newer programmable-privacy platforms point in the same direction. For example, Aleo and Circle announced USDCx, a private and programmable USDC-backed stablecoin, on December 8, 2025.

route for applications such as stress tests, financial disclosure, credit ratings, coordination problems, investor communication, and algorithmic recommendation (Goldstein and Leitner, 2018; Goldstein and Huang, 2016; Inostroza and Pavan, 2025; Orlov et al., 2023; Alonso and Zachariadis, 2024; Bacchiocchi et al., 2022; Wu et al., 2022).

The remainder of the paper proceeds as follows. Section 2 presents the model and defines Receiver-Private Certified Bayesian Persuasion, together with its public-abort variant and simulation-based implementation notion. Section 3 proves the equivalence in cryptographic power between Receiver-Private Certified Bayesian Persuasion and secure two-party computation. Section 4 concludes.

2. The Model

2.1. Computational Bayesian Persuasion environments

We now translate the economic benchmark into a computational environment. The only new ingredient is a security parameter λ, which lets the state and signal spaces grow in the cryptographic reductions. The standard finite Bayesian Persuasion model is recovered when these objects do not vary with λ.

Here and throughout, a function

$$\varepsilon : \mathbb{N} \to \mathbb{R}_+$$

is negligible if, for every polynomial p, there exists $\bar{\lambda}$ such that

$$\varepsilon(\lambda) < \frac{1}{p(\lambda)}$$

for all $\lambda \geq \bar{\lambda}$. For two distribution ensembles $\{X_\lambda\}_{\lambda\in\mathbb{N}}$ and $\{Y_\lambda\}_{\lambda\in\mathbb{N}}$, we write

$$X_\lambda \approx_c Y_\lambda$$

if they are computationally indistinguishable: for every probabilistic polynomial-time distinguisher D,

$$|\Pr[D(X_\lambda) = 1] - \Pr[D(Y_\lambda) = 1]|$$

is negligible in λ.

Definition 1 (Computational Bayesian Persuasion environment)**.** A computational Bayesian Persuasion environment is a sequence

$$\mathcal{E} = \left\{(\Omega_\lambda, S_\lambda, A_\lambda, \mu_\lambda, u_\lambda^S, u_\lambda^R)\right\}_{\lambda\in\mathbb{N}},$$

where Ω_λ is a finite state space, S_λ is a finite signal space, A_λ is a finite action space, and

$\mu_\lambda \in \Delta(\Omega_\lambda)$ is a prior. The functions

$$u_\lambda^S, u_\lambda^R : \Omega_\lambda \times A_\lambda \to \mathbb{R}$$

are polynomial-time computable and uniformly bounded.

A signaling rule is a map

$$\pi_\lambda : \Omega_\lambda \to \Delta(S_\lambda).$$

The rule π_λ is efficiently samplable if there exists a probabilistic polynomial-time algorithm Samp_π such that, for every $\omega \in \Omega_\lambda$,

$$\mathsf{Samp}_\pi(1^\lambda, \omega) \sim \pi_\lambda(\cdot \mid \omega).$$

The prior μ_λ is efficiently samplable if there exists a probabilistic polynomial-time algorithm Samp_μ such that

$$\mathsf{Samp}_\mu(1^\lambda) \sim \mu_\lambda.$$

The standard finite Bayesian Persuasion model is the special case in which $\Omega_\lambda, S_\lambda, A_\lambda, \mu_\lambda, u_\lambda^S, u_\lambda^R$ do not vary with λ. We allow these objects to vary with the security parameter because the cryptographic reductions below use state and signal spaces whose size grows with λ.

For a fixed security parameter λ, the direct-commitment benchmark is as follows. The sender commits ex ante to a signaling rule

$$\pi_\lambda : \Omega_\lambda \to \Delta(S_\lambda).$$

Nature draws

$$\omega \sim \mu_\lambda,$$

and the sender observes ω. A signal is then drawn according to

$$s \sim \pi_\lambda(\cdot \mid \omega).$$

The receiver observes s, forms the posterior

$$\mu_\lambda(\omega \mid s) = \frac{\mu_\lambda(\omega)\pi_\lambda(s \mid \omega)}{\sum_{\omega' \in \Omega_\lambda} \mu_\lambda(\omega')\pi_\lambda(s \mid \omega')}$$

whenever the denominator is positive, and chooses an action $a \in A_\lambda$.

As usual, the sender's choice can also be represented as a choice of a distribution over posterior beliefs. A signaling rule π_λ, together with the prior μ_λ, induces a distribution

$$\tau_\lambda \in \Delta(\Delta(\Omega_\lambda))$$

over posterior beliefs. Any such induced distribution satisfies Bayes plausibility:

$$\mathbb{E}_{\hat{\mu}\sim\tau_\lambda}[\hat{\mu}] = \mu_\lambda.$$

Conversely, any Bayes-plausible distribution over posterior beliefs can be induced by some signaling rule. The paper's question is not how to solve the sender's Bayes-plausible optimization problem. The question is whether the commitment technology assumed by that benchmark can be implemented cryptographically.

2.2. Receiver-Private Certified Bayesian Persuasion

For compactness, write Receiver-Private CBP for Receiver-Private Certified Bayesian Persuasion.

We now define the ideal primitive. Unlike a concrete protocol, the ideal object does not include commitments, certificates, proofs, ciphertexts, or transcripts. Those are implementation devices. In a real protocol, the receiver's transcript may contain a certificate or proof that the disclosed signal was generated correctly. In the ideal functionality, that validity is built in, so the object specifies only the information that each party should learn.

Definition 2 (Receiver-Private CBP)**.** Fix an efficiently samplable prior μ_λ and an efficiently samplable signaling rule π_λ. The Receiver-Private Certified Bayesian Persuasion functionality

$$\mathcal{F}^{\mathrm{rpCBP}}_{\mu,\pi}$$

proceeds as follows.

1. The functionality samples
$$\omega \sim \mu_\lambda$$
and sends ω to the sender.

2. The functionality samples
$$s \sim \pi_\lambda(\cdot \mid \omega)$$
and sends s to the receiver.

3. The sender receives no output about s.

Thus Receiver-Private CBP implements the direct-commitment Bayesian Persuasion experiment with one additional privacy restriction: the realized signal is hidden from the sender. The receiver learns the signal s, and no more than s. The sender learns the state ω, and no more about the random branch of the signaling rule than is already implied by (ω, π_λ).

This receiver-privacy requirement is the anti-grinding condition. If the sender learns s before delivery, then a strategic sender can disclose favorable realizations and suppress unfavorable realizations. In the ideal Receiver-Private functionality, that deviation is impossible because the sender never observes the realized signal.

This privacy requirement is also natural in the bank-regulator-depositor application. A bank can commit its balance-sheet information, and a regulator can verify that committed state and invoke the agreed disclosure rule, but the realized stress-test signal should then be delivered directly to depositors, or to a public bulletin board, without first passing through the regulator's discretionary hands. Otherwise the regulator could observe a favorable or unfavorable draw and decide whether to release it, which would collapse the randomized disclosure rule into a selective disclosure policy.

For applications, it is useful to allow public non-delivery. This captures the fact that a cryptographic protocol by itself cannot force an unwilling party to participate.

Definition 3 (Public-abort Receiver-Private CBP)**.** Fix a default outcome

$$\bot \notin S_\lambda.$$

The public-abort functionality

$$\mathcal{F}_{\mu,\pi}^{\mathrm{rpCBP},\bot}$$

proceeds as follows.

1. The functionality samples

   $$\omega \sim \mu_\lambda$$

   and sends ω to the sender.

2. After observing ω, the sender sends either continue or abort to the functionality.

3. If the sender sends abort, the functionality sends the public outcome $\bot$ to the receiver and halts.

4. If the sender sends continue, the functionality samples

   $$s \sim \pi_\lambda(\cdot \mid \omega)$$

   and sends s to the receiver.

5. The sender receives no output about s.

The default outcome $\bot$ is public. It is not a hidden deviation. In this functionality, the sender may condition participation on the state ω, because the sender observes ω. But the sender cannot condition delivery on the realized signal s, because s is sampled only after the continuation decision and is not revealed to the sender.

2.3. *Implementation*

A real protocol replaces the ideal functionality with messages, commitments, ciphertexts, proofs, and verification rules. We use a simulation-based implementation notion because it makes precise the claim that cryptography implements, rather than changes, the underlying economic benchmark. The *ideal world* is the thought experiment in which a trusted mediator directly runs the functionality $\mathcal{F}_{\mu,\pi}^{\mathrm{rpCBP}}$: the mediator samples the state, delivers to the sender exactly the information the model says the sender should learn, delivers to the receiver exactly the signal the model says the receiver should learn, and reveals nothing else. The *real world* is the actual protocol run by strategic parties who exchange commitments, ciphertexts, and proofs instead of interacting with a trusted mediator.

The role of the simulator is to translate any real-world attack into an ideal-world attack. Suppose, for example, that the receiver is corrupted. In the real protocol, the corrupted receiver sees a full transcript and may try to extract information from its cryptographic form. In the ideal world, by contrast, the corrupted receiver should see only the signal s prescribed by the functionality, together with the public description of the environment. A simulator is an efficient algorithm that is given only that ideal information and must manufacture a fake transcript that looks like the real one. If this can be done, then the transcript carries no economically relevant information beyond the signal itself. The same logic applies when the sender is corrupted: any real-world deviation must correspond to some ideal-world strategy with the same effective information and the same effective ability to manipulate outcomes. The distinguisher $\mathcal{Z}$ is simply an outside observer that chooses inputs, watches the interaction, and tries to tell whether it is seeing the real protocol or the ideal functionality. If no polynomial-time distinguisher can tell the difference, then the protocol is an implementation of the ideal commitment technology in the only sense that matters for computationally bounded agents.

Let $\Pi = \{\Pi_\lambda\}_{\lambda\in\mathbb{N}}$ be a two-party protocol between a sender and a receiver, with public input descriptions of

$$(\mu_\lambda, \pi_\lambda).$$

The receiver's output is either a signal $s \in S_\lambda$, or in the public-abort variant the default outcome $\perp$. The sender's view consists of its input, internal randomness, and all messages it observes during the protocol.

Definition 4 (Implementation of Receiver-Private CBP)**.** A protocol family Π implements

$$\mathcal{F}_{\mu,\pi}^{\mathrm{rpCBP}}$$

if, for every probabilistic polynomial-time adversary $\mathcal{A}$ corrupting either party in the real protocol, there exists a probabilistic polynomial-time simulator $\mathcal{S}$ in the ideal world such

that, for every probabilistic polynomial-time distinguisher $\mathcal{Z}$,

$$\mathsf{Real}_{\Pi,\mathcal{A},\mathcal{Z}}(1^\lambda) \approx_c \mathsf{Ideal}_{\mathcal{F}_{\mu,\pi}^{\mathrm{rpCBP}},\mathcal{S},\mathcal{Z}}(1^\lambda).$$

The analogous definition applies to

$$\mathcal{F}_{\mu,\pi}^{\mathrm{rpCBP},\perp}$$

in the public-abort model.

Equivalently, the implementation must satisfy the following four economic and cryptographic requirements.

Completeness. If both parties follow the protocol, then the receiver outputs a signal distributed as

$$s \sim \pi_\lambda(\cdot \mid \omega), \qquad \omega \sim \mu_\lambda,$$

except with negligible error.

Soundness. A malicious sender cannot cause the receiver to accept an output whose distribution is inconsistent with the ideal functionality. In the public-abort variant, this means that for every malicious sender there is an ideal-world strategy that, after observing ω, either aborts and produces $\perp$, or continues and allows the functionality to sample

$$s \sim \pi_\lambda(\cdot \mid \omega).$$

Thus any abort may be state-contingent, but not signal-contingent.

Selective disclosure. For every malicious receiver, the receiver's real-world view can be simulated given only the receiver's ideal output s, or $\perp$ in the public-abort variant, together with the public description of $(\mu_\lambda, \pi_\lambda)$. Thus the receiver learns no more about ω than what is implied by the realized signal.

Receiver privacy. For every malicious sender, the sender's real-world view can be simulated given only the sender's ideal information: the public description of $(\mu_\lambda, \pi_\lambda)$ and the realized state ω. In the public-abort variant, the simulator is also given the sender's own continuation or abort decision. Thus the sender learns no more about the realized signal than what is already implied by (ω, π_λ).

The last condition is what rules out signal grinding. Ordinary certification can prove that a disclosed signal was computed correctly. Receiver privacy ensures that the sender does not learn the realized branch before delivery, and therefore cannot selectively retain favorable draws.

Proposition 1 (Honest-path payoff equivalence)**.** Fix an efficiently samplable signaling rule π_λ and a polynomial-time receiver strategy α_λ. Suppose Π implements

$$\mathcal{F}_{\mu,\pi}^{\mathrm{rpCBP}}.$$

Then, under honest play, the sender's and receiver's expected payoffs in the real protocol differ from their expected payoffs in the direct-commitment Bayesian Persuasion benchmark by at most a negligible function of λ.

The same conclusion holds for the public-abort functionality conditional on the sender choosing continue.

Proof. Under honest play, implementation implies that the joint distribution of the receiver's output in the real protocol is computationally indistinguishable from the ideal distribution generated by

$$\omega \sim \mu_\lambda, \qquad s \sim \pi_\lambda(\cdot \mid \omega).$$

Selective disclosure implies that any additional transcript information observed by the receiver can be simulated from s. Therefore any polynomial-time receiver strategy induces the same action distribution in the real and ideal executions, up to negligible error. Since utilities are uniformly bounded and polynomial-time computable, any non-negligible payoff difference would yield a polynomial-time distinguisher between the real and ideal executions. Hence the payoff difference is negligible. □

2.4. Random oblivious transfer

The formal proof uses Random oblivious transfer as its canonical cryptographic primitive. For economists, the relevant implication is that Receiver-Private Certified Bayesian Persuasion is equivalent in cryptographic power to secure two-party computation: Random OT is the standard complete primitive through which that equivalence is proved (Goldreich et al., 1987; Kilian, 1988).

Definition 5 (Random oblivious transfer)**.** For message length $k = k(\lambda)$, the Random OT functionality

$$\mathcal{F}_k^{\mathrm{Random-OT}}$$

samples

$$X_0, X_1 \leftarrow \{0,1\}^k, \qquad B \leftarrow \{0,1\},$$

independently and uniformly at random. It sends

$$(X_0, X_1)$$

to the sender and sends

$$(B, X_B)$$

to the receiver. The sender receives no information about B, and the receiver receives no information about X_{1-B} beyond what is implied by (B, X_B).

The public-abort version

$$\mathcal{F}_k^{\text{Random}-\text{OT},\perp}$$

is defined analogously: after receiving (X_0, X_1), the sender may either continue or abort. If the sender aborts, the receiver receives $\perp$. If the sender continues, the receiver receives (B, X_B), while the sender learns nothing about B.

3. Main Theorem: The Cryptographic Content of Certified Persuasion

Our main theorem shows that Receiver-Private Certified Bayesian Persuasion and Random OT can be derived from one another. Since Random OT is complete for secure two-party computation, this is the formal sense in which the benchmark is equivalent in cryptographic power to secure two-party computation (Goldreich et al., 1987; Kilian, 1988).

Theorem 1 (Receiver-Private CBP and Random OT)**.** *Receiver-Private Certified Bayesian Persuasion and Random OT are equivalent as cryptographic primitives.*

First, for every message length $k = k(\lambda)$, a single instance of Receiver-Private CBP realizes k-bit Random OT.

Second, Random OT realizes Receiver-Private CBP for every efficiently samplable computational Bayesian Persuasion environment.

The same equivalence holds for the public-abort variants

$$\mathcal{F}_{\mu,\pi}^{\text{rpCBP},\perp} \qquad \textit{and} \qquad \mathcal{F}_k^{\text{Random}-\text{OT},\perp}.$$

Before giving the proof of the main equivalence theorem, we isolate one direction as a separate proposition, which follows immediately from the theorem. This is the implementation direction: once Random OT is available, Receiver-Private CBP can be implemented. The proof is useful because it explains why the primitive is economically natural. Random OT first gives the parties a private way to choose one item from a menu; secure computation then lets them use that private choice technology to emulate the trusted Bayesian Persuasion mediator.

Proposition 2 (Random OT implements Receiver-Private CBP)**.** Fix an efficiently samplable computational Bayesian Persuasion environment

$$\mathcal{E}_\lambda = (\Omega_\lambda, S_\lambda, A_\lambda, \mu_\lambda, u_\lambda^S, u_\lambda^R)$$

with polynomial-length encodings of states and signals. Fix an efficiently samplable signaling rule

$$\pi_\lambda : \Omega_\lambda \to \Delta(S_\lambda).$$

Suppose that the parties have access to a secure implementation of Random OT in the standard simulation-based model, with public abort. Then there exists a probabilistic polynomial-time protocol that securely realizes the public-abort Receiver-Private CBP functionality

$$\mathcal{F}_{\mu,\pi}^{\mathrm{rpCBP},\perp}.$$

In particular, any cryptographic assumption that yields a secure implementation of Random OT also yields a secure implementation of Receiver-Private CBP for every efficiently samplable prior and signaling rule.

OT is known to be implementable under standard public-key assumptions. Classical implementations can be obtained from factoring-style assumptions, such as Rabin-type oblivious transfer and its factoring-based refinements, or from assumptions such as decisional Diffie-Hellman, quadratic residuosity, and decisional composite residuosity. Post-quantum implementations can be obtained from lattice assumptions, including the Learning with Errors assumption. Thus the Random OT assumption used below can be instantiated either from classical public-key assumptions or from post-quantum lattice assumptions, depending on the desired security model. For this paper, we do not need to commit to a particular implementation.

Proof of Theorem 1. We prove the two reductions.

Receiver-Private CBP implies Random OT. Fix a message length $k = k(\lambda)$. Define a Bayesian Persuasion environment by

$$\Omega_\lambda = \{0,1\}^k \times \{0,1\}^k,$$

and let μ_λ be the uniform distribution on Ω_λ. A state is therefore

$$\omega = (X_0, X_1),$$

where X_0 and X_1 are independent uniform k-bit strings.

Let the signal space be

$$S_\lambda = \{0,1\} \times \{0,1\}^k.$$

Define the signaling rule $\pi_\lambda^{\mathrm{Random-OT}}$ by

$$\pi_\lambda^{\mathrm{Random-OT}}((b,x) \mid (x_0,x_1)) = \begin{cases} 1/2 & \text{if } x = x_b, \\ 0 & \text{otherwise.} \end{cases}$$

Equivalently, the rule draws

$$B \leftarrow \{0,1\}$$

uniformly at random and outputs

$$s = (B, X_B).$$

Now run the Receiver-Private CBP functionality

$$\mathcal{F}^{\mathrm{rpCBP}}_{\mu,\pi^{\mathrm{Random-OT}}}.$$

The functionality first draws

$$\omega = (X_0, X_1) \sim \mu_\lambda$$

and sends (X_0, X_1) to the sender. It then samples

$$s = (B, X_B)$$

according to $\pi_\lambda^{\mathrm{Random-OT}}(\cdot \mid X_0, X_1)$ and sends s to the receiver. Receiver privacy guarantees that the sender does not learn B. Selective disclosure guarantees that the receiver learns no more than (B, X_B).

Thus the joint distribution of outputs is exactly the Random OT distribution:

$$\text{sender receives } (X_0, X_1), \qquad \text{receiver receives } (B, X_B).$$

Therefore Receiver-Private CBP realizes Random OT.

For the public-abort variant, the same construction gives

$$\mathcal{F}_k^{\mathrm{Random-OT},\perp}.$$

The sender may abort after receiving (X_0, X_1). If it does not abort, the receiver receives (B, X_B), and the sender does not learn B.

Random OT implies Receiver-Private CBP. We prove the reverse direction in three steps. The first step explains how Random OT can be used to implement ordinary one-out-of-two OT. The second recalls the standard fact that ordinary OT is sufficient to implement any efficient two-party computation. The third applies that fact to the Receiver-Private CBP mediator.

Step 1: Random OT implements ordinary OT. Ordinary one-out-of-two OT is the following primitive. A sender has two messages

$$M_0, M_1 \in \{0,1\}^k.$$

A receiver has a choice bit

$$b \in \{0,1\}.$$

At the end, the receiver should learn M_b, and should learn nothing about M_{1-b}. The sender should learn nothing about b.

A single call to Random OT gives the sender two random strings

$$X_0, X_1 \leftarrow \{0,1\}^k,$$

and gives the receiver one of those strings, together with its index:

$$(B, X_B), \qquad B \leftarrow \{0,1\}.$$

Thus Random OT creates a hidden one-time-pad menu.[2] The sender holds two random pads, X_0 and X_1. The receiver holds exactly one pad, X_B, but the sender does not know which pad the receiver holds.

To turn this into ordinary OT, the receiver sends the sender the shifted index

$$d = b \oplus B,$$

where $\oplus$ denotes addition modulo two. This message does not reveal b. From the sender's perspective, B is an independent uniform bit, so

$$d = b \oplus B$$

is also uniform, regardless of whether $b = 0$ or $b = 1$.

The sender then uses d to align the messages with the random pads. For each $i \in \{0,1\}$, the sender sends

$$C_i = M_{i \oplus d} \oplus X_i.$$

Equivalently,

$$C_0 = M_d \oplus X_0, \qquad C_1 = M_{1 \oplus d} \oplus X_1.$$

[2] A one-time-pad X is a uniformly random binary string of length k. Given a binary string message $M \in \{0,1\}^k$, the bit-wise modulo 2 addition $C = X \oplus M$ is a perfectly encrypted ciphertext. Since bit-wise addition modulo 2 is its own inverse, the ciphertext C can be decrypted by computing $X \oplus C = X \oplus X \oplus M = M$.

The receiver knows B and X_B, so it can open the B-th ciphertext:

$$C_B \oplus X_B = (M_{B\oplus d} \oplus X_B) \oplus X_B = M_{B\oplus d}.$$

Since

$$d = b \oplus B,$$

we have

$$B \oplus d = B \oplus b \oplus B = b.$$

Therefore the receiver obtains

$$C_B \oplus X_B = M_b.$$

The privacy properties are exactly those of ordinary OT. The sender sees only

$$d = b \oplus B,$$

which is uniformly random from the sender's perspective, so the sender learns nothing about b. The receiver sees both ciphertexts C_0, C_1, but it knows only one pad, X_B. The other pad X_{1-B} is uniform and unknown to the receiver, so the other ciphertext remains one-time-padded. Hence the receiver learns M_b and not M_{1-b}. Thus Random OT realizes ordinary one-out-of-two OT.

Step 2: Ordinary OT implements secure computation. The next step uses a standard completeness theorem from cryptography: ordinary OT is complete for secure two-party computation (Goldreich et al., 1987; Kilian, 1988). For our purposes, the theorem can be read as a statement about trusted mediators.

Suppose two parties want to emulate a mediator who does the following:

takes private inputs, uses private randomness, computes an efficient function,

and then sends different outputs to the two parties. If the mediator's computation can be written as a polynomial-size circuit, then an OT-based secure-computation protocol can emulate that mediator. The real protocol reveals to each party only what that party would have learned from the mediator.

The economic interpretation is useful. Secure computation lets the parties replace a trusted mediator with a protocol. The protocol may involve many encrypted messages, but the final information structure is the same as in the mediator model. A party who should receive only its own output can simulate its entire protocol transcript from that output. Thus the protocol does not reveal the hidden inputs, hidden random coins, or hidden intermediate variables of the computation.

We use this theorem as a black box. Since Step 1 showed that Random OT implements

ordinary OT, Random OT also implements any efficient two-party mediator.

Step 3: The Receiver-Private CBP mediator is an efficient two-party computation. Now fix an efficiently samplable prior μ_λ and an efficiently samplable signaling rule

$$\pi_\lambda : \Omega_\lambda \to \Delta(S_\lambda).$$

Because μ_λ is efficiently samplable, there is a polynomial-time algorithm

$$\mathsf{Samp}_\mu(1^\lambda; R_\mu)$$

that outputs a state distributed as μ_λ. Because π_λ is efficiently samplable, there is a polynomial-time algorithm

$$\mathsf{Samp}_\pi(1^\lambda, \omega; R_\pi)$$

that outputs a signal distributed as $\pi_\lambda(\cdot \mid \omega)$.

Consider the following trusted mediator. It has two stages.

State stage. The mediator draws private random coins R_μ, computes

$$\omega = \mathsf{Samp}_\mu(1^\lambda; R_\mu),$$

and sends ω to the sender. The receiver receives nothing at this stage. The mediator also keeps an internal authenticated copy of ω, so that the same state will be used in the next stage.

Disclosure stage. After observing ω, the sender sends a continuation decision

$$q \in \{\mathsf{continue}, \mathsf{abort}\}$$

to the mediator. If

$$q = \mathsf{abort},$$

the mediator sends the public default outcome

$$\perp$$

to the receiver and halts. If

$$q = \mathsf{continue},$$

the mediator draws fresh private random coins R_π, computes

$$s = \mathsf{Samp}_\pi(1^\lambda, \omega; R_\pi),$$

and sends s to the receiver. The sender receives no output about s.

This mediator is exactly the public-abort Receiver-Private CBP functionality

$$\mathcal{F}_{\mu,\pi}^{\mathrm{rpCBP},\perp}.$$

It is also an efficient randomized two-party functionality: the only computations it performs are the efficient samplers for μ_λ and π_λ, together with a simple branch on the sender's continuation decision.

Therefore, by OT completeness, ordinary OT securely implements this mediator. Since Random OT implements ordinary OT by Step 1, Random OT securely implements

$$\mathcal{F}_{\mu,\pi}^{\mathrm{rpCBP},\perp}.$$

Why this gives receiver privacy. The secure-computation implementation has the same information structure as the mediator. A corrupted receiver's view can be simulated from its ideal output,

$$s \quad \text{or} \quad \perp,$$

and the public description of $(\mu_\lambda, \pi_\lambda)$. Thus the receiver learns no more about ω than what is implied by the realized signal.

Similarly, a corrupted sender's view can be simulated from its ideal information:

$$\omega, \qquad q, \qquad \text{and the public description of } (\mu_\lambda, \pi_\lambda).$$

The simulator is not given s, and does not need s, because in the ideal Receiver-Private functionality the sender never learns the realized signal. Thus the sender learns no more about the random branch of the signaling rule than what is already implied by (ω, π_λ).

This is the anti-grinding property. The sender can condition its continuation decision on the state ω, because the sender observes ω. But it cannot condition that decision on the realized signal s, because s is generated only after continuation and is delivered only to the receiver. Any failure to continue is therefore state-contingent, not signal-contingent, and appears publicly as $\perp$.

Thus Random OT realizes public-abort Receiver-Private CBP. The no-abort version follows as the special case in which the sender's continuation decision is fixed to continue.

□

3.1. Interpretation

The theorem shows that Receiver-Private Certified Bayesian Persuasion is not merely Bayesian Persuasion plus a certificate. It has exactly the cryptographic content of Random OT.

The reduction from Receiver-Private CBP to Random OT uses the prior μ_λ in an essential way. The two random OT messages are the components of the randomly drawn state:

$$\omega = (X_0, X_1) \sim \mu_\lambda.$$

The random choice bit is the random branch of the signaling rule:

$$B \leftarrow \{0, 1\}.$$

The receiver's OT output is the realized signal:

$$s = (B, X_B).$$

Thus the Bayesian ingredients map directly into the Random OT ingredients:

Random OT	Receiver-Private CBP
(X_0, X_1)	$\omega \sim \mu_\lambda$
sender receives (X_0, X_1)	sender observes ω
B	random branch of $\pi_\lambda(\cdot \mid \omega)$
(B, X_B)	$s \sim \pi_\lambda(\cdot \mid \omega)$
sender does not learn B	sender does not learn s
receiver does not learn X_{1-B}	receiver learns no more than s

This equivalence explains why receiver privacy is the natural anti-grinding condition. A sender-visible certificate can certify that a disclosed signal was computed correctly, but it cannot prevent the sender from observing the realized signal and suppressing unfavorable draws. Receiver-Private CBP removes the realized signal from the sender's information set. The sender may still condition participation on the state in the public-abort version, but it cannot condition delivery on the realized branch of the random experiment.

4. Conclusion

This paper argues that the right implementation problem in Bayesian Persuasion is not sender-visible certification, but anti-grinding commitment. A sender who sees the realized draw can selectively release favorable signals and suppress unfavorable ones, even when every disclosed signal is verifiably correct. The relevant benchmark is therefore Receiver-Private Certified Bayesian Persuasion: the receiver obtains the signal and its certificate, while the sender does not observe the realized branch of the random experiment.

The main theorem shows that this benchmark is equivalent in cryptographic power to secure two-party computation. This equivalence identifies the cryptographic content of commitment in persuasion. In economic settings where the sender must be prevented from

changing the signal sent to the receiver, the signal must be hidden from the sender itself. In applications such as stress tests, audits and supervisory processes still authenticate the underlying state; cryptography enforces the committed disclosure rule after that state has been authenticated. Computational hardness is therefore not merely a constraint on institutional design. In this setting, it is the resource that makes welfare-improving disclosure credible.

REFERENCES

Ricardo Alonso and Konstantinos E. Zachariadis. Persuading large investors. *Journal of Economic Theory*, 222:105933, 2024. doi: 10.1016/j.jet.2024.105933.

Francesco Bacchiocchi, Matteo Castiglioni, Alberto Marchesi, Giulia Romano, and Nicola Gatti. Public signaling in bayesian ad auctions. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, IJCAI-22, pages 39–45. International Joint Conferences on Artificial Intelligence Organization, 2022. doi: 10.24963/ijcai.2022/6. Main Track.

James Best and Daniel Quigley. Persuasion for the long run. *Journal of Political Economy*, 132(5):1740–1791, 2024. doi: 10.1086/727282.

Sean Bowe, Alessandro Chiesa, Matthew Green, Ian Miers, Pratyush Mishra, and Howard Wu. Zexe: Enabling decentralized private computation. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 947–964. IEEE, 2020. doi: 10.1109/SP40000.2020.00050.

Vincent P. Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 50 (6):1431–1451, 1982. doi: 10.2307/1913390.

Constantinos Daskalakis, Paul W. Goldberg, and Christos H. Papadimitriou. The complexity of computing a nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259, 2009. doi: 10.1137/070699652.

Oded Goldreich, Silvio Micali, and Avi Wigderson. How to play any mental game or a completeness theorem for protocols with honest majority. In *Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing*, STOC '87, pages 218–229, New York, NY, USA, 1987. Association for Computing Machinery. doi: 10.1145/28395.28420.

Itay Goldstein and Chong Huang. Bayesian persuasion in coordination games. *American Economic Review*, 106(5):592–596, 2016. doi: 10.1257/aer.p20161047.

Itay Goldstein and Yaron Leitner. Stress tests and information disclosure. *Journal of Economic Theory*, 177:34–69, 2018. doi: 10.1016/j.jet.2018.05.013.

Jason D. Hartline and Tim Roughgarden. Simple versus optimal mechanisms. In *Proceedings of the 10th ACM Conference on Electronic Commerce*, EC '09, pages 225–234, New York, NY, USA, 2009. Association for Computing Machinery. doi: 10.1145/1566374.1566407.

Nicolas A. Inostroza and Alessandro Pavan. Adversarial coordination and public information design. *Theoretical Economics*, 20(2):763–813, 2025. doi: 10.3982/TE5768.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011. doi: 10.1257/aer.101.6.2590.

Joe Kilian. Founding cryptography on oblivious transfer. In *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing*, STOC '88, pages 20–31, New York, NY, USA, 1988. Association for Computing Machinery. doi: 10.1145/62212.62215.

Laurent Mathevet, David G. Pearce, and Ennio Stacchetti. Reputation and information design. Working paper, 2024.

Noam Nisan and Amir Ronen. Algorithmic mechanism design. *Games and Economic Behavior*, 35(1–2):166–196, 2001. doi: 10.1006/game.1999.0790.

Dmitry Orlov, Pavel Zryumov, and Andrzej Skrzypacz. The design of macroprudential stress tests. *Review of Financial Studies*, 36(11):4460–4501, 2023. doi: 10.1093/rfs/hhad040.

Tim Roughgarden. *Selfish Routing and the Price of Anarchy*. MIT Press, Cambridge, MA, 2005. ISBN 9780262182430.

Eli Ben Sasson, Alessandro Chiesa, Christina Garman, Matthew Green, Ian Miers, Eran Tromer, and Madars Virza. Zerocash: Decentralized anonymous payments from bitcoin. In *2014 IEEE Symposium on Security and Privacy*, pages 459–474, 2014. doi: 10.1109/SP.2014.36.

Maria Titova and Kun Zhang. Persuasion with verifiable information. *Journal of Economic Theory*, 230:106102, 2025. doi: 10.1016/j.jet.2025.106102.

Jibang Wu, Zixuan Zhang, Zhe Feng, Zhaoran Wang, Zhuoran Yang, Michael I. Jordan, and Haifeng Xu. Sequential information design: Markov persuasion process and its efficient reinforcement learning. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, EC '22, pages 471–472, New York, NY, USA, 2022. Association for Computing Machinery. doi: 10.1145/3490486.3538313.