Self-Revealing Renegotiation

Andrea Attar, Lorenzo Bozzoli, Roland Strausz*

May 4, 2026

Abstract

We revisit the tension between the legal doctrine of renegotiation and economic efficiency. We introduce *self-revealing mechanisms* that combine bidirectional communication (the agent sends and receives information) with conditional disclosure (communication remains private during renegotiation but becomes verifiable at contract execution). Within the canonical Fudenberg and Tirole (1990) framework, we design a self-revealing mechanism that fully mitigates the renegotiation threat by uniquely implementing the second-best allocation. Thus, we achieve the full-commitment outcome while satisfying renegotiation-proofness. Our optimal mechanism is structurally simple, and exploits signal disclosures to the agent to construct incentive-compatible punishments, which she activates upon observing a renegotiation offer.It satisfies standard commitment assumptions by only conditioning decisions on public information, without requiring any third-party enforcement. In practical terms, it can be implemented using existing smart-contract techniques. Our results extend to general settings of renegotiation. (*JEL* D43, D82, D86)

*Attar: CNRS, Toulouse School of Economics University of Toulouse Capitole, and Università degli Studi di Roma "Tor Vergata" (email: andrea.attar@tse-fr.eu); Bozzoli: Toulouse School of Economics University of Toulouse Capitole' (email: lorenzo.bozzoli@tse-fr.eu); Strausz: School of Business and Economics, Humboldt-Universität zu Berlin (email: strauszr@hu-berlin.de). This paper supersedes a previous version circulated under the title "Mediated Renegotiation". We thank Eloisa Campioni, Dino Gerardi, Samuel Häfner, Johannes Horner, Fahad Khalil, Daniel Krähmer, Elliot Lipnowski, Alessandro Pavan, Soenje Reiche, Francois Salanié, Steve Tadelis, and Takuro Yamashita for very thoughtful comments. We also thank seminar audiences at Berkeley University, Bonn University, Collegio Carlo Alberto, Northwestern University, Università degli Studi di Roma "Tor Vergata", Toulouse School of Economics, Washington University, Yale University, as well as conference participants at the 2024 Conference on Mechanism and Institution Design (Budapest), at the 2024 Conference in honor of Françoise Forges, at the 2024 Game Theory and Information Economics Conference (Rio de Janeiro), at the 2025 SAET Conference (Ischia), and at the 2025 Unibg IO Workshop for many useful discussions. Andrea Attar and Lorenzo Bozzoli acknowledge financial support from the Agence Nationale de la Recherche (ANR) (Programme d'Investissements d'Avenir ANR-17-EURE- 0010 and project ANR-23-CE26-0006), and from Ministero dell' Università e della Ricerca, (project PRIN-2022-PXE3B7). Roland Strausz acknowledges financial support from the European Union through the ERC-grant PRIVDIMA (project number 101096682) and the Deutsche Forschungsgemeinschaft through CRC-TRR 190 (project number 280092119).

1 Introduction

The threat of renegotiation is ubiquitous in contracting. As Dewatripont (1989) first pointed out, this threat arises because contracts that optimally resolve incentive problems typically implement allocations that are inefficient ex post. When parties cannot contractually prevent renegotiating away these inefficiencies, the resulting collective opportunism pushes them towards allocations that are inefficient from an ex-ante perspective.

The perverse effects of collective opportunism reflect a tension between economic efficiency and legal doctrine (Jolls, 1997; Davis, 2006). Courts generally refuse to enforce no-renegotiation clauses, viewing them as violations of the freedom of contract principle.[1] This refusal prevents direct contractual solutions to the renegotiation problem, requiring economic solutions that work within existing legal frameworks.

We provide a novel solution within the standard framework of mechanism design: *self-revealing* mechanisms, that rely on two features. First, bidirectional communication: a mechanism determines final allocations through the reports it receives from the agent and the signals it sends back to her. Second, conditional disclosure: communication within a mechanism remains private during the interaction but becomes verifiable at the contract execution stage. We show that a simple communication architecture, with only a binary message and a coin flip, fully eliminates the threat of renegotiation without resorting to any external commitment device, such as a trusted third party. This result overturns the conventional wisdom that the inability to contractually prevent renegotiation constrains incentive provision in contracting.

We develop our analysis in Fudenberg and Tirole (1990)'s canonical framework: a moral-hazard principal-agent relationship in which ex-ante optimality under full commitment entails imperfect risk-sharing. The corresponding second-best allocation is threatened by the principal's opportunity to renegotiate the original mechanism once the agent's effort is sunk. In this context, we construct a self-revealing mechanism that *uniquely* implements the second-best allocation: the threat of renegotiation is fully mitigated by design. Thus, renegotiation-proofness and second-best efficiency are simultaneously attainable.

While Fudenberg and Tirole (1990) characterize optimal renegotiation-proof arrangements by restricting attention to revelation mechanisms, in which the agent only reports

[1]For instance, the US Code on contract law under Title 42,§1981 declares the right of all persons to "the making, performance, modification, and termination of contracts". Jolls (1997) and Davis (2006) cite multiple applications of this code voiding contractual clauses limiting collective renegotiation. A notable example is Beatty v. Guggenheim Exploration Co. 225 N.Y. 380, 1919, where in his judgment Justice Cardozo voided an explicit contractual clause forbidding future modification stating that "Those who make a contract, may unmake it. The clause which forbids a change, may be changed like any other."

her private information, we consider a richer class of communication mechanisms that may both receive private reports from the agent and send her private signals. Specifically, we introduce self-revealing mechanisms, which consist of a *communication protocol* governing the bidirectional communication with the agent and a *decision rule* mapping communication histories to final transfers.

We construct a mechanism in this class that exploits the privacy of the agent's communication to deter renegotiation. This approach draws on the multistage communication mechanisms of Myerson (1986) and Forges (1986) but repurposes signals: rather than correlating players' behaviors, our mechanism uses them to generate off-path punishments that render renegotiation unprofitable.

The mechanism's structure is straightforward. After observing a renegotiation offer, the agent privately submits one of two reports: *status quo* or *renegotiation*. The mechanism then privately reveals to the agent the outcome of a fair coin toss, committing to the following payment rule: (i) if she reports *status quo*, it executes the second-best transfers; (ii) if she reports *renegotiation*, it implements one of two contracts, depending on the coin toss.

Intuitively, reporting *renegotiation* enables the agent to trigger, within the original mechanism, a random fallback lottery whose outcome she privately observes. She then accepts the principal's renegotiation offer only when the lottery outcome is unfavorable to her, and rejects it otherwise — so the original mechanism is retained precisely when its terms are costly to the principal, and replaced when they are not. This selective rejection makes any renegotiation attempt prohibitively costly to the principal in expectation, and the mechanism thereby implements the second-best allocation.

The randomness of the fallback lottery is key: with positive probability, the lottery yields the agent strictly more than the second-best. Anticipating this, she optimally activates the punishment against the renegotiating principal. Privacy of communication ensures that the renegotiation offer cannot be made contingent on the lottery's realization.

We formally establish the existence of such a self-revealing mechanism in Proposition 1. While offering a straightforward way to implement the full-commitment outcome under renegotiation, the result does not address equilibrium uniqueness. Indeed, because our optimal mechanism leaves the agent indifferent over efforts and reports, the subgame induced by the mechanism supports a continuum of equilibrium allocations. We achieve unique implementation in Proposition 2: the second-best allocation is the *unique* equilibrium outcome of the overall game in which the principal selects the mechanism at the initial stage. The result exploits a perturbation argument: we construct a perturbed mechanism that breaks all the agent's indifferences while leaving the punishment logic against renegotiation intact. The principal can then secure a utility arbitrarily close to

the second-best one under any belief about the agent's effort.

The construction underlying Propositions 1 and 2 requires no third party: the agent activates the punishment by reporting the occurrence of a renegotiation offer within the original mechanism. Bidirectional communication is crucial to this construction: the private signal sent back to the agent generates the asymmetric information that makes her selective rejection of renegotiation offers sequentially rational, rendering such offers unprofitable for the principal. Yet, bidirectional communication alone does not guarantee the enforceability of our mechanism: courts must be able to verify the contract's execution without relying on any external commitment device. To ensure this, the mechanism must also resolve a *verifiability paradox*. The mechanism's communication cannot be publicly verifiable when renegotiation is proposed—otherwise the principal could condition his renegotiation offer on it, undermining the punishment. Yet, this communication must become verifiable for enforcement when the original mechanism executes. To resolve this paradox, we exploit the second feature of self-revealing mechanisms, *conditional disclosure*. Bidirectional communication remains private during renegotiation, but becomes verifiably disclosed if and only if the original mechanism executes. On-path, where no renegotiation occurs, the disclosure has no strategic effect. Off-path, where renegotiation offers are made, privacy prevents the principal from conditioning on the agent's communication, keeping the punishment effective. A subtle question concerns off-path privacy: how can the agent's message remain hidden from the principal during renegotiation without relying on a third party? We show in Section 4 that a smart contract with a standard commit-and-reveal cryptographic scheme can deliver both message privacy and its conditional disclosure. This implements the off-path deterrent without any trusted third party.

Finally, while our mechanism exploits a specific timing of communication (the agent makes a report in the original mechanism *before* accepting a renegotiation offer), this timing requires no external enforcement: if the agent is free to choose when to communicate, her self-interest ensures she reports at the prescribed stage.

Our approach fits squarely within the mechanism design tradition: like Fudenberg and Tirole (1990), we take as given the sequence of contractible events and let the principal design a communication mechanism at the outset. The design is constrained by the legal doctrine of renegotiation: transfers cannot be conditioned on the occurrence or content of any future renegotiation offer, and any modification of the original mechanism requires the mutual consent of both parties.

In line with Myerson (1986), Forges (1986), and Bester and Strausz (2007), our optimal mechanism commits to a signal-sending protocol, but with a novel role: signals are exploited to deter renegotiation off the equilibrium path. While the principal may be tempted to circumvent the protocol at the renegotiation stage, the mutual consent re-

quirement rules this out: modifying the protocol is itself a contract modification requiring the agent's agreement, which her self-interest ensures she withholds. Restricting attention to standard revelation mechanisms at the renegotiation stage is therefore without loss, as in Fudenberg and Tirole (1990).

The power of self-revealing mechanisms is not confined to the Fudenberg and Tirole (1990) framework. To substantiate this claim, we first consider alternative extensive forms for the renegotiation game, extending our result to infinite rounds of renegotiation. We develop an infinite-horizon setting in the spirit of Strulovici (2017): renegotiation breaks down with positive probability at each round, in which case the last accepted contract executes. In our moral hazard context, second-best efficiency involves imperfect risk sharing, which leaves room for renegotiation after each round. We construct a self-revealing mechanism, offered at the ex-ante stage, that implements the second-best allocation, showing that backward-induction reasoning is inessential to our approach.

Second, we examine alternative specifications of the renegotiation process. In particular, we consider situations in which renegotiation complements rather than replaces the original mechanism. This generates a richer set of renegotiation opportunities than the replacement view: a new offer may exploit the observability of the original mechanism's transfers to undo any potential punishment. We show, however, that a modified version of our mechanism prevents this exploitation and implements the second-best allocation even under this supplementary view of renegotiation.

Mechanism design is thus flexible enough to accommodate legal constraints on renegotiation: by exploiting bidirectional communication and strategically timed information disclosure, contracting parties can achieve full-commitment outcomes within existing contract law, without requiring courts to enforce no-renegotiation clauses.

Related literature. The renegotiation literature, starting with Dewatripont (1989), shares a common principle identified by Bolton (1990): optimal renegotiation-proof mechanisms require private information for the agent at the renegotiation stage. Under moral hazard, Fudenberg and Tirole (1990) generate this private information through an equilibrium randomization over the agent's efforts; under incomplete information, Hart and Tirole (1988), Laffont and Tirole (1990) rely on the agent's exogenous private type, which the principal learns only gradually. In either case, maintaining this private information imposes allocative costs that prevent second-best efficiency. Maestri (2017) and Strulovici (2017) extend the analysis to infinite-horizon settings but share the same fundamental constraint. Our self-revealing mechanism circumvents this constraint: private information at the renegotiation stage is generated by signals sent to the agent rather than by costly randomization, eliminating the allocative cost and achieving full second-best efficiency.

Formally, our approach draws on the tradition of mechanism design for multistage games initiated by Myerson (1986) and Forges (1986). We complement their framework by explicitly considering an extensive-form game in which the principal has commitment power. This allows us to exploit the signals privately sent by the mechanism to target a new objective: generating off-equilibrium punishments rather than correlating players' strategies.

Hart and Moore (1999) and Maskin and Tirole (1999) analyze whether the parties can commit not to renegotiate by signing a clause under which any deviation from the original agreement triggers a large monetary transfer to a third party, making renegotiation jointly unprofitable. This approach, however, is vulnerable to multilateral renegotiation: if the third party can be brought to the table, all parties can jointly agree to waive the penalty, undoing the commitment device. Qi et al. (2024) revisit this approach in the Coase-conjecture framework and construct a mechanism that virtually implements the full-commitment allocation by secretly appointing a third party to collect a penalty if renegotiation takes place. This requires introducing a third party who can monitor renegotiation and credibly report it to a court. By contrast, we fully mitigate the renegotiation threat without any third party, achieving exact rather than virtual implementation.

Rahman and Obara (2010) achieve virtual implementation in a full-commitment setting through a disinterested mediator who makes confidential recommendations to agents and reveals them publicly at the payment stage. They do not address how to implement this conditional disclosure without a trusted third party. We achieve full implementation under renegotiation and explicitly construct an enforcement mechanism that, through self-revelation, requires no mediator.

Bester and Strausz (2007) develop the general idea that, in the absence of full commitment, mechanisms featuring private communication with an agent may have a welfare-enhancing role. The subsequent literature has mainly focused on the class of pure limited-commitment settings, in which contracts can be unilaterally voided by the principal. In this context, Doval and Skreta (2022) and Lomys and Yamashita (2022) establish different versions of a revelation principle under noisy communication. Recent works by Brzustowski et al. (2023) and Doval and Skreta (2024) focus on the Coase-conjecture environment and characterize optimal allocations under different contracting assumptions (long-term vs short-term contracts). Yet, they typically do not achieve second-best efficiency. We analyze mechanism design under the threat of renegotiation, providing a new rationale for private communication. Key to our construction is a defining feature of renegotiation environments: until both parties agree on new terms, the agent retains access to the options available in the original mechanism. This enables mechanisms sending private signals to generate a new set of punishments and, ultimately, to achieve unique (Perfect

Bayesian equilibrium) implementation of the second-best allocation without introducing any noisy communication device that garbles messages.

Renegotiation can also be interpreted as competition taking place between the principal at the ex-ante stage and his future self at the renegotiation stage to trade with the agent. This suggests a close relationship with common agency games, which analyze the competition among several principals who post mechanisms to deal with a common agent. In line with common agency, we let a mechanism delegate the implementation of any punishment – against renegotiation – to the agent. In our construction, such punishments correspond to (random) options that are offered but not activated by the agent on the equilibrium path. They hence serve the same role of the *latent* contracts, which are used to deter principals' deviation in common agency.[2]

Finally, our work contributes to literature on implementing mechanisms through smart contracts (Townsend, 2020, Chapter 6; Akbarpour and Li, 2020; Roughgarden, 2021). Brzustowski et al. (2023) appeal to smart contracts for implementing mechanisms that receive private messages without sending signals. We extend this idea by explicitly showing that smart contracts can also implement the reverse: mechanisms sending private signals to agents. This extension is crucial for demonstrating how current technologies enable full implementation of self-revealing mechanisms without mediators or third parties. This eliminates potential manipulation risks and achieves practical feasibility, bridging our theoretical innovation with real-world applicability.

The paper proceeds as follows. Section 2 presents the Fudenberg-Tirole framework. Section 3 constructs the self-revealing mechanism and establishes unique implementation of the second-best allocation. Section 4 addresses enforcement requirements and demonstrates practical implementation through smart contracts. Section 5 extends the analysis to other contracting environments. Section 6 concludes. The main proofs are in Appendix 7. Appendix 8 develops the smart contracts implementation and Appendix 9 collects additional results, including the extension to CRRA preferences, the case of a self-enforced communication protocol, and that of supplementary renegotiation.

2 The Benchmark

We consider the canonical framework of Fudenberg and Tirole (1990) (FT, henceforth), in which a risk-neutral principal (he) contracts with a risk-averse agent (she), who chooses an unobservable effort. There are two outputs $\omega \in \{g, b\}$, a good one g and a bad one b,

[2]See Bisin and Guaitoli (2004); Attar and Chassagnon (2009); Attar et al. (2011); Attar et al. (2019). Attar et al. (2025) analyze bidirectional communication in games with multiple principals and multiple agents.

where $g > b > 0$. The probability distribution over outputs depends on the binary effort $e \in E \equiv \{H, L\}$. Let $p_e \equiv \mathbb{P}(g|e)$ represent the probability of the good output given effort $e \in E$ with $p_H > p_L$ so that $\Delta p \equiv p_H - p_L > 0$. The effort e yields expected output $Y_e \equiv p_e g + (1 - p_e)b$.

Preferences and Allocations. The agent's utility is additively separable in income $w \in \mathbb{R}$ and effort $e \in E$, expressed as $u(w) - D(e)$. The function u exhibits $u'(w) > 0$ and $u''(w) < 0$ for each $w \in \mathbb{R}$, and is unbounded over its domain, i.e., $\lim_{w \to -\infty} u(w) = -\infty$ and $\lim_{w \to \infty} u(w) = \infty$. Consequently, the inverse $\Phi = u^{-1}$ is well-defined on the range of u, strictly increasing, $\Phi'(u) > 0$, and strictly convex, $\Phi''(u) > 0$. The low effort cost is normalized to $D(e = L) = 0$ and the high effort cost is $D(e = H) = d > 0$.[3]

Final payoffs are determined by the output-contingent transfers that the principal makes to the agent. A *contract* is a pair $(w_g, w_b) \in \mathbb{R}^2$ of such transfers. For notational convenience, we also write a contract as $c = (u_g, u_b)$, with $u_g = u(w_g)$ and $u_b = u(w_b)$. A (deterministic) *allocation* is a pair $(e, c) \in E \times \mathbb{R}^2$ of payoff-relevant decisions.

The agent's expected utility from (e, c) is

$$U_e(c) = p_e u_g + (1 - p_e)u_b - D(e),$$

where U^0 is her reservation utility.[4] The principal's expected utility from (e, c) is

$$V_e(c) = Y_e - p_e \Phi(u_g) - (1 - p_e)\Phi(u_b).$$

Efficient and Incentive-Compatible Allocations. Because the agent is risk-averse, while the principal is risk-neutral, efficient risk-sharing between the parties requires full insurance. For any $e \in E$, let $c_e^{FI}(U) \equiv (U + D(e), U + D(e))$ denote the full-insurance contract that yields the agent the expected utility $U \in \mathbb{R}$. We also define, for each $e \in E$, the function $V_e^{FI} : \mathbb{R} \to \mathbb{R}$ where

$$V_e^{FI}(U) \equiv V_e(c_e^{FI}(U)) = Y_e - \Phi(U + D(e))$$

identifies the principal's utility associated to the full-insurance contract leaving an expected utility U to the agent. Since $\Phi' > 0$, V_e^{FI} is strictly decreasing in U for any $e \in E$.

If effort is observable, the principal's optimal contract induces efficient risk-sharing while guaranteeing the agent her reservation utility U^0. We hence refer to $c^{FB} \equiv c_H^{FI}(U^0)$

[3] These assumptions follow FT directly. In Appendix 9, we discuss how FT's unboundedness assumption can be relaxed to accommodate CRRA utility with bounded transfers.

[4] In FT, it holds $U^0 = 0$. Writing the outside option as U^0 is more insightful for interpreting results.

as the first-best contract. The first-best allocation (H, c^{FB}) yields $V^{FB} \equiv V_H^{FI}(U^0)$ to the principal, and $U = U^0$ to the agent.[5]

If, instead, effort is unobservable, any feasible allocation must be incentive-compatible. Then, the optimal contract for the principal, which we denote the second-best contract, is the unique solution of:

$$\begin{aligned} \arg\max_{c \in \mathbb{R}^2} \quad & V_H(c) = p_H(g - \Phi(u_g)) + (1 - p_H)(b - \Phi(u_b)) \\ \text{s.t.} \quad & p_H u_g + (1 - p_H)u_b - d \geq p_L u_g + (1 - p_L)u_b \qquad \text{(IC)} \\ & p_H u_g + (1 - p_H)u_b - d \geq U^0. \qquad \text{(PC)} \end{aligned}$$

At the solution, the agent's incentive constraint (IC) binds. Accordingly, let $c^{IC}(U) \equiv (u_g^{IC}(U), u_b^{IC}(U))$ denote the contract leaving expected utility U to the agent, while satisfying the incentive constraint (IC) with equality:

$$u_g^{IC}(U) \equiv U + \frac{1 - p_L}{\Delta p} d \quad \text{and} \quad u_b^{IC}(U) \equiv U - \frac{p_L}{\Delta p} d. \tag{1}$$

Hence, $u_g^{IC}(U) > u_b^{IC}(U)$ for all $U \in \mathbb{R}$. It is convenient to define, for each $e \in E$, the function $V_e^{IC} : \mathbb{R} \to \mathbb{R}$, which denotes the principal's utility from the allocation $(e, c^{IC}(U))$:

$$V_e^{IC}(U) \equiv V_e(c^{IC}(U)) = Y_e - p_e \Phi\left(U + \frac{1 - p_L}{\Delta p} d\right) - (1 - p_e)\Phi\left(U - \frac{p_L}{\Delta p} d\right).$$

Since V_H^{IC} is decreasing in U, the agent's participation constraint (PC) also binds at the solution. We follow FT in assuming that $V_H^{IC}(U^0) > V_L^{FI}(U^0)$, so that inducing $e = H$ is optimal in the second-best. The second-best contract is $c^{SB} \equiv c^{IC}(U^0)$, and the second-best, ex-ante efficient, allocation (H, c^{SB}) yields $V^{SB} \equiv V_H^{IC}(U^0)$ to the principal, and $U_H(c^{SB}) = U^0$ to the agent.

The Renegotiation Threat. Any contract agreed upon ex-ante can be renegotiated at the *interim* stage, i.e., after effort is chosen but before output is realized. Following established tradition, the impact of this renegotiation threat is assessed in a non-cooperative game between the principal at the contract design stage, his future self at the interim stage and the agent. The timing of this game is as follows:

(i) The principal publicly offers a contract $c \in \mathbb{R}^2$.

(ii) The agent publicly accepts or rejects c. If she rejects, the game ends and the outside options accrue. If she accepts, the game continues as follows:

[5] Because we follow FT in focusing on the non-trivial case that $e = H$ is optimal in the second-best, we have that $e = H$ is also optimal in the first-best.

(*iii*) The agent privately chooses $e \in E$.

(*iv*) Without observing e, the principal makes a public renegotiation offer $c^r \in \mathbb{R}^2 \cup \{\emptyset\}$, where $\emptyset$ represents the principal's decision not to renegotiate.

(*v*) If $c^r \neq \emptyset$, the agent publicly accepts or rejects c^r by declaring $\rho \in \{y, n\}$.

(*vi*) Nature publicly draws the output realization g or b. If $c^r = \emptyset$, or $\rho = n$, transfers are determined by c. If $\rho = y$, transfers are determined by the renegotiated c^r.

Stages $(i) - (vi)$ define the *primitive* game G, which formalizes the sequence of events unfolding under renegotiation. The following assumptions reflect the relevant institutional environment:

A.1. (*Anti-renegotiation clauses are not enforceable*). The contract c *cannot* condition on the renegotiation offer c^r or on the agent's decision ρ at stage (v).

A.2. (*Renegotiation requires mutual consent*). If the original offer c is accepted by the agent at stage (ii), it cannot be unilaterally voided. If $c^r = \emptyset$ or $\rho = n$, both parties remain bound to the original contract c. This feature distinguishes renegotiation settings from pure limited-commitment ones.

A.3. (*Renegotiation as a replacement of contracts*). If $c^r \neq \emptyset$, the agent accepts or rejects it by declaring $\rho \in \{y, n\}$ at stage (v). Contract c^r then replaces contract c if and only if $\rho = y$, in which case, c becomes irrelevant. By contrast, if c^r is rejected at stage (v), i.e. $\rho = n$, c^r becomes irrelevant. Contracts c and $c^r \neq \emptyset$ are thus exclusive: at most one executes at stage (vi).[6]

The FT's Renegotiation Game. FT show that, for any probability $x \in (0, 1)$ with which the agent selects $e = H$ at stage (iii), the renegotiation stage (iv) corresponds to the setting of Stiglitz (1977): a monopolistic insurer facing a privately informed consumer. Hence, following Stiglitz (1977) and appealing to the revelation principle, FT let the principal offer revelation mechanisms $\gamma_C : E \to \mathbb{R}^2$, which map each effort report to a contract.

Denoting by C the set of all revelation mechanisms, FT thus modify the primitive game G into a renegotiation game G_C that allows the principal to design revelation mechanisms to deal with the agent's private information and the renegotiation threat. The modified game G_C is as follows. First, the principal offers a revelation mechanism $\gamma_C \in C$ at

[6]This "replacement" view of renegotiation is commonly adopted in the renegotiation literature (Bolton, 1990). In Section 5.2 we discuss the alternative "supplementary" view of renegotiation.

stage (i) and may renegotiate to $\gamma_C^r \in C$ at stage (iv). Second, the agent, after taking her participation decision at stage (v), sends a message $m \in E$ in the mechanism she participates in.

In G_C, any mechanism γ_C accepted by the agent at stage (ii) yields a subgame $G_C(\gamma_C)$ starting at stage (iii). In any such subgame, choosing $x = 1$ is not part of a Perfect Bayesian equilibrium. To see this, suppose the agent takes $e = H$ with probability one. Then, the principal's best reply is to offer the full-insurance contract $c_H^{FI}(U^0)$ in stage (iv) that is accepted by the agent. But against this renegotiation offer, the agent would be strictly better off choosing $e = L$.

Restricting attention to revelation mechanisms, FT show that the unique equilibrium allocation of G_C implements $e = H$ with probability $x^{FT} < 1$: the renegotiation threat constrains incentive provision. The characterization exploits the renegotiation-proofness principle: any equilibrium allocation can be replicated by one in which the mechanism offered at stage (i) is not renegotiated on path.

FT further show that revelation mechanisms at stage (i) yield no gain over simple contracts: the same allocation obtains irrespective of whether the principal offers a mechanism $\gamma_C \in C$ or a single contract c, with renegotiation taking place on the equilibrium path in the latter case. ~~(with on-path in the latter case)~~.[7] This suggests that mechanism design cannot resolve the conflict between ex-ante and interim efficiency. By contrast, we show that designing mechanisms with *bidirectional* communication—where the mechanism both receives messages and sends signals—fully eliminates the renegotiation threat, thereby reconciling renegotiation-proofness with second-best efficiency.

3 Self-Revealing Mechanisms and Renegotiation

In this section, we construct a simple mechanism that uniquely implements the second-best allocation, fully mitigating the renegotiation threat. Like FT, we let the principal design mechanisms within the event sequence $(i) - (vi)$ of the primitive game G. Unlike FT, we exploit the dynamic nature of G by introducing bidirectional communication, following Myerson (1986) and Forges (1986).

Separating the design of communication from the determination of final transfers, we write a mechanism $(\mathcal{C}, \tau)$ as a pair: a *communication protocol* $\mathcal{C}$ that specifies the (possibly bidirectional) communication exchanged at each stage and a *decision rule* τ that maps the communication history to final transfers. The mechanism $(\mathcal{C}, \tau)$ is publicly observed at stage (i), while all messages and signals exchanged through $\mathcal{C}$ remain *private* during the communication phase. However, we design our mechanisms to publicly reveal the full

[7] See Section 5.B in Fudenberg and Tirole (1990).

communication history *at the final payout stage*. This defines the *self-revealing* property of a mechanism. Any mechanism $(\mathcal{C}, \tau)$ satisfying the self-revealing property is called self-revealing. Thus, a self-revealing mechanism conditions the public revelation of its (possibly multistage) communication on its execution.

In our framework, these mechanisms serve two purposes. First, they generate private information during the game that enables off-equilibrium punishments. Second, by publicly revealing communication at execution, they ensure that conditional transfers are enforceable in standard contract-theoretic terms and do not require any third-party mediation.[8] Importantly, because execution halts upon renegotiation, self-revelation occurs only if the original mechanism is retained.

Rather than considering all possible self-revealing mechanisms, we focus on a simple class that suffices for achieving unique implementation of the second-best. Specifically, we fix a communication protocol with the following features:

1. Bidirectional communication occurs only with the agent and only at the beginning of stage (v).

2. At stage (v), the agent sends a message m from a message set $\mathcal{M} = \{N, R\}$ where N indicates "no renegotiation proposed" and R indicates "renegotiation proposed".

3. At stage (v), the agent also receives a signal s from the signal set $\mathcal{S} = \{h, t\}$ representing a fair coin toss, with probability ~~with~~ $\sigma(h|m) = \sigma(t|m) = 1/2$ for each $m \in \mathcal{M}$.

4. The agent sends m and receives s *before* her participation decision ρ.[9]

We denote such a protocol by $\mathcal{C} = (\mathcal{M}, \mathcal{S}, \sigma)$. The corresponding decision rule $\tau : \mathcal{M} \times \mathcal{S} \to \mathbb{R}^2$ maps each (m, s) pair to a contract $c = (u_g, u_b)$. We denote the set of all such mechanisms by Γ.

In the remainder of this section, we let the principal design mechanisms in the class Γ under the threat of renegotiation. The design problem is structurally simple: only four transfer pairs $\{\tau(N,h), \tau(N,t), \tau(R,h), \tau(R,t)\}$ require specification. We next formalize the induced renegotiation game G_Γ.

[8] Section 4.3 explicitly discusses how, with the use of cryptographic tools, existing "smart contracting" technologies provide a concrete way to implement self-revelation ~~without the need for any third-party mediation~~.

[9] Given that signal s does not condition on message m, the sequential structure of first sending m and then receiving s is strategically equivalent to the message and the signal being exchanged simultaneously.

3.1 The Self-Revealing Renegotiation Game G_Γ

Allowing the principal to choose self-revealing mechanisms from the set Γ modifies the primitive game G into the extensive-form game G_Γ as follows:

(i) The principal publicly offers a *self-revealing mechanism* $\gamma \in \Gamma$. That is, he chooses the four transfer pairs that determine the decision rule $\tau : \mathcal{M} \times \mathcal{S} \to \mathbb{R}^2$.

(ii) The agent publicly accepts or rejects γ. If she rejects, the game ends and the outside options accrue. If she accepts, the game continues as follows:

(iii) The agent privately chooses $e \in E$.

(iv) Without observing e, the principal makes a public renegotiation offer $\gamma^r \in C \cup \{\emptyset\}$, where $\emptyset$ represents the principal's decision not to renegotiate.

(v) The agent sends a private message $m \in \mathcal{M} = \{N, R\}$ and receives a private random signal $s \in \{h, t\}$. If $\gamma^r \neq \emptyset$, the agent publicly accepts or rejects γ^r by declaring $\rho \in \{y, n\}$.

(vi) If $\gamma^r \neq \emptyset$ and $\rho = y$, the agent sends a private message $m^r \in E$ in γ^r. Nature publicly draws the output realization g or b and transfers are determined by $\gamma^r(m^r)$. If, instead, either $\gamma^r = \emptyset$ or $\rho = n$, then γ executes. Nature publicly draws the output realization g or b, the communication (m, s) from stage (v) is publicly revealed, and transfers are determined by $\tau(m, s)$.

In G_Γ, the principal selects a self-revealing mechanism $\gamma \in \Gamma$ at stage (i) but is restricted to revelation mechanisms $\gamma^r \in C$ at the renegotiation stage (iv). As in FT's analysis, this restriction is without loss of generality. To see this, note that offering a self-revealing mechanism γ at stage (i) constrains the design of a renegotiation offer γ^r at stage (iv) as follows.

No offer γ^r can condition on the private communication (m, s) within γ. Thus, the renegotiation stage becomes a mechanism design problem where the principal faces an agent with private information (e, m, s). If the agent accepts γ^r, the original mechanism γ does not execute, so (m, s) are never publicly revealed.[10] In addition, her preferences over the transfers to implement in γ^r depend solely on e. Indeed, if the agent accepts γ^r, the transfers in γ are payoff-irrelevant, which guarantees that an agent with a given e but different (m, s) has the same set of optimal reports in γ^r, regardless of the message space of γ^r.

[10] That is, the new offer *replaces* the original one, as assumed in *A*.3.

At the renegotiation stage, the principal therefore cannot screen on (m, s) through the design of γ^r. Different realizations of (m, s) do generate different outside options for the agent in the original mechanism γ, but this heterogeneity only affects *whether* she accepts γ^r, not her behavior within it. Then, standard monopolistic-screening arguments ~~guarantee that any decision rule offered at the renegotiation stage can be reproduced by restricting to revelation mechanisms in the class C.~~ imply that any allocation supported by an indirect mechanism at the renegotiation stage is also supported by a direct revelation mechanism in the class C in which the agent reports her effort.[11]

This conclusion concerns the message space at the renegotiation stage, not the communication protocol. However, since γ^r screens only on e, the revelation principle for multistage games (Myerson, 1986; Forges, 1986) guarantees that any richer protocol, involving multi-stage communication or signal-sending, provides no additional screening power beyond what revelation mechanisms already achieve.

When attempting to renegotiate, the principal may also be tempted to alter the communication architecture of the original mechanism γ, by acting, for instance, on its signal-disclosure rule. This, however, requires the agent's consent (Assumption A.2), which her self-interest ensures she withholds, as we discuss in Sections 4.1 and 4.2.

Together, these two arguments imply that any feasible renegotiation opportunity is represented, with no loss of generality, in the game G_Γ.

A (pure) strategy of the principal in G_Γ consists of a mechanism $\gamma \in \Gamma$ followed by a renegotiation offer $\gamma^r \in C \cup \{\emptyset\}$ for any $\gamma \in \Gamma$. An agent's (behavioral) strategy λ in G_Γ has four components. First, for any $\gamma \in \Gamma$, a probability $x \in [0, 1]$ of choosing $e = H$. Second, for any history (γ, e, γ^r), a probability distribution over messages $m \in \mathcal{M}$. Third, for any $(\gamma, e, \gamma^r, m, s)$ with $\gamma^r \neq \emptyset$, a participation choice $\rho \in \{y, n\}$. Fourth, for any subsequent history with $\rho = y$, a message $m^r \in E$ in the renegotiated mechanism γ^r.

We consider the perfect Bayesian equilibria (henceforth equilibria) of G_Γ.[12] We denote $G_\Gamma(\gamma)$ the subgame induced by $\gamma \in \Gamma$ starting at stage (iii). In this game, $\lambda(\gamma)$ represents the agent's continuation strategy while the principal's strategy is a renegotiation offer $\gamma^r(\gamma) \in C \cup \{\emptyset\}$. Because $G_\Gamma(\gamma)$ is an extensive form game with imperfect information, any equilibrium of G_Γ must induce an equilibrium in each $G_\Gamma(\gamma)$. Therefore,

[11] The argument parallels Rochet and Stole (2002), who study nonlinear pricing in the presence of a buyer whose preferences over the seller's offers depend only on her (private) valuation, not on her (private) reservation utility. They show that, conditional on participation, restricting attention to (nonlinear) tariffs that screen the buyer's valuation but not her reservation utility is without loss of generality (Rochet and Stole, 2002, p. 282). In our setting, this structure is mirrored by the fact that the agent's preferences over renegotiation contracts depend only on e, not on (m, s).

[12] The principal has only one information set in the game G_Γ, where his belief $x \in [0, 1]$ is formulated on the probability that $e = H$. This is unambiguously pinned down in any equilibrium by the agent's equilibrium strategy. Thus, off-path belief-updating rules are irrelevant, and equilibrium refinements beyond PBE are superfluous for our analysis.

in an equilibrium of G_Γ, the principal chooses an optimal mechanism γ anticipating that the continuation play will constitute an equilibrium of $G_\Gamma(\gamma)$. We say that a mechanism $\gamma \in \Gamma$ is *renegotiation-proof* if the continuation game $G_\Gamma(\gamma)$ admits an equilibrium in which renegotiation does not occur, i.e. $\gamma^r(\gamma) = \emptyset$.

The game G_Γ differs from FT's game G_C only in the mechanisms available at stage (i). Both games share the same event sequence (i)-(vi) and renegotiation threat $\gamma^r \in C$. Mechanisms in Γ add bidirectional communication: next to the agent sending messages, she may also receive signals. Crucially, this bidirectional communication takes place *before* the agent's acceptance decision of a renegotiation offer.

3.2 Implementing the Second Best

We next show that self-revealing mechanisms fully mitigate the renegotiation threat. We proceed in two steps. In this subsection, we identify a specific renegotiation-proof mechanism $\gamma^* \in \Gamma$ that implements the second-best allocation. In the next subsection, we show that this allocation is the unique equilibrium outcome of G_Γ: in *any* equilibrium, the principal obtains V^{SB} and the agent picks $e = H$ with probability one.

We consider the self-revealing mechanism $\gamma^* \in \Gamma$ with the following decision rule:

$$\tau^*(N,h) = \tau^*(N,t) = c^{SB}; \quad \tau^*(R,h) = c^{IC}(U^0 - \Delta U); \quad \tau^*(R,t) = c^{IC}(U^0 + \Delta U).$$

The mechanism γ^* sets the second-best contract c^{SB} as the default, triggered by $m = N$. It also allows the agent to trigger a random counter-offer by sending $m = R$: the realized contract then either increases or decreases her utility by ΔU, each with probability $1/2$. The counter-offer preserves the agent's expected utility at U^0 for any $e \in E$, but costs the principal strictly more by Jensen's inequality, since Φ is convex. The principal views the counter-offer as random, whereas the agent privately observes its realization after sending $m = R$ but before deciding whether to accept the principal's renegotiation offer.

The next lemma establishes that a sufficiently large ΔU makes any renegotiation unprofitable for the principal.

Lemma 1 *There exists $\Delta U \in (0, \infty)$ such that for all $e \in E$:*

$$V_e^{IC}(U^0) > \max\left\{V_e^{FI}(U^0 + \Delta U),\ \frac{1}{2}V_e^{FI}(U^0 - \Delta U) + \frac{1}{2}V_e^{IC}(U^0 + \Delta U)\right\}. \tag{2}$$

The lemma states that, for any $e \in E$, the principal prefers the second-best contract, $c^{SB} = c^{IC}(U^0)$, to a full-insurance contract that leaves an extra utility of ΔU to the agent.[13] Additionally, the principal prefers c^{SB} to a 50-50 lottery between the full-insurance contract leaving ΔU less to the agent, and the incentive-compatible one leaving

[13]The proof of Lemma 1 constructs a $U^n > U^0$ such that any $\Delta U > U^n - U^0$ satisfies (2).

the agent an extra utility ΔU. This validates our construction: the principal attains the left-hand side of (2) when he does not renegotiate. The first term in the maximum bounds his utility from offers the agent always accepts; the second bounds his utility from offers accepted only when $s = h$.

The lemma allows us to establish the following result.

Proposition 1 *The second-best allocation* (H, c^{SB}) *is supported in an equilibrium of the subgame* $G_\Gamma(\gamma^*)$.

Proof. For any effort $e \in E$ and renegotiation offer $\gamma^r \in C$, let $\hat{m}_e^r \in E$ denote an agent's optimal report in γ^r and let $\hat{U}_e^r$ denote her corresponding utility upon acceptance:

$$\hat{m}_e^r \in \arg\max_{m^r \in E} U_e(\gamma^r(m^r)) \quad \text{and} \quad \hat{U}_e^r \equiv U_e(\gamma^r(\hat{m}_e^r)). \tag{3}$$

We now construct a strategy profile $\{\gamma^r(\gamma^*), \lambda(\gamma^*)\}$ that supports the second-best equilibrium allocation. The principal does not renegotiate, i.e. $\gamma^r(\gamma^*) = \emptyset$. The agent's strategy $\lambda(\gamma^*)$ is as follows:

1. The agent chooses $e = H$ with probability $x = 1$.

2. For any $e \in E$, her message $m \in \{N, R\}$ in γ^* depends on the principal's offer γ^r through $\hat{U}_e^r$:

   (*i*) If $\gamma^r = \emptyset$, the agent sends $m = N$ in γ^*.

   (*ii*) If $\gamma^r \neq \emptyset$ and $\hat{U}_e^r \leq U^0 - \Delta U$, the agent sends $m = N$ in γ^*.

   (*iii*) If $\gamma^r \neq \emptyset$ and $\hat{U}_e^r > U^0 - \Delta U$, the agent sends $m = R$ in γ^*.

3. For any $e \in E$, $\gamma^r \in C$, $m \in \{N, R\}$ and $s \in \{h, t\}$, her participation decisions are the following:

   (*i*) If $\hat{U}_e^r < U^0 - \Delta U$, the agent selects $\rho = n$ for any $(m, s) \in \{N, R\} \times \{h, t\}$.

   (*ii*) If $\hat{U}_e^r \in [U^0 - \Delta U, U^0)$, the agent selects: when $m = N$, $\rho = n$ for all $s \in \{h, t\}$; when $m = R$, $\rho = y$ if $s = h$ and $\rho = n$ if $s = t$.

   (*iii*) If $\hat{U}_e^r \in [U^0, U^0 + \Delta U)$, the agent selects: when $m = N$, $\rho = y$ for all $s \in \{h, t\}$; when $m = R$, $\rho = y$ if $s = h$ and $\rho = n$ if $s = t$.

   (*iv*) If $\hat{U}_e^r \geq U^0 + \Delta U$, the agent selects $\rho = y$ for any $(m, s) \in \{N, R\} \times \{h, t\}$.

4. For any $e \in E$, $\gamma^r \in C$, $m \in \{N, R\}$ and $s \in \{h, t\}$, after $\rho = y$, the agent sends $\hat{m}_e^r$ to γ^r as defined in (3).

To summarize, the agent reports $m = R$ in γ^* in response to any renegotiation offer yielding her more than $U^0 - \Delta U$, and takes her subsequent participation decisions by comparing her continuation utility under the two mechanisms.

We show that the strategy profile $\{\gamma^r(\gamma^*), \lambda(\gamma^*)\}$, together with the principal's belief that the agent picked $e = H$ with probability $x = 1$, constitutes an equilibrium of $G_\Gamma(\gamma^*)$, supporting the second-best allocation (H, c^{SB}).

Note first that the only non-trivial information set of the principal in $G_\Gamma(\gamma^*)$ is at the renegotiation stage, when he offers γ^r. The only belief consistent with the strategies $\{\gamma^r(\gamma^*), \lambda(\gamma^*)\}$ is, indeed, $x = 1$, as $\lambda(\gamma^*)$ prescribes $e = H$ for the agent.

We verify the sequential rationality of $\{\gamma^r(\gamma^*), \lambda(\gamma^*)\}$ given $x = 1$, in two lemmas whose proofs appear in Appendix 7:

Lemma 2 *In the subgame $G_\Gamma(\gamma^*)$, the agent's strategy $\lambda(\gamma^*)$ is sequentially rational given the principal's strategy $\gamma^r(\gamma^*) = \emptyset$.*

The proof of Lemma 2 establishes that it is optimal for the agent to report $m = R$ in γ^* after an offer γ^r with $\hat{U}_e^r \in (U^0 - \Delta U, U^0 + \Delta U)$. In so doing, she obtains $U^0 + \Delta U$ when $s = t$ by rejecting, and $\hat{U}_e^r$ (which exceeds $U^0 - \Delta U$) when $s = h$ by accepting. This exceeds in expected terms the utility associated with the report $m = N$. If, instead, $\hat{U}_e^r \geq U^0 + \Delta U$, the agent is indifferent over reports in γ^* as she eventually accepts γ^r under any communication profile; thus $m = R$ is therefore (weakly) optimal. Finally, if $\gamma^r = \emptyset$ or $\hat{U}_e^r \leq U^0 - \Delta U$, the agent's continuation utility from γ^* equals her reservation utility U^0 regardless of her communication, ~~$\hat{U}_e^r < U^0 - \Delta U$, the agent eventually rejects the offer. This yields her the reservation utility U^0 regardless of her communications in γ^*,~~ which guarantees that $m = N$ is optimal.

In addition, the proof of Lemma 2 establishes the optimality of $e = H$. In particular, since $\gamma^r(\gamma^*) = \emptyset$, the agent expects U^0 from either effort level.

The next lemma establishes the absence of profitable deviations for the principal.

Lemma 3 *In the subgame $G_\Gamma(\gamma^*)$, the principal's strategy $\gamma^r(\gamma^*) = \emptyset$ is sequentially rational given his (Bayes-consistent) belief $x = 1$, and the agent's strategy $\lambda(\gamma^*)$.*

The proof of Lemma 3 shows how the agent's equilibrium strategy implements an effective punishment against renegotiation. In particular, the utility $U^0 + \Delta U$ that she gets with probability 1/2 by reporting $m = R$ in γ^* makes any attempted renegotiation too costly to the principal.

The strategies $\{\lambda(\gamma^*), \gamma^r(\gamma^*)\}$ and the principal's belief $x = 1$ therefore constitute a perfect Bayesian equilibrium of $G_\Gamma(\gamma^*)$. In this equilibrium, the agent chooses $e = H$ with probability one and the contract c^{SB} is implemented, establishing Proposition 1. ■

~~Since the principal cannot obtain more in a game with renegotiation than under full commitment, and the agent receives her reservation utility U^0, Proposition 1 implies that the game G_Γ has *an* equilibrium in which the renegotiation threat is fully mitigated.~~ Proposition 1 shows that the subgame $G_\Gamma(\gamma^*)$ has a second-best efficient equilibrium in which no renegotiation takes place. The result contrasts with FT, who restrict the principal to offer revelation mechanisms.

In FT, renegotiation-proofness and second-best efficiency are incompatible: achieving one precludes the other. The mechanism γ^* reconciles this conflict by exploiting signals to generate off-equilibrium punishments that deter renegotiation. To see how, consider the full-insurance offer $c_H^{FI}(U^0)$ that undermines $x = 1$ in FT. Against γ^*, the agent reveals this offer by sending $m = R$, triggering the counter-offer lottery. She then rejects the principal's offer with probability $1/2$ (when $s = t$), retaining the favorable counter-offer terms. This random rejection makes renegotiation unprofitable for the principal: for any offer yielding the agent a utility in $(U^0 - \Delta U, U^0 + \Delta U]$, the expected cost of the counter-offer exceeds the gain from renegotiation.

Because this punishment hinges on the random signal s, it effectively implements a randomized contract: the agent's actual utility depends on the coin flip outcome. This raises a natural question: could a purely stochastic mechanism — one that directly assigns randomized contracts to any agent message — also implement the second-best allocation? The answer is no. To see why intuitively, note that standard insurance-theory arguments (Chade and Schlee, 2012) guarantee that the optimal renegotiation offer against any distribution over efforts chosen by the agent is deterministic. Thus, anticipating such deterministic renegotiation, the principal cannot benefit from committing ex ante to stochastic mechanisms.[14]

Indeed, a standard stochastic mechanism is random for both the principal and agent — neither party can condition their decisions on the randomness realization. By contrast, the contract implemented by γ^* conditional on receiving $m = R$ appears random only to the principal. The agent privately observes s and conditions her acceptance on it. For offers with $\hat{U}^r \in (U^0 - \Delta U, U^0 + \Delta U)$, acceptance occurs only when $s = h$, yielding the principal an expected utility of $V_H^{FI}(\hat{U}^r)/2 + V_H^{IC}(U^0 + \Delta U)/2$, which Lemma 1 ranks strictly below V^{SB}. ~~The agent's private observation of the signal is therefore essential for deterrence.~~

Starting with Bester and Strausz (2007), the idea that a principal may benefit by making his decision rule contingent on the realizations of some endogenous signal has been extensively employed in mechanism design without full commitment. Yet, the off-

[14]The result is formally established in Appendix 9. We show, in particular, that stochastic mechanisms do not play any strategic role in the FT construction either.

equilibrium role of signals we document crucially exploits the features of the renegotiation problem and cannot in general be reproduced under other forms of limited commitment. For instance, in settings where parties can unilaterally withdraw contracts, a new offer forces withdrawal of the original one (e.g., Doval and Skreta, 2022; Brzustowski et al., 2023). The agent cannot then communicate within the original mechanism or solicit counter-offers, narrowing the strategic role of signals. In this context, Doval and Skreta (2022) argue that signals only provide Bayes-plausible updates of the principal's beliefs about the agent's type rather than generating new private information as in our approach. Thus, our construction establishes an entirely novel application for endogenous information disclosures under commitment frictions.

To conclude, observe that, in line with FT, we have taken the agent's utility over monetary transfers u to be defined on the entire real line and unbounded. These features are key to establish Lemma 1, and, ultimately, to identify the relevant punishments against renegotiation, which we exploit in the Proof of Proposition 1. Although this approach greatly simplifies presentation, it does not allow us to consider a range of situations of economic relevance, most notably those in which the agent is subject to limited liability, and her monetary transfers are therefore bounded. To cope with this issue, we show in Appendix 9 that Proposition 1 extends to cases where the agent's utility is CRRA, accommodating limited liability constraints that bound transfers below.

3.3 Unique Implementation of the Second Best

Proposition 1 shows that the self-revealing mechanism γ^* induces a subgame supporting the second-best allocation at equilibrium. Because this outcome yields the agent the utility U^0, it is also incentive-compatible for her to accept γ^* at stage (ii), as she cannot strictly gain by rejecting it. Moreover, the principal cannot attain a utility greater than V^{SB} in the benchmark without renegotiation. Thus, the game G_Γ, which includes the principal's design of the mechanism $\gamma \in \Gamma$ at stage (i), admits an equilibrium yielding the second-best allocation.

Under the standard selection convention in mechanism design where the designer can target an equilibrium of the chosen mechanism, existence suffices for implementability. The stricter notion of unique implementation requires ruling out other equilibrium outcomes. Indeed, γ^* makes the agent indifferent over her messages as well as over her effort choices, implying that the subgame $G_\Gamma(\gamma^*)$ supports a continuum of equilibria. For instance, any $x \in [0,1]$ can be supported in an equilibrium of $G_\Gamma(\gamma^*)$ where the principal does not renegotiate and the agent reports $m = R$ following any off-path renegotiation offer. Yet, although $G_\Gamma(\gamma^*)$ admits multiple equilibrium allocations, the next proposition shows that *only the second-best one* is supported at equilibrium in the overall game G_Γ.

Proposition 2 *The game G_Γ has a unique equilibrium allocation, which coincides with the second-best one (H, c^{SB}).*

The proof of Proposition 2 in Appendix 7 constructs a mechanism γ_ε by perturbing γ^* in a way that allows us to break all the agent's indifferences at the root of equilibrium multiplicity. Specifically, γ_ε penalizes low effort while preserving incentives for high effort. Thus, in the subgame $G_\Gamma(\gamma_\varepsilon)$ the agent strictly prefers to choose $e = H$ with probability $x = 1$, and to report $m = N$ in γ_ε as long as no renegotiation attempt is made. As for the principal, for any belief $x \in [0, 1]$, choosing not to renegotiate turns out to be the unique best response to any sequentially rational behavior of the agent.

By offering a perturbed mechanism γ_ε at the initial stage, the principal can hence guarantee himself a utility arbitrarily close to V^{SB}, which obtains under full commitment. Since V^{SB} is also an upper bound, uniqueness of the equilibrium allocation follows.

Since γ^* operates after effort is chosen, Propositions 1 and 2 extend directly to any countable or continuous effort space, paralleling the extension in FT.[15]

4 A New Approach to Renegotiation Proofness

In this section, we examine two central properties of our optimal mechanism: its enforceability by external courts and the commitment power required to implement it.

4.1 The Self-Revealing Mechanism γ^*: Enforcement

The mechanism $\gamma^* \in \Gamma$ combines a communication protocol with a decision rule that maps a pair of messages and signals to transfers. Its enforceability relies on two features: conditional disclosure of private communication and delegation of punishments to the agent. Together, these ensure contractability—courts can verify compliance with contractual obligations.

The verifiability paradox. The mechanisms we consider face an apparent tension. On one hand, communication must remain private during renegotiation: if the principal observes the agent's message and signal, he can condition his renegotiation offer on them, undermining the punishment mechanism. On the other hand, communication must be verifiable at enforcement: courts need to verify that transfers match the contractually specified rule $\tau(m, s)$.

The mechanism γ^* resolves this *verifiability paradox* through its self-revealing property. Communication remains private throughout the renegotiation stage but becomes publicly

[15] See Fudenberg and Tirole (1990, Section 5.A).

revealed if and only if the original contract executes (i.e. if $\rho = n$ at stage (vi)). This conditional disclosure satisfies both requirements simultaneously.

First, privacy during renegotiation: if the principal attempts to renegotiate and the agent accepts ($\rho = y$), the original mechanism does not execute, so (m, s) are never revealed. Second, verifiability at enforcement: whenever the original mechanism executes ($\rho = n$)—either because no renegotiation was attempted or because the agent rejected it—the mechanism publicly reveals (m, s), allowing courts to verify that actual transfers correspond to $\tau(m, s)$ as contractually specified.

In our construction, the self-revealing property guarantees that all relevant communication becomes public on the equilibrium path, where renegotiation does not take place. When, instead, a renegotiation offer is made, the communication occurring within γ^* is made public at no strategic cost whenever the agent is led to reject the new offer. Thus, while we share with recent approaches to mechanism design under limited commitment the focus on private communication, our mechanism does not rely on noisy communication as in Bester and Strausz (2007) or direct belief recommendations as in Doval and Skreta (2022). The court can therefore directly enforce the equilibrium transfers without requiring any trusted third party to observe or execute them.

A self-enforcing communication protocol. The mechanism γ^* is structurally simple, featuring a binary message space $\{N, R\}$ and a binary signal $\{h, t\}$ generated by a fair coin toss. Its complexity is comparable to that of FT's revelation mechanisms, which induce the agent to randomize over binary reports $\{H, L\}$ but do not involve signals.

In FT, the stage at which the agent sends her message is immaterial, so the construction is compatible with several communication protocols.[16] Our construction, by contrast, exploits the agent communicating at stage (v), after observing a renegotiation offer but before accepting it. This raises a question: does our mechanism effectively require courts to verify adherence to this communication protocol?

The agent's self-interest ensures compliance with the protocol. Even if courts cannot verify when the agent communicates, she finds it optimal to report at stage (v): observing γ^r before sending her message allows her to condition on whether renegotiation was attempted. In Appendix 9, we formalize this intuition by constructing a protocol that delegates the timing of communication to the agent. In any pure strategy equilibrium of the subgame induced by such a generalized mechanism, she communicates *after* the principal's renegotiation offer and before her participation decision, i.e., exactly at stage (v).

This self-enforcing property has practical implications. Courts need only verify that executed transfers match $\tau(m, s)$ for the revealed (m, s), not *when* communication oc-

[16] See Fudenberg and Tirole (1990, p. 1283)

curred. The generalized mechanism operates *under the shadow of the court*: by delegating the timing choice to the agent, it aligns her strategic interests with the required protocol.

4.2 The Commitment Requirements in G_Γ

Our approach is rooted in the mechanism design tradition. Like FT, we take as given the sequence of events (i)–(vi), but we let the principal design a self-revealing mechanism $\gamma \in \Gamma$. The optimal mechanism γ^* makes its transfers and disclosure policy conditional on both the contractible variables in (i)–(vi) and the communication privately exchanged with the agent.

The mechanism γ^* incorporates at the outset a commitment to the signals sent to the agent, a feature it shares with standard approaches to mechanism design, both under full commitment (Myerson, 1982, 1986; Forges, 1986) and under limited commitment (Bester and Strausz, 2007). Unlike in these settings, however, we exploit signals off the equilibrium path to deter future renegotiation attempts, rather than to correlate players' strategies on path. This gives the principal an incentive to modify or suppress the signal-sending protocol at the renegotiation stage, since it is precisely this protocol that generates the punishments against renegotiation. The mutual consent Assumption A.2, however, rules this out: modifying the signal-sending protocol is itself a form of contract modification and therefore requires the agent's explicit agreement. Indeed, as discussed in the previous subsection, the agent's self-interest ensures she prefers to maintain the protocol. The protocol therefore remains intact throughout the renegotiation stage. In particular, the principal cannot circumvent the mechanism through "exploding offers" that demand immediate acceptance. Even if such offers were legally permissible, they cannot prevent the agent from communicating within γ^* before responding. Because renegotiation requires mutual consent, the principal cannot unilaterally revoke the agent's right to send m and receive s before deciding on any renegotiation offer. Moreover, because the agent's communication is private, the principal cannot condition his offer on the agent not having communicated.

The renegotiation game G_Γ also reveals the specific commitment assumptions we exploit at the renegotiation stage. As noted in Section 3, restricting attention to deterministic revelation mechanisms at the renegotiation stage involves no loss of generality. For a given self-revealing mechanism $\gamma \in \Gamma$ offered at stage (i), making a renegotiation offer that conditions on the content or occurrence of the agent's communication in γ is infeasible by construction.

Overall, in the game G_Γ, the design of an initial mechanism is fundamentally constrained by the legal doctrine of renegotiation. The mechanism cannot commit to the content or occurrence of a renegotiation offer, and can be modified only in the presence of

mutual consent. At the same time, when renegotiating, the principal operates under the same commitment assumptions as in FT. In particular, we do not allow for any trusted third party observing communication and executing transfers on behalf of the parties.

A natural concern is whether conditional disclosure constitutes a form of commitment that courts would refuse to enforce, analogous to the no-renegotiation clauses ruled out by Assumption A.1. It does not: a disclosure clause is part of the communication protocol, on which the parties can freely contract. As established in Section 4.1, the disclosure clause of γ^* triggers only when γ^* executes, which occurs precisely when renegotiation has not taken place. Conditional disclosure therefore places no constraint on the parties in any state where renegotiation occurs, and requires no legal enforcement beyond that of the output-contingent payments themselves.

The commit-and-reveal cryptographic scheme we describe in the next section renders conditional disclosure self-executing.

4.3 From Theory to Practice: Smart Contracts Implementation

Smart contracts—self-executing programs deployed on public blockchains—are a natural tool for implementing our optimal mechanism. They provide two features that align with our construction. First, once deployed, a smart contract commits to its programmed rules, including any randomization protocol; this delivers the commitment to signals that γ^* requires. Second, because all variables, inputs, and computations within a smart contract are publicly observable on the blockchain, conditional transfers become enforceable by standard means once the mechanism self-reveals its private information. The implementation challenge therefore centers entirely on the communication protocol.

The public nature of blockchains, however, implies that smart contracts cannot send private signals to players: any value known to the contract is accessible to all observers of the blockchain. This conflicts with the privacy requirements of our mechanism γ^*.[17]

We solve this issue by modifying γ^* to work with public signals, while keeping the agent's message private. The key idea is to give the agent multiple (private) message options that interact differently with the public coin flip, allowing her to effectively choose which version of randomness to face.

Formally, consider the modified mechanism $\gamma^{**} = (\mathcal{M}^{**}, \mathcal{S}^*, \sigma^{**}, \tau^{**})$ with three private messages $\mathcal{M}^{**} = \{N, R_1, R_2\}$, and, as in γ^*, a signal $s \in \mathcal{S}^*$ representing a fair coin toss:

[17]Note that if the signal s were public rather than private, the principal could make signal-conditional renegotiation offers that undermine the effectiveness of γ^* as follows: provide attractive terms only when $s = t$ but terrible terms when $s = h$. This would induce the agent to send message $m = N$ and accept renegotiation when $s = t$, allowing the principal to avoid the punishment mechanism and gain from renegotiation.

$\sigma^{**}(h|m) = \sigma^{**}(t|m) = 1/2$ for all $m \in \mathcal{M}^{**}$.[18] Unlike in γ^*, we now consider the signal s to be publicly observable. To address the challenge that a renegotiation offer can condition on its realization, we let the decision rule τ^{**} depend on the two messages R_1 and R_2 as follows

$$\begin{aligned}
&\tau^{**}(N,h) = \tau^{**}(N,t) = c^{IC}(U^0) = c^{SB};\\
&\tau^{**}(R_1,t) = c^{IC}\left(U^0 + \Delta U\right); \quad \tau^{**}(R_1,h) = c^{IC}(U^0 - \Delta U);\\
&\tau^{**}(R_2,t) = c^{IC}\left(U^0 - \Delta U\right); \quad \tau^{**}(R_2,h) = c^{IC}(U^0 + \Delta U).
\end{aligned}$$

Effectively, γ^{**} gives the agent a choice between two random counter-offers that differ only in which face of the coin flip yields the favorable contract. Thus γ^{**} requires only (i) privacy for a 3-symbol message and (ii) a public fair coin; it does not rely on contract-provided private randomness.

The modified mechanism γ^{**} still implements the second-best allocation.[19] When facing a renegotiation offer, the agent selects between the private messages R_1 and R_2, each creating a different lottery over favorable and unfavorable terms. Regardless of her choice, she faces a 50-50 chance of receiving highly favorable terms (utility $U^0 + \Delta U$) that make rejecting renegotiation optimal. This random rejection punishes the principal in expectation, deterring renegotiation just as in the original mechanism γ^*. For instance, in the intermediate region, the principal's expected utility under renegotiation equals

$$V_H^{FI}(\hat{U}^r)/2 + V_H^{IC}(U^0 + \Delta U)/2,$$

which remains strictly below V^{SB} by Lemma 1. Thus, implementation also obtains with an observable signal, yet at the complexity cost of adding an extra message.

To circumvent the verifiability paradox, the modified mechanism γ^{**} must require that the agent's messages initially remain private. If messages were public, the principal could make message-conditional renegotiation offers that defeat the mechanism. For instance, he could offer attractive terms only for message N while making R_1 and R_2 lead to terrible outcomes. This would induce the agent to send message N and accept renegotiation, eliminating the punishment mechanism entirely.

We now show that, despite its dependence on private messages, γ^{**} can be implemented via smart contracts that are self-executing programs on transparent blockchains.[20] While this may seem paradoxical given that blockchain transactions are publicly recorded,

[18] In practice, s can be instantiated via a verifiable on-chain randomness source (e.g., a verifiable random function (Micali et al., 1999) or reputable randomness oracle); the choice determines trust and liveness assumptions. If the randomness source fails, a two-party commit-and-reveal coin toss between principal and agent can serve as a fallback.

[19] This is shown formally in Appendix 9.

[20] For an extensive definition of a smart contract see Szabo (1996) and Catalini and Gans (2020) for a discussion of potential economic applications for smart contracts. We here emphasize however that,

cryptographic techniques allow us to achieve the required privacy within this transparent environment.

In particular, the commit-and-reveal technique solves this privacy challenge by allowing parties to record information that remains hidden initially but can be publicly verified later. Technically, the technique is a cryptographic protocol with two phases. In the commit phase, a party uses a hash function to create a cryptographic commitment to her message without revealing it. In the reveal phase, she can publicly disclose the original message, which others can verify matches the earlier commitment.[21]

The technique relies on hash functions that are one-way and collision-resistant, making it impossible to derive the original message from the commitment or to create fake commitments. This ensures the message remains secret until revealed while preventing later manipulation. This enables us to implement self-revealing mechanisms on transparent blockchains by emulating their defining property: recording secret messages that are revealed only later. During the commit phase, the agent's message remains hidden while the commitment is publicly recorded. During the reveal phase, the agent discloses her message, which the smart contract verifies against the stored commitment. This process maintains message secrecy until the designated reveal time while ensuring the message cannot be altered after commitment.

To demonstrate the practical feasibility concretely, we present in Appendix 8 a complete Solidity smart contract that implements γ^{**} using the commit-and-reveal technique for a fully parameterized version of our framework. The implementation shows that self-revealing mechanisms can indeed be deployed on current blockchain technologies, bridging the gap between theoretical mechanism design and real-world contracting.

While smart contracts are often seen as immune to renegotiation,[22] in practice they commonly include functions allowing termination or modification. For instance, DeFi protocols often feature *emergency stop* or *circuit breaker* functions that automatically freeze execution when pre-set risks are met. Others, such as OpenZeppelin's Pausable module or MakerDAO's Emergency Shutdown, allow authorized parties to manually halt operations through governance control. Modules allowing built-in *modification* rights are also common: for example, proxy-based upgrades used by Compound and OpenZeppelin allow preserving the state while replacing the contract's code logic (see Ebrahimi et al.,

in general, an enforcement of smart contracts depends on the shadow of the law. To see this in our specific context of γ^{**}, note that because its transfers condition on the realized output value $Y \in \{g, b\}$, the realized output value must somehow be reported to the smart contract. This can be done by, for instance, the principal, but only the verifiability by a court ensures that the principal will do so truthfully, anticipating its prohibitively large punishment when misreporting.

[21] See Narayanan et al. (2016, Chapter 1) for a more in-depth introduction to cryptographic hash functions and the reveal-and-commit technique.

[22] See, for example, the discussion in Chapter 6 in Townsend (2020).

2024).

By explicitly allowing both contract termination and modification, these adaptability functions reintroduce classic time-consistency concerns in the smart contracts paradigm.[23] We regard our results as relevant in this respect: the finding that blockchain-compatible mechanisms can replicate full-commitment outcomes under a traditional renegotiation constraint suggests that, by careful structuring of the smart contract's transfers, one can preserve contractual flexibility while neutralizing the inefficient modification incentives that adaptability functions create.

This connects our work to concrete efforts to design governance mechanisms deterring harmful upgrades while preserving adaptability in smart contracts, such as: multi-signature authorization, DAO voting systems,[24] and *timelocks* between the approval and implementation of upgrades, which give users time to assess and exit the contract before changes take effect.[25]

5 The Power of Self-Revealing Mechanisms

The power of self-revealing mechanisms extends beyond the baseline framework in two main directions: alternative extensive forms for the renegotiation game (Section 5.1) and alternative specifications of the renegotiation process itself (Section 5.2).

In addition, our approach does not exploit the principal's specific objective function. A utilitarian planner can rely on a modified version of γ^* to implement second-best insurance under renegotiation threats, addressing the government failure emphasized by Netzer and Scheuer (2010).

5.1 Alternative Extensive Forms

Propositions 1 and 2 extend to situations in which the agent, rather than the principal, initiates renegotiation, as analyzed by Ma (1994). In that framework, our approach yields unique implementation even when renegotiation threats originate from the agent, contrasting with the equilibrium multiplicity in Ma (1994).[26]

[23] See also, on this topic, Salehi et al. (2022); Wang et al. (2025) and the Ethereum guide on upgrading smart contracts.

[24] See OpenZeppelin's on-chain governance framework.

[25] "Timelocks give users some time to exit the system if they disagree with a proposed change (e.g., logic upgrade or new fee schemes). Without timelocks, users need to trust developers not to implement arbitrary changes in a smart contract without prior notice. The drawback here is that timelocks restrict the ability to quickly patch vulnerabilities" (source).

[26] A formal argument is available from the authors. The proof constructs a modified version of γ^* where the principal reports the renegotiation offers he receives from the agent. This extends the punishment logic of γ^* to the Ma (1994) setting, deterring the agent from renegotiating under *every* system of principal's beliefs.

The punishment logic of our self-revealing mechanism is forward-looking: the agent's incentive to report a renegotiation attempt depends on the fallback lottery she activates, not on the number of remaining renegotiation rounds. This suggests that backward induction reasoning is not essential to our approach. To verify this, we extend the analysis to an infinite horizon.

Our original construction follows FT in assuming that the renegotiation offer γ^r is made only once, i.e., at stage (iv). Adding any *finite* number k of renegotiation rounds introduces no new strategic effects: all bargaining would occur in the last round, making the analysis equivalent to the single-round case.[27] However, the second-best allocation implemented by γ^* involves inefficient risk sharing, leaving room for further renegotiation after the last round. Under infinite renegotiation, this room persists indefinitely, and the backward-induction logic that collapses finite rounds to a single round no longer applies.

We consider an infinite-horizon setting in the spirit of Strulovici (2017): parties interact over rounds $T = 1, 2, \ldots$, agreeing ex-ante on a mechanism that can be renegotiated any number of times. Renegotiation breaks down with probability $\eta \in (0, 1)$ in each round $T \geq 1$, at which point output $\omega \in \{g, b\}$ realizes and the last accepted contract executes.

Thus, the breakdown round T^* follows a geometric distribution: $\mathbb{P}(T^* = T) = (1 - \eta)^{T-1} \cdot \eta$ and $\mathbb{P}(T^* = T' \mid T^* \geq T) = (1 - \eta)^{T'-T} \cdot \eta$. For both players, the time-T expectation of a unit of utility is:

$$\sum_{T' \geq T} (1 - \eta)^{T'-T} \cdot \eta = \frac{\eta}{1 - (1 - \eta)} = 1.$$

For a given η, we denote G^η the corresponding primitive game, which extends the game G by allowing for infinite renegotiation rounds.

We construct a self-revealing mechanism ξ^{0*}, which implements the second-best allocation. While Strulovici (2017) shows that infinite renegotiation erodes incentives by driving parties toward Coasian full-insurance outcomes, our mechanism prevents this erosion, sustaining the second-best allocation across every round.

The mechanism ξ^{0*}, offered by the principal and accepted by the agent at the onset of the relationship, induces the subgame $G^\eta_\Xi(\xi^{0*})$ (that is, the game G^η with renegotiation offers selected from the class Ξ of self-revealing mechanisms after ξ^{0*} is chosen):

- At $T = 0$: The agent privately selects the effort level $e \in \{H, L\}$.

- At any $T \geq 1$ the following sequence of events is involved:

$T.i)$ The principal offers $\xi^T \in \Xi \cup \{\emptyset\}$.

[27] See Section 6B in Fudenberg and Tirole (1990) for a formal proof of this result relative to renegotiation game G_C.

T.ii) The agent makes a report in the last accepted mechanism. Simultaneously, the mechanism privately discloses a signal to the agent.

T.iii) The agent accepts ($\rho^T = y$) or rejects ($\rho^T = n$) the renegotiation offer ξ^T, with the convention that $\rho^T = n$ if $\xi^T = \emptyset$.

T.iv) If $\rho^T = y$, the agent submits a report $\hat{e}^T \in \{H, L\}$ to ξ^T. Then, if renegotiation breaks down, $\omega \in \{g, b\}$ realizes, the last accepted mechanism publicly reveals its communication history and executes transfers; otherwise the game continues to $T + 1$.

In $G^{\eta}_{\Xi}(\xi^{0*})$, after the agent chooses effort, T^* rounds of renegotiation take place, in which the actions $T.i) - T.iv)$ are iterated at each $T : 1 \leq T \leq T^*$. The parties are uncertain about the realization of T^* until renegotiation breaks down and the game ends. The mechanism ξ^{0*} requires the agent to submit a report $m_T^{0*} \in \{N, R\}$ in each round $T.ii)$, i.e. after a renegotiation offer ξ^T is made and before the agent decides to accept it.

The report N maintains the status quo (inducing the transfers c^{SB}), while R irreversibly triggers a lottery over full-insurance transfers at different utility levels, the outcome of which is privately disclosed to the agent via a fair coin toss. This communication protocol naturally extends that of γ^* to an infinite horizon.

The principal may attempt to renegotiate ξ^{0*} at any $T \geq 1$, until T^* realizes. We next show that the implementation result of Proposition 1 extends to this setting, which suggests that backward induction reasoning is *not* key to our approach.

Specifically, we establish the following:

Proposition 3 *The second-best allocation* (H, c^{SB}) *is supported in an equilibrium of* $G^{\eta}_{\Xi}(\xi^{0*})$.

The proof of Proposition 3, provided in Appendix 7, constructs an equilibrium of $G^{\eta}_{\Xi}(\xi^{0*})$ in which the principal offers $\xi^T = \emptyset$ for all $T \geq 1$ and the agent selects $e = H$ sending the status-quo report $m_T^{0*} = N$ for all $T \geq 1$. The proof shows that any renegotiation offer $\xi^T \in \Xi$ that the agent accepts and that yields her a continuation utility $\hat{U} \in \mathbb{R}$ is unprofitable. If $\hat{U} \leq U^0 - (\Delta U - d)$ the agent optimally rejects it; if $\hat{U} > U^0 - (\Delta U - d)$, she optimally sends the report R, activating a punishment similar to that constructed in the proof of Proposition 1. Crucially, the argument does not hinge on the specific features of the communication protocol associated with the renegotiation offer, only on the continuation utility $\hat{U}$. Restricting to renegotiation offers in the class Ξ therefore eases exposition without affecting the general validity of our approach.

In addition, the proof does not exploit any restrictions on the principal's off-path beliefs. We show that, for any off-equilibrium-path history in which the original mechanism is not renegotiated, there is *no* system of beliefs under which the principal would profitably renegotiate.[28] This implies that Proposition 2 extends to the infinite-horizon setting. That is, one can construct a perturbed version of ξ^{0*}, for which choosing not to renegotiate is the principal's unique best response, yielding him a utility arbitrarily close to the full-commitment utility. Propositions 2 and 3 thus deliver unique implementation in the infinite-horizon setting.

5.2 The Supplementary View of Renegotiation

Under the primitive game G that underlies both FT and our framework, self-revealing mechanisms uniquely implement the second-best allocation, fully mitigating renegotiation at no efficiency loss. This subsection examines the robustness of this result to an alternative specification of the renegotiation process itself.

Thus far, we adopted the *replacement* view of renegotiation, following the standard approach in the literature that a renegotiation offer replaces the original mechanism (Assumption $A.3$).[29] Under this view, the agent cannot combine γ^* with a renegotiation offer—contracts are exclusive. We now examine *supplementary* renegotiation, where such combinations are possible. This introduces a new strategic possibility: the principal can condition his renegotiation offer on the transfers realized under γ^*.[30]

Consider the following supplementary offer. The principal proposes a mechanism γ_+^r that, when *combined* with γ^*, yields the full-insurance contract $c_H^{FI}(U^0+\varepsilon)$ for some $\varepsilon > 0$. The key feature is that γ_+^r conditions on γ^*'s realized transfers: it pays a transfer only if γ^* implements $(u_g^{IC}(U^0), u_b^{IC}(U^0))$ as defined by (1). Specifically, γ_+^r pays the difference between $c_H^{FI}(U^0+\varepsilon)$ and c^{SB}.

Upon observing γ_+^r, the agent finds it optimal to report $m = N$ in γ^*, triggering the second-best transfers (as reporting $m = R$ would trigger the punishment lottery, causing γ_+^r to pay nothing since its transfer is contingent on γ^* implementing c^{SB}). She then accepts γ_+^r, which offsets these transfers and implements $c_H^{FI}(U^0+\varepsilon)$, guaranteeing her a

[28] If, at any such history, he believes that R has been sent by the agent in the past, he already expects to face some first-best transfers. If, instead, he believes that R has never been sent, he expects the agent to follow her on-path behavior, which, as already argued, makes any attempt to renegotiate unprofitable.

[29] As Bolton (1990, p. 304) notes: "[...] For once the contracting parties reach the point where an inefficient outcome is suggested by the contract, they can always tear up the initial contract and write a new Pareto-improving contract. As a result, when the contracting parties are unable to commit not to renegotiate they will have to abandon these contracts designed to be executed without renegotiation".

[30] Observability of final transfers could also be exploited under the replacement view. However, the exclusivity assumption (Assumption $A.3$) that defines this view makes any such conditional offer strategically irrelevant: the profitability of the above renegotiation offer hinges on the ability to combine it with the original mechanism.

utility $U^0 + \varepsilon > U^0$. The principal also gains: for ε sufficiently small, the reduction in expected transfers from eliminating the risk premium in c^{SB} outweighs the rent transferred to the agent.

Both parties are then strictly better off. Thus, γ^* is vulnerable to supplementary renegotiation: the principal can profitably deviate by conditioning on γ^*'s realized transfers.

However, alternative mechanisms can restore the second-best also under supplementary renegotiation. Consider a modified self-revealing mechanism with the following structure: when output $\omega = b$ realizes, it pays a flat (non-contingent) transfer; when $\omega = g$ realizes, it pays an (m, s)-conditional transfer. Because the transfer is flat when $\omega = b$, the principal cannot infer the agent's communication from the realized payment in that state. This prevents the principal from inferring the agent's message by observing both the transfer and the realized output, eliminating the vulnerability demonstrated above.

We formalize this argument in Proposition 4, provided in Appendix 9, which establishes that the second-best allocation remains implementable under supplementary renegotiation for CRRA preferences. The result obtains without any tie-breaking assumption: the agent's sequential rationality alone is sufficient to deter any renegotiation offer. The perturbation arguments of Proposition 2 are then expected to extend to supplementary renegotiation.

The result illustrates an important point about the relationship between supplementary and replacement renegotiation. Under the replacement view, the verifiability paradox is the primary challenge: the principal cannot observe the agent's communication within γ^*, creating the uncertainty that deters renegotiation. Under supplementary renegotiation, an additional difficulty arises: while communication remains unverifiable, the principal can observe its consequences through the realized transfers from γ^*. The modified mechanism addresses both challenges by revealing communication through transfers only when $\omega = g$, not when $\omega = b$. This selective revelation conceals the agent's message in the bad state while preserving the incentive structure in the good state.

To summarize, self-revealing mechanisms enable second-best implementation both under the standard replacement view and under the supplementary view of renegotiation, though the details of the required construction differ.

6 Conclusion

We revisit the tension between the legal doctrine of renegotiation and economic efficiency (Dewatripont, 1989). We show that the threat of renegotiation can be fully mitigated by self-revealing mechanisms with bidirectional communication that keep messages private at the renegotiation stage yet verifiable at execution. The combination of bidirectional

communication and conditional ~~its strategically timed~~ disclosure enables off-equilibrium punishments that restore the (full-commitment) second-best allocation without distorting incentives on the equilibrium path.

We establish these results in the canonical renegotiation framework of Fudenberg and Tirole (1990), and we show that they also obtain in alternative settings of renegotiation under moral hazard. In particular, the second-best allocation can be implemented under infinite rounds of renegotiation, showing that our approach does not rely on backward-induction reasoning. We have also verified that our approach extends to renegotiation under incomplete information, in particular to the setting of Laffont and Tirole (1990). In their procurement framework, a buyer contracts with a seller whose marginal cost $\theta_i \in \{\theta_1, \theta_2\}$ (with $\theta_2 > \theta_1$) is private information. The second-best allocation involves rationing: type θ_2 sells less than the first-best quantity. This is fragile to renegotiation: in the final period, the buyer can exploit remaining gains from trade with type θ_2. A self-revealing mechanism can be designed to prevent this renegotiation without slowing down the revelation of the agent's private information. The key insight is that our approach requires generating private information only off the equilibrium path, where it is needed to activate punishments against renegotiation. Pre-existing private information on the equilibrium path does not interfere: a self-revealing mechanism generates an additional private signal that type θ_2 exploits exclusively in response to a renegotiation offer. We omit the formal treatment because it requires analyzing an environment with a distinct information structure.

Our results carry significant implications. Self-revealing mechanisms reframe renegotiation-proofness as a problem of communication architecture: the law's refusal to enforce no-renegotiation clauses need not bind efficiency once private signals and conditional revelation are available.

The institutional message is that standard court enforcement suffices when contracts embed this timing of information, aligning legal doctrine with economic efficiency rather than requiring some external commitment devices or third-party mediation. Practically, commit-and-reveal cryptographic tools operationalize the required conditional disclosure, indicating that algorithmic contracting can implement the information structure that eliminates renegotiation incentives.

More broadly, the analysis suggests a design principle for contract theory: when ex-post inefficiencies create scope for opportunism, engineering when and to whom information is disclosed can substitute for formal commitment, with implications for environments beyond the canonical model and for the governance of digital markets.

7 Main Proofs

This appendix collects the proofs.

Proof of Lemma 1. For a given $e \in E$, define the function $\tilde{V}_e : [U^0, \infty) \to \mathbb{R}$ as

$$\tilde{V}_e(U) \equiv \frac{1}{2} V_e^{FI}(2U^0 - U) + \frac{1}{2} V_e^{FI}(U).$$

The function satisfies the following properties:

a) $\tilde{V}_e(U)$ is well-defined, continuous and twice differentiable for $U \in [U^0, \infty)$, because $\Phi(U)$, and thus $V_e^{FI}(U)$, are defined for every $U \in (-\infty, +\infty)$ and, moreover, are continuous and twice differentiable.

b) $\tilde{V}_e(U)$ is strictly decreasing since

$$\frac{\partial \tilde{V}_e(U)}{\partial U} = \frac{1}{2} \frac{\partial V_e^{FI}(U)}{\partial U} - \frac{1}{2} \frac{\partial V_e^{FI}(2U^0 - U)}{\partial U} < 0$$

for any $U \in (U^0, \infty)$, where the inequality obtains since $U > 2U^0 - U$, and because $V_e^{FI}(U)$ is concave so that $\partial V_e^{FI}/\partial U$ is decreasing.

c) $\tilde{V}_e(U)$ is strictly concave since

$$\frac{\partial^2 \tilde{V}_e(U)}{\partial U^2} = \frac{1}{2} \frac{\partial^2 V_e^{FI}(U)}{\partial U^2} + \frac{1}{2} \frac{\partial^2 V_e^{FI}(2U^0 - U)}{\partial U^2} < 0,$$

where the inequality follows because $\partial^2 V_e^{FI}(U)/\partial U^2 < 0$.

d) It follows from (b) and (c) that $\lim_{U \to \infty} \tilde{V}_e(U) = -\infty$.

e) For each $e \in E$, there is a $\underline{U}_e \in (U^0, \infty)$ such that

$$V_e^{IC}(U^0) = \tilde{V}_e(\underline{U}_e) \quad \text{and} \quad V_e^{IC}(U^0) > \tilde{V}_e(U) \quad \forall U \in (\underline{U}_e, \infty).$$

This holds since $\tilde{V}_e(U^0) = V_e^{FI}(U^0) > V_e^{IC}(U^0) > \lim_{U \to \infty} \tilde{V}_e(U) = -\infty$, where the first inequality follows from the convexity of Φ. Because $\tilde{V}_e(U)$ is continuous, the intermediate value theorem guarantees that there is a $\underline{U}_e \in (U^0, \infty)$: $\tilde{V}_e(\underline{U}_e) = V_e^{IC}(U^0)$. Because $\tilde{V}_e(U)$ is strictly decreasing, we have $\tilde{V}_e(U) < \tilde{V}_e(\underline{U}_e) = V_e^{IC}(U^0)$ for all $U > \underline{U}_e$.

It follows from (e) that, for any $U^n > \max\{\underline{U}_H, \underline{U}_L\}$, we have

$$V_e^{IC}(U^0) > \tilde{V}_e(U^n). \tag{4}$$

Since $U^n > U^0 \Leftrightarrow U^n > 2U^0 - U^n$, it follows from $V_e^{FI}(U)$ decreasing and Φ convex that:

$$\tilde{V}_e(U^n) = \frac{1}{2}V_e^{FI}(2U^0 - U^n) + \frac{1}{2}V_e^{FI}(U^n) > \max\left\{V_e^{FI}(U^n), \frac{1}{2}V_e^{FI}(2U^0 - U^n) + \frac{1}{2}V_e^{IC}(U^n)\right\}. \tag{5}$$

Taking $\Delta U = U^n - U^0 > 0$ together with both (4) and (5) imply (2). ■

Proof of Lemma 2. By (3), sending $\hat{m}_e^r$ is sequentially rational at any history $(e, \gamma^r \neq \emptyset, m, s, y)$, where $m \in \{N, R\}$ denotes the agent's message in γ^*. Comparing her utility $U_e(\tau^*(m, s))$ from remaining in γ^* with her utility $\hat{U}_e^r$ from accepting γ^r then implies that the participation behavior we construct is optimal at each history $(e, \gamma^r \neq \emptyset, m, s)$. ~~for any $(e, \gamma^r \neq \emptyset, m, s, y)$ with $m \in \{N, R\}$ the agent's message in the original self-revealing mechanism γ. From comparing her utility $U_e(\tau^*(m, s))$ of remaining in γ^* with her utility $\hat{U}_e^r$ (defined in (3)) of accepting γ^r, it follows that, at each history $(e, \gamma^r \neq \emptyset, m, s)$, the agent's participation behavior we construct is optimal.~~ Next, consider any history $(e, \gamma^r \neq \emptyset)$. Because the agent observes s before making her decision ρ, her continuation value under $m = R$ equals

$$\frac{1}{2}\max\left\{U^0 - \Delta U, \hat{U}_e^r\right\} + \frac{1}{2}\max\{U^0 + \Delta U, \hat{U}_e^r\},$$

while under $m = N$ it equals

$$\max\{U^0, \hat{U}_e^r\}.$$

Hence, it is optimal for the agent to send $m = N$ in γ^* if

$$\max\{U^0, \hat{U}_e^r\} \geq \frac{1}{2}\max\left\{U^0 - \Delta U, \hat{U}_e^r\right\} + \frac{1}{2}\max\{U^0 + \Delta U, \hat{U}_e^r\}. \tag{6}$$

From (6), it follows that the agent's reporting behavior we construct is optimal:

(*i*) If $\hat{U}_e^r \leq U^0 - \Delta U$, then (6) is satisfied because it reduces to $U^0 \geq U^0$ since $\hat{U}_e^r \leq U^0 - \Delta U < U^0$. Sending $m = N$ in γ^*, followed by $\rho = n$ for both $s \in \{h, t\}$ as prescribed by $\lambda(\gamma^*)$, is hence optimal.

(*ii*) If $\hat{U}_e^r \in (U^0 - \Delta U, U^0 + \Delta U)$, then, upon sending $m = R$ as prescribed by $\lambda(\gamma^*)$, the agent selects $\rho = y$ when $s = h$ and $\rho = n$ when $s = t$. We now argue that sending $m = R$ in γ^* is indeed optimal. To show this, we need to establish that the reverse of inequality (6) holds, where we note that, due to $\hat{U}_e^r \in (U^0 - \Delta U, U^0 + \Delta U]$, its RHS reduces to $\hat{U}_e^r/2 + (U^0 + \Delta U)/2$. Namely, we verify that, for every $\hat{U}_e^r$ in the interval,

$$\max\{U^0, \hat{U}_e^r\} \leq \frac{1}{2}\hat{U}_e^r + \frac{1}{2}(U^0 + \Delta U). \tag{7}$$

To get the result, it is sufficient to observe that:

(a) If $\hat{U}_e^r < U^0$, then (7) rewrites as $U^0 - \Delta U \leq \hat{U}_e^r$, which is satisfied by assumption.

(b) If $\hat{U}_e^r \geq U^0$, then (7) rewrites as $\hat{U}_e^r \leq U^0 + \Delta U$, which is also satisfied by assumption.

(*iii*) If $\hat{U}_e^r \geq U^0 + \Delta U$, then we have $U^0 < U^0 + \Delta U \leq \hat{U}_e^r$ and the agent's continuation value under $m = R$ equals $\hat{U}_e^r$, the same obtained under $m = N$ (followed by $\rho = y$ for any $s \in \{h, t\}$). Hence, it is rational for the agent, as prescribed by $\lambda(\gamma^*)$, to send $m = R$ in γ^* and then accept γ^r for any received signal.

Therefore, in every history $(e, \gamma^r \neq \emptyset)$, the prescribed choices in $\lambda(\gamma^*)$—$m = N$ in case (*i*); $m = R$ with $\rho = y$ when $s = h$ and $\rho = n$ when $s = t$ in case (*ii*); and $m = R$ with $\rho = y$ for any s in case (*iii*) —are optimal.

Consider now the agent's behavior at each history $(e, \emptyset)$ where the principal does *not* renegotiate. Sending $m = N$ is optimal for the agent since she obtains the same utility U^0 under any report in γ^*. Finally, at her starting node, she optimally selects $e = H$ against $\gamma^r = \emptyset$, since she anticipates that no renegotiation takes place on path and $c^{SB} = c^{IC}(U^0)$ is eventually implemented, which makes her indifferent between effort levels. This completes the proof of the agent's sequential rationality. ■

Proof of Lemma 3: Start by noting that the principal obtains V^{SB} under his equilibrium behavior $\gamma^r(\gamma^*) = \emptyset$: indeed, the agent's behavior $\lambda(\gamma^*)$ prescribes her to report $m = N$ in γ^* when $\gamma^r = \emptyset$, inducing the implementation of c^{SB}, which, together with the agent's equilibrium effort decision $e = H$, yields $V^{SB} = V_H(c^{SB})$ to the principal. We now verify that, given the principal's belief $x = 1$ and the agent's behavior $\lambda(\gamma^*)$, the principal's expected utility after every deviation $\gamma^r \neq \emptyset$ does not exceed V^{SB}.

Because of the principal's degenerate belief and the agent's risk-aversion, any accepted γ^r that maximizes the principal's utility yields full insurance to the agent of type $e = H$. So, conditional on acceptance, his utility from an optimal offer equals $V_H^{FI}(\hat{U}_H^r)$ for some scalar $\hat{U}_H^r \in \mathbb{R}$. This noted, we distinguish three classes of offers according to the value of $\hat{U}_H^r$, and we show that no offer yields more than V^{SB} to the principal:

(*i*) If $\hat{U}_H^r \leq U^0 - \Delta U$ then $\lambda(\gamma^*)$ prescribes $(m = N, \rho = n)$ and the principal gets V^{SB}.

(*ii*) If $\hat{U}_H^r \in (U^0 - \Delta U, U^0 + \Delta U)$ then $\lambda(\gamma^*)$ prescribes ($m = R$, $\rho = y$ when $s = h$, and $\rho = n$ when $s = t$), and the principal gets

$$\frac{1}{2}V_H^{FI}(\hat{U}_H^r) + \frac{1}{2}V_H^{IC}(U^0 + \Delta U) < \frac{1}{2}V_H^{FI}(U^0 - \Delta U) + \frac{1}{2}V_H^{IC}(U^0 + \Delta U) < V^{SB}, \quad (8)$$

where the first inequality follows from V_H^{FI} decreasing, and the second from Lemma 1.

(*iii*) If $\hat{U}_H^r \geq U^0 + \Delta U$ then $\lambda(\gamma^*)$ prescribes $(m = R,\, \rho = y)$ for any $s \in \{h, t\}$, and the principal gets

$$V_H^{FI}(\hat{U}_H^r) < V_H^{FI}(U^0 + \Delta U) < V^{SB} \tag{9}$$

$$V_H^{FI}(\hat{U}_H^r) \leq V_H^{FI}(U^0 + \Delta U) < V^{SB} \tag{10}$$

where the first inequality follows from V_H^{FI} decreasing, and the second from Lemma 1.

Thus, the principal cannot profitably deviate by offering any $\gamma^r \neq \emptyset$, making the constructed behavior $\gamma^r(\gamma^*) = \emptyset$ sequentially rational. ■

Proof of Proposition 2. We construct a mechanism γ_ε that uniquely implements $e = H$ and yields a principal's utility arbitrarily close to V^{SB}.

Define for any $\varepsilon \in (0, \bar{\varepsilon})$ with $\bar{\varepsilon} > 0$, the contract

$$c_\varepsilon^{SB} \equiv \left(U^0 + \frac{(1-p_L)d + (1-p_H)\varepsilon}{\Delta p},\; U^0 - \frac{p_L d + p_H \varepsilon}{\Delta p} \right).$$

Note that c_ε^{SB} yields the agent the utility U^0 if she selects $e = H$, and $U^0 - \varepsilon$ if $e = L$.

Mechanism $\gamma_\varepsilon = \{\mathcal{M}^*, \mathcal{S}^*, \sigma^*, \tau_\varepsilon\}$ coincides with γ^*, except for τ_ε:

$$\tau_\varepsilon(N,h) = \tau_\varepsilon(N,t) = c_\varepsilon^{SB};\;\; \tau_\varepsilon(R,t) = c^{IC}(U^0 + \Delta U);\;\; \tau_\varepsilon(R,h) = c^{IC}(U^0 - \Delta U - \kappa\varepsilon)$$

for any arbitrary $\kappa > 2$. We consider the subgame $G_\Gamma(\gamma_\varepsilon)$, and construct $\bar{\varepsilon} > 0$ so that, for any belief $x \in [0,1]$ and any $\varepsilon \in (0, \bar{\varepsilon})$, the principal is strictly worse off from any renegotiation offer that the agent accepts with a strictly positive probability.

Fixing an arbitrary behavior $\gamma^r(\gamma_\varepsilon)$ of the principal, we first characterize all the agent's reporting and participation decisions that are sequentially rational in the subgame $G_\Gamma(\gamma_\varepsilon)$ for every renegotiation offer γ^r. We start from the terminal nodes of $G_\Gamma(\gamma_\varepsilon)$.

Recalling (3), note that in any history (e, γ^r, m, s, y) with $\gamma^r \neq \emptyset$, the agent sends any $m^r \in E$ (or distribution over reports) that satisfies the left-hand side of (3), expecting to obtain $\hat{U}_e^r$ from accepting γ^r as expressed in the right-hand side of (3).

In any history $(e, \gamma^r \neq \emptyset, m, s)$, the agent's optimal acceptance behavior $(\rho(h), \rho(t))$ follows from comparing the agent's utility $U_e(\tau_\varepsilon(m,s))$ of staying in γ_ε with the utility $\hat{U}_e^r$ of accepting γ^r:

(a) For $(e,m) = (H,R)$ and $(e,m) = (L,R)$, we have

$$\rho(h) \in \begin{cases} \{y\} & \text{if } \hat{U}_e^r > U^0 - \Delta U - \kappa\varepsilon; \\ \{n\} & \text{if } \hat{U}_e^r < U^0 - \Delta U - \kappa\varepsilon; \\ \{n,y\} & \text{if } \hat{U}_e^r = U^0 - \Delta U - \kappa\varepsilon; \end{cases} \quad \text{and } \rho(t) \in \begin{cases} \{y\} & \text{if } \hat{U}_e^r > U^0 + \Delta U; \\ \{n\} & \text{if } \hat{U}_e^r < U^0 + \Delta U; \\ \{n,y\} & \text{if } \hat{U}_e^r = U^0 + \Delta U. \end{cases}$$

(b) For $(e,m)=(H,N)$, we have

$$\rho(h)\in\begin{cases}\{y\} & \text{if } \hat{U}_H^r>U^0;\\ \{n\} & \text{if } \hat{U}_H^r<U^0;\\ \{n,y\} & \text{if } \hat{U}_H^r=U^0;\end{cases}\quad\text{and } \rho(t)\in\begin{cases}\{y\} & \text{if } \hat{U}_H^r>U^0;\\ \{n\} & \text{if } \hat{U}_H^r<U^0;\\ \{n,y\} & \text{if } \hat{U}_H^r=U^0.\end{cases}$$

(c) For $(e,m)=(L,N)$, we have

$$\rho(h)\in\begin{cases}\{y\} & \text{if } \hat{U}_L^r>U^0-\varepsilon;\\ \{n\} & \text{if } \hat{U}_L^r<U^0-\varepsilon;\\ \{n,y\} & \text{if } \hat{U}_L^r=U^0-\varepsilon;\end{cases}\quad\text{and } \rho(t)\in\begin{cases}\{y\} & \text{if } \hat{U}_L^r>U^0-\varepsilon;\\ \{n\} & \text{if } \hat{U}_L^r<U^0-\varepsilon;\\ \{n,y\} & \text{if } \hat{U}_L^r=U^0-\varepsilon.\end{cases}$$

Fixing any optimal participation behavior as characterized above, we now derive the agent's optimal reporting behavior in any history $(e,\gamma^r\neq\emptyset)$, where γ^r yields $\hat{U}_e^r$ to the agent if accepted. For $e=H$, $m=N$ is optimal if

$$\max\{U^0,\hat{U}_H^r\}\geq\frac{1}{2}\max\left\{\hat{U}_H^r,U^0-\Delta U-\kappa\varepsilon\right\}+\frac{1}{2}\max\{\hat{U}_H^r,U^0+\Delta U\},\tag{11}$$

while $m=R$ is optimal if the opposite weak inequality holds. For $e=L$, $m=N$ is optimal if

$$\max\{U^0-\varepsilon,\hat{U}_L^r\}\geq\frac{1}{2}\max\left\{\hat{U}_L^r,U^0-\Delta U-\kappa\varepsilon\right\}+\frac{1}{2}\max\{\hat{U}_L^r,U^0+\Delta U\},\tag{12}$$

while $m=R$ is optimal if the opposite weak inequality holds.

At any history $(e,\gamma^r=\emptyset)$, the agent's unique optimal report in γ_ε is $m=N$. To see this, note that $m=N$ yields U^0 if $e=H$ and $U^0-\varepsilon$ if $e=L$, whereas $m=R$ yields $U^0-\frac{\kappa}{2}\varepsilon$ regardless of e. Since $\varepsilon>0$ and $\kappa>2$, we have $U^0-\frac{\kappa}{2}\varepsilon<U^0-\varepsilon<U^0$, confirming that $m=N$ is strictly preferred to $m=R$ for both effort levels

We now derive the principal's optimal behavior in the subgame $G_\Gamma(\gamma_\varepsilon)$. We show in particular that, for any effort probability $x\in[0,1]$ and any optimal reporting and participation behavior of the agent as characterized above, every sequentially rational principal's behavior involves either $\gamma^r=\emptyset$ or, equivalently, any offer that the agent rejects with probability one.

First, suppose $\lambda(\gamma_\varepsilon)$ specifies that the agent selects $e=H$ with probability $x\in\{0,1\}$ and the principal holds a deterministic, consistent belief $x\in\{0,1\}$ over the agent's effort. In this case, by not renegotiating, given the agent's subsequent report $m=N$, the principal expects $V_H(c_\varepsilon^{SB})$ if $x=1$ or $V_L(c_\varepsilon^{SB})$ if $x=0$. To verify that the principal cannot do better by offering $\gamma^r\neq\emptyset$, we start by indexing every possible offer γ^r by $\hat{U}_x^r\in(-\infty,+\infty)$, the agent's optimal expected utility which she obtains from accepting it. When $x=1$, the relevant utility is that of the agent who chose $e=H$; when $x=0$, that of the agent who chose $e=L$. Recall also that under degenerate beliefs, as noted

in the proof of Lemma 3, the principal's utility from an optimal offer yielding $\hat{U}_e^r$ to the agent, conditional on its acceptance, is $V_e^{FI}(\hat{U}_e^r)$. Using the agent's sequentially rational behavior as derived above by substituting $\hat{U}_H^r = \hat{U}_1^r$ and $\hat{U}_L^r = \hat{U}_0^r$, we derive the optimal utility that the principal expects from every $\hat{U}_x^r \in \mathbb{R}$:

1. For $\hat{U}_1^r < U^0 - \Delta U$, the principal expects utility $V_H(c_\varepsilon^{SB})$ and for $\hat{U}_0^r < U^0 - \Delta U - 2\varepsilon$, the principal expects utility $V_L(c_\varepsilon^{SB})$. This follows because the principal expects the agent to consider her strategy $(m, \rho(t), \rho(h)) = (N, n, n)$ uniquely optimal. To see this, note first that conditional on sending $m = N$, $\rho(t) = \rho(h) = n$ is strictly optimal, because

$$\hat{U}_1^r < U^0 - \Delta U < U^0 \quad \text{and} \quad \hat{U}_0^r < U^0 - \Delta U - 2\varepsilon < U^0 - \varepsilon.$$

   To see instead why the principal expects the agent to strictly prefer $m = N$ over $m = R$, consider the two subcases:

   (a) If $\hat{U}_1^r \leq U^0 - \Delta U - \kappa\varepsilon$, then (11) with $\hat{U}_H^r = \hat{U}_1^r$ becomes $U^0 \geq U^0 - \frac{\kappa}{2}\varepsilon$; likewise, if $\hat{U}_0^r \leq U^0 - \Delta U - \kappa\varepsilon$, then (12) with $\hat{U}_L^r = \hat{U}_0^r$ becomes $U^0 - \varepsilon \geq U^0 - \frac{\kappa}{2}\varepsilon$. Both inequalities are strictly satisfied since $\varepsilon > 0$ and $\kappa > 2$.

   (b) If $\hat{U}_1^r \in (U^0 - \Delta U - \kappa\varepsilon, U^0 - \Delta U)$, then (11) with $\hat{U}_H^r = \hat{U}_1^r$ becomes $\hat{U}_H^r \leq U^0 - \Delta U$; likewise, if $\hat{U}_0^r \in (U^0 - \Delta U - \kappa\varepsilon, U^0 - \Delta U - 2\varepsilon)$, then (12) with $\hat{U}_L^r = \hat{U}_0^r$ becomes $\hat{U}_0^r \leq U^0 - \Delta U - 2\varepsilon$. Both inequalities are strictly satisfied in case (b) by assumption.

2. For $\hat{U}_1^r = U^0 - \Delta U$ or $\hat{U}_0^r = U^0 - \Delta U - 2\varepsilon$, the principal expects the agent to consider only the strategies $(m, \rho(h), \rho(t)) = (N, n, n)$ and $(m, \rho(h), \rho(t)) = (R, y, n)$ as optimal, because, in this case, (11) and (12) both hold with equality. For any randomization over the agent's decisions, the principal expects a utility that is a convex combination of $V_L(c_\varepsilon^{SB})$ and $\frac{1}{2}V_L^{FI}(U^0 - \Delta U - 2\varepsilon) + \frac{1}{2}V_L^{IC}(U^0 + \Delta U)$ for $x = 0$, and of $V_H(c_\varepsilon^{SB})$ and $\frac{1}{2}V_H^{FI}(U^0 - \Delta U) + \frac{1}{2}V_H^{IC}(U^0 + \Delta U)$ for $x = 1$.

3. For $\hat{U}_1^r \in (U^0 - \Delta U, U^0 + \Delta U)$ or $\hat{U}_0^r \in (U^0 - \Delta U - 2\varepsilon, U^0 + \Delta U)$, both (11) and (12) are violated for $(\hat{U}_H^r, \hat{U}_L^r) = (\hat{U}_1^r, \hat{U}_0^r)$ so that the principal expects the agent to consider only $(m, \rho(h), \rho(t)) = (R, y, n)$ optimal. Hence, the principal expects the utility $\frac{1}{2}V_H^{FI}(\hat{U}_1^r) + \frac{1}{2}V_H^{IC}(U^0 + \Delta U)$ for $x = 1$, and utility $\frac{1}{2}V_L^{FI}(\hat{U}_0^r) + \frac{1}{2}V_L^{IC}(U^0 + \Delta U)$ for $x = 0$.

4. For $\hat{U}_1^r = U^0 + \Delta U$ or $\hat{U}_0^r = U^0 + \Delta U$, the principal expects the agent to consider exactly the three strategies $(m, \rho(h), \rho(t)) = (N, y, y)$, $(m, \rho(h), \rho(t)) = (R, y, y)$, and $(m, \rho(h), \rho(t)) = (R, y, n)$ optimal. For any mixture over these strategies, the

principal obtains a convex combination between $V_H^{FI}(U^0 + \Delta U)$ and $\frac{1}{2}V_H^{FI}(U^0 + \Delta U) + \frac{1}{2}V_H^{IC}(U^0 + \Delta U)$ for $x = 1$; and between $V_L^{FI}(U^0 + \Delta U)$ and $\frac{1}{2}V_L^{FI}(U^0 + \Delta U) + \frac{1}{2}V_L^{IC}(U^0 + \Delta U)$ for $x = 0$.

5. For $\hat{U}_e^r > U^0 + \Delta U$, the principal expects the agent to consider exactly strategies $(m, \rho(h), \rho(t)) = (N, y, y)$ and $(m, \rho(h), \rho(t)) = (R, y, y)$ optimal. For any mixture over these strategies, the principal obtains $V_H^{FI}(\hat{U}_1^r)$ for $x = 1$ and $V_L^{FI}(\hat{U}_0^r)$ for $x = 0$.

The analysis above implies that, with belief $x = 1$ ~~and since Φ is strictly increasing~~, the following inequalities guarantee that the principal expects to be strictly worse off from every renegotiation offer that the agent accepts with a strictly positive probability:

$$V_H(c_\varepsilon^{SB}) - \frac{1}{2}V_H^{FI}(U^0 - \Delta U) - \frac{1}{2}V_H^{IC}(U^0 + \Delta U) > 0, \tag{13}$$

and

$$V_H(c_\varepsilon^{SB}) - V_H^{FI}(U^0 + \Delta U) > 0. \tag{14}$$

Observe that, if $\varepsilon = 0$, (13) and (14) are strictly satisfied because they coincide with (8) and (10), respectively. Since $V_H(c_\varepsilon^{SB})$ is continuous in ε, there is a $\varepsilon^H > 0$ such that (13) and (14) are strictly satisfied for any $\varepsilon \in (0, \varepsilon^H)$. If, instead, $x = 0$, the principal believes to be strictly worse off from the agent accepting a renegotiation offer with a strictly positive probability when

$$V_L(c_\varepsilon^{SB}) - \frac{1}{2}V_L^{FI}(U^0 - \Delta U - 2\varepsilon) - \frac{1}{2}V_L^{IC}(U^0 + \Delta U) > 0 \tag{15}$$

and

$$V_L(c_\varepsilon^{SB}) - V_L^{FI}(U^0 + \Delta U) > 0. \tag{16}$$

Again, since $V_L(c_\varepsilon^{SB})$ is continuous in ε, there is a $\varepsilon^L > 0$ such that (15) and (16) are strictly satisfied for any $\varepsilon \in (0, \varepsilon^L)$. Defining $\bar{\varepsilon} \equiv \min\{\varepsilon^L, \varepsilon^H\}$ implies that if the principal holds a degenerate belief, then, for any $\varepsilon \in (0, \bar{\varepsilon})$, he believes that he is strictly worse off from a renegotiation offer that the agent accepts with a strictly positive probability.

We next argue that the polar cases $x \in \{0, 1\}$ as studied above imply that, also for an intermediate belief $x \in (0, 1)$, the principal expects to be strictly worse off from the agent accepting a renegotiation offer with strictly positive probability. To see this, note that the principal's expected utility by not renegotiating is linear in x:

$$V_x(c_\varepsilon^{SB}) = xV_H(c_\varepsilon^{SB}) + (1 - x)V_L(c_\varepsilon^{SB}),$$

since, regardless of her previous effort, the unique optimal report of the agent when $\gamma^r = \emptyset$ is $m = N$, inducing the transfers c_ε^{SB}.

Moreover, note that by offering $\gamma^r \neq \emptyset$, fixing any sequentially rational behavior $\lambda(\gamma_\varepsilon)$ by the agent and denoting $V_e^*(\gamma^r, \lambda(\gamma_\varepsilon))$ the principal's expected equilibrium utility in the continuation of $(\gamma_\varepsilon, e, \gamma^r)$, he would instead get

$$V_x^*(\gamma^r, \lambda(\gamma_\varepsilon)) \equiv xV_H^*(\gamma^r, \lambda(\gamma_\varepsilon)) + (1-x)V_L^*(\gamma^r, \lambda(\gamma_\varepsilon)).$$

As the agent's behavior is independent of the principal's belief x, this is also linear in x.

We have already established that, for any sequentially rational behavior of the agent, $V_e(c_\varepsilon^{SB}) > V_e^*(\gamma^r, \lambda(\gamma_\varepsilon))$ holds if x is degenerate on $e \in E$. To extend $V_x(c_\varepsilon^{SB}) > V_x^*(\gamma^r, \lambda(\gamma_\varepsilon))$ to intermediate beliefs $x \in (0,1)$, observe that under such beliefs the principal can offer a screening menu $\gamma^r \in C$ with $\gamma^r(H) \neq \gamma^r(L)$. For any such γ^r, the agent with effort e optimally reports in γ^r to obtain $\hat{U}_e^r = \max_{m^r \in E} U_e(\gamma^r(m^r))$ upon acceptance. Her optimal participation decisions ρ in γ^r and her optimal reporting decisions m in γ_ε vary with $\hat{U}_e^r$ in the same way as under degenerate beliefs. Hence the inequality $V_e^*(\gamma^r, \lambda(\gamma_\varepsilon)) < V_e(c_\varepsilon^{SB})$ holds for each $e \in E$, and taking the x-weighted average yields $V_x(c_\varepsilon^{SB}) > V_x^*(\gamma^r, \lambda(\gamma_\varepsilon))$ for every $x \in [0,1]$. Any renegotiation offer $\gamma^r \neq \emptyset$ is therefore ~~To see why $V_x(c_\varepsilon^{SB}) > V_x^*(\gamma^r, \lambda(\gamma_\varepsilon))$ extends to intermediate beliefs, observe that under $x \in (0,1)$ the principal could optimally offer a screening menu $\gamma^r \in C$ with $\gamma^r(H) \neq \gamma^r(L)$. For any γ^r as such, the agent with effort e, after accepting renegotiation, optimally reports in γ^r to obtain $\hat{U}_e^r = \max_{m^r \in E} U_e(\gamma^r(m^r))$. Also, her optimal participation decisions ρ in γ^r, and her optimal reporting decisions m in γ_ε vary according to $\hat{U}_e^r$ in the same way as already characterized under degenerate beliefs. Hence, screening does not improve the principal's utility: $V_e^*(\gamma^r, \lambda(\gamma_\varepsilon)) < V_e(c_\varepsilon^{SB})$ is verified for each $e \in E$ under every sequentially rational behavior of the agent, regardless of x being non-degenerate, and the inequality $V_x(c_\varepsilon^{SB}) > V_x^*(\gamma^r, \lambda(\gamma_\varepsilon))$ holds for every $x \in [0,1]$. This renders any renegotiation attempt $\gamma^r \neq \emptyset$~~ unprofitable under every belief of the principal.

Let us now consider the agent's choice of effort in the subgame induced by γ_ε. From the previous considerations, at any equilibrium of $G_\Gamma(\gamma_\varepsilon)$, the agent must anticipate, when selecting her effort, that the principal does not make a renegotiation offer inducing acceptance with positive probability for both $e \in E$. As already argued, the agent's unique optimal report in this case is $m = N$ regardless of e. This leads to the implementation of transfers c_ε^{SB}, which satisfy $U_H(c_\varepsilon^{SB}) > U_L(c_\varepsilon^{SB})$ by construction. Hence, by anticipating this, the agent finds $e = H$ her only optimal effort decision. But then, at any equilibrium of $G_\Gamma(\gamma_\varepsilon)$, the agent selects $x = 1$ and no renegotiation takes place. Hence, the principal's unique equilibrium utility in $G_\Gamma(\gamma_\varepsilon)$ is $V_H(c_\varepsilon^{SB})$. Equilibrium existence in the subgame is ensured by the fact that $e = H$, $m = N$ and $\gamma^r = \emptyset$ are optimal behaviors on the equilibrium path.

We now turn to the entire game G_Γ. Note first that, once the principal offers γ_ε,

the agent is indifferent between accepting it or not. Standard tie-breaking arguments, however, guarantee that the only participation decision consistent with equilibrium is acceptance.[31] Consequently, in any equilibrium of G_Γ, the principal must obtain at least the utility V^{SB}: any inferior utility $V' < V^{SB}$ is not sequentially rational since the principal could deviate to some γ_ε and uniquely obtain $V_H(c_\varepsilon^{SB}) \in (V', V^{SB})$. The existence of an appropriate γ_ε is guaranteed for any choice of V' since

$$\lim_{\varepsilon \to 0} V_H(c_\varepsilon^{SB}) = V^{SB}.$$

Given that the principal cannot obtain more than V^{SB} (the full-commitment upper bound), every equilibrium of G_Γ yields the principal a utility of exactly V^{SB}. It remains to show that the equilibrium *allocation* is unique. In the static second-best problem, V^{SB} is achieved only when both (IC) and (PC) bind with $e = H$, which pins down the contract as $c^{SB} = c^{IC}(U^0)$. This characterization extends to G_Γ: any mechanism $\gamma \in \Gamma$ achieving principal utility V^{SB} must (i) implement $e = H$ with probability one (since $V^{SB} = V_H^{IC}(U^0) > V_L^{FI}(U^0)$ by the maintained assumption that high effort is optimal in the second-best), (ii) leave the agent exactly U^0 (since V_H^{IC} is strictly decreasing), and (iii) execute transfers c^{SB} on path (since c^{SB} is the unique incentive-compatible contract for $e = H$ at U^0). Hence the equilibrium allocation (H, c^{SB}) is unique. ■

Proof of Proposition 3. We start by specifying the transfers implemented by ξ^{0*}. They are defined by the sequence of decision rules $\tau^{0*} = (\tau_{T'}^{0*})_{T' \geq 1}$. Each mapping $\tau_{T'}^{0*}$ associates any history of communication between the agent and ξ^{0*} from $T = 1$ to $T = T'$, which we denote $Z_{T'}^{0*} \in \{N, R\}^{T'} \times \{h, t\}^{T'}$, to the transfers $\tau_{T'}^{0*}(Z_{T'}^{0*}) \in \mathbb{R}^2$ to be paid if renegotiation breaks down at $T^* = T'$. Specifically, for any $Z_{T'}^{0*}$ with $T' \geq 1$, we let s^R be the signal extracted in the first round in which R is reported by the agent. Then:

$$\tau_T^{0*}(Z_T^{0*}) = \begin{cases} c^{SB} & \text{if } R \notin Z_T^{0*}, \\ c_H^{FI}\left(U^0 - \Delta U - d\right) & \text{if } R \in Z_T^{0*} \text{ and } s^R = h, \\ c_H^{FI}(U^0 + \Delta U - d) & \text{if } R \in Z_T^{0*} \text{ and } s^R = t, \end{cases}$$

where ΔU is such that:[32]

$$\Delta U > d \quad \text{and} \quad \frac{1}{2} V_H^{FI}\left(U^0 - \Delta U - d\right) + \frac{1}{2} V_H^{FI}(U^0 + \Delta U - d) < V^{SB}. \tag{17}$$

Here, the terms $U^0 - \Delta U - d$ and $U^0 + \Delta U - d$ in the punishment lottery are chosen so that, under $e = L$, the agent's saving of the effort cost d is exactly offset. Indeed, since

[31] One can construct another tie-breaking mechanism identical to γ_ε except for yielding $U^0 + \varepsilon$ to the agent if she accepts.

[32] Existence of ΔU satisfying (17) follows from continuity and the fact that, for large ΔU, the LHS in the second inequality diverges to $-\infty$ while V^{SB} is finite. The condition $\Delta U > d$ is then satisfiable for ΔU in an appropriate range.

$U_L(c_H^{FI}(U)) = U + d$, the two utilities generated by the punishment lottery under $e = L$ are

$$U^0 - \Delta U \quad \text{and} \quad U^0 + \Delta U.$$

Hence, under $e = L$, the expected utility from reporting R is exactly U^0. That is, reporting R is not profitable relative to reporting N for each effort level; the agent strictly prefers N when $e = H$ and is indifferent when $e = L$.

At each round T, ξ^{0*} requires the agent to either send a status-quo report N or a punishment report R. The decision rule τ_T^{0*} is such that, as long as R is *not* reported, the second-best transfers c^{SB} are implemented. If R is sent at some round T, then the random punishment associated with τ_T^{0*} is implemented and all future reports become payoff-irrelevant. Observe, in addition, that the report R induces a lottery over *first-best* efficient contracts, whose outcomes are therefore not improvable by any renegotiation.

The agent's strategies in $G_\Xi^\eta(\xi^{0*})$. At $T = 0$, the agent chooses $e \in \{H, L\}$. Then, for any $T' \geq 1$, the agent's histories have a recursive structure. At round $T'.ii)$, she makes a report in the last accepted mechanism, which we denote $m_{T'}^{0*} \in \{N, R\}$ (if this is ξ^{0*}) or $m_{T'}^T \in \{N, R\}$ (if this is ξ^T with $T < T'$), and she hence observes either $s_{T'}^{0*} \in \{h, t\}$ or $s_{T'}^T \in \{h, t\}$. Then, at stage $T'.iii)$, she selects $\rho^{T'} \in \{y, n\}$, and, if $\rho^{T'} = y$, she reports $\hat{e}^{T'} \in \{H, L\}$ in $\xi^{T'}$ in round $T'.iv)$. We denote $\mathcal{H}_A^{T'+1}$ a history of the agent up to $T'.iv)$.

The principal's strategies in $G_\Xi^\eta(\xi^{0*})$. The principal may attempt to renegotiate the mechanism ξ^{0*} at any round $T \geq 1$, until T^* realizes. A renegotiated mechanism ξ^T offered at round T requires the agent to submit a report $\hat{e}^T \in \{H, L\}$ at $T.iv)$ after accepting it. Further, at any $T' > T$ where ξ^T is still active, the mechanism requires the agent to send a report $m_{T'}^T \in \{N, R\}$ while privately disclosing the realization $s_{T'}^T \in \{h, t\}$ of a fair coin toss to her. The sequence of decision rules $\tau^T = (\tau_{T'}^T)_{T' \geq T}$ maps this communication to transfers, with $\tau_{T'}^T$ being the rule for round $T' \geq T$. This rule associates any sequence of reports and signals $Z_{T'}^T \in \{H, L\} \times \{N, R\}^{T'-T-1} \times \{h, t\}^{T'-T-1}$ exchanged between the agent and ξ^T up to round T' to the transfers $\tau_{T'}^T(Z_{T'}^T) \in \mathbb{R}^2$ to be paid if renegotiation breaks down at $T^* = T'$.

We denote $\mathcal{H}_P^1$ the initial history of the principal at $T = 1$. We then let $\mathcal{H}_P^T \equiv (\xi^1, \rho^1, \ldots, \xi^{T-1}, \rho^{T-1})$ be a principal's history at the end of stage $T - 1$. Thus, a (pure) behavioral strategy of the principal in $G_\Xi^\eta(\xi^{0*})$ associates with each history $\mathcal{H}_P^T$, for all $T \geq 1$, a renegotiated offer $\xi^T \in \Xi \cup \{\emptyset\}$.

Furthermore, we denote $\mathcal{P}_T^{0*}$ the set of principal's histories, and $\mathcal{A}_T^{0*}$ the set of agent's histories, such that $\rho^{T'} = n$ for all $T' : 1 \leq T' < T$. At any such history, ξ^{0*} is still in place at round T. At any history $\mathcal{H}_P^T \in \mathcal{P}_T^{0*}$, the principal may either offer a mechanism ξ^T or decide not to renegotiate. At any history $(\mathcal{H}_A^T, \xi^T)$ with $\mathcal{H}_A^T \in \mathcal{A}_T^{0*}$, the agent reports $m_T^{0*} \in \{N, R\}$ in ξ^{0*} and privately observes the signal $s_T^{0*} \in \{h, t\}$; then, at any history

$(\mathcal{H}_A^T, \xi^T, m_T^{0*}, s_T^{0*})$ she selects $\rho^T \in \{y, n\}$ and, at any history $(\mathcal{H}_A^T, \xi^T, m_T^{0*}, s_T^{0*}, y)$, she selects $\hat{e}^T \in \{H, L\}$.

We next construct players' equilibrium strategies supporting the second-best allocation and verify their sequential rationality. The proof is developed in three steps.

Step 1. Equilibrium strategies and beliefs. We first describe the agent's equilibrium behavior in $G_\Xi^\eta(\xi^{0*})$. At $T = 0$ she takes $e = H$ with probability one. Then, we explicitly characterize her reporting and participation behavior only at the histories $(\mathcal{H}_A^T, \xi^T)$ such that $e = H$ and $\mathcal{H}_A^T \in \mathcal{A}_T^{0*}$. Relative to all other histories, we only require that the agent behaves in a sequentially rational way given the principal's equilibrium behavior.

Consider any history $(\mathcal{H}_A^T, \xi^T, m_T^{0*}, s_T^{0*}, y)$ with $\mathcal{H}_A^T \in \mathcal{A}_T^{0*}$, in which $(m_T^{0*}, s_T^{0*}) \in \{N, R\} \times \{h, t\}$ is the communication entertained by the agent with the mechanism ξ^{0*} at time T. Before constructing the agent's strategy, observe that her continuation utility corresponding to $(\mathcal{H}_A^T, \xi^T, m_T^{0*}, s_T^{0*}, y)$ must be independent of her communication (m_T^{0*}, s_T^{0*}) within ξ^{0*}. This follows from the facts that the renegotiation offer ξ^T cannot condition on this private communication, and that any accepted renegotiation offer must replace the original one rendering it payoff-irrelevant. We denote this utility by $\hat{U}_H^{0*}(\xi^T)$. For a given offer ξ^T, we specify the agent's equilibrium behavior only in terms of $\hat{U}_H^{0*}(\xi^T)$.

To construct the agent's reporting and participation behavior, we distinguish two mutually exclusive cases, according to the relevant round T and the history of the communication Z_{T-1}^{0*} between the agent and the original mechanism ξ^{0*} up to this round.

1. $T > 1$ and $R \in Z_{T-1}^{0*}$. That is, the history of communication within ξ^{0*} contains at least a report R. In any such case, we let the agent report $m_T^{0*} = R$ in ξ^{0*}. In addition, her participation decision depends on the signal s^R received from ξ^{0*} in the first round in which R was sent. Specifically:

- If $s^R = h$, then she selects $\rho^T = y$ iff $\hat{U}_H^{0*}(\xi^T) \geq U^0 - \Delta U - d$,
- If $s^R = t$, then she selects $\rho^T = y$ iff $\hat{U}_H^{0*}(\xi^T) \geq U^0 + \Delta U - d$.

2. Either $T = 1$, or $R \notin Z_{T-1}^{0*}$. In any such case, the agent's report m_T^{0*} in ξ^{0*} is determined as follows:

- She reports $m_T^{0*} = N$ if either $\hat{U}_H^{0*}(\xi^T) \leq U^0 - (\Delta U - d)$ or $\xi^T = \emptyset$,
- She reports $m_T^{0*} = R$ if $\hat{U}_H^{0*}(\xi^T) > U^0 - (\Delta U - d)$.

Finally, at any $(\mathcal{H}_A^T, \xi^T, m_T^{0*}, s_T^{0*})$, the agent's participation within ξ^T is such that:

- If $m_T^{0*} = N$, then, for all $s_T^{0*} \in \{h, t\}$, she selects $\rho^T = y$ iff $\hat{U}_H^{0*}(\xi^T) \geq U^0$,

- If $m_T^{0*} = R$ and $s_T^{0*} = h$, then she selects $\rho^T = y$ iff $\hat{U}_H^{0*}(\xi^T) \geq U^0 - \Delta U - d$,
- If $m_T^{0*} = R$ and $s_T^{0*} = t$, then she selects $\rho^T = y$ iff $\hat{U}_H^{0*}(\xi^T) \geq U^0 + \Delta U - d$.

We next specify the principal's equilibrium behavior and beliefs in $G_\Xi^\eta(\xi^{0*})$.

The principal chooses $\xi^T = \emptyset$ at any history $\mathcal{H}_P^T \in \mathcal{P}_T^{0*}$. Relative to all other histories, we only require that the principal behaves in a sequentially rational way given his beliefs and the agent's equilibrium behavior.

Concerning his beliefs, we assume that, at any on-the-equilibrium-path history $\mathcal{H}_P^T \in \mathcal{P}_T^{0*}$, the principal believes that $e = H$ with probability one, and that $m_{T'}^{0*} = N$ for all $T' < T$, while he assigns probability one-half to each $s_{T'}^{0*} \in \{h, t\}$. Thus, the principal's on-path beliefs are Bayes-consistent given the agent's behavior.

We also require that, at any history $\mathcal{H}_P^T \in \mathcal{P}_T^{0*}$ which is *off* the equilibrium path (i.e. the agent has rejected some previous non-empty renegotiation offers), the principal still believes that $e = H$ with probability one~~, and that~~ and his beliefs about the agent's past communications are consistent with the agent's equilibrium behavior (e.g. the principal believes $m_{T'}^{0*} = N$ at each previous round where $\xi^{T'} = \emptyset$). At all remaining histories, beliefs can be arbitrarily selected.

Step 2. The agent's sequential rationality. We establish the sequential rationality of the agent's effort and communication behavior.

—*Effort choice.* Given the principal's equilibrium behavior, choosing $e = H$ and reporting $m_T^{0*} = N$ at every $T \geq 1$ yields the agent her reservation utility U^0. Suppose, instead, that she takes $e = L$ at $T = 0$. Then, any subsequent reporting strategy yields her again U^0. Indeed, reporting $m_T^{0*} = N$ in ξ^{0*} at every $T \geq 1$ yields the second-best transfers c^{SB}. By reporting instead R in any round $T \geq 1$, the agent triggers the punishment lottery yielding her

$$\frac{1}{2}U_L(c_H^{FI}(U^0 - \Delta U - d)) + \frac{1}{2}U_L(c_H^{FI}(U^0 + \Delta U - d)) = \frac{1}{2}(U^0 - \Delta U) + \frac{1}{2}(U^0 + \Delta U) = U^0,$$

since $U_L(c_H^{FI}(U)) = U + d$. Thus, choosing $e = L$ does not constitute a profitable deviation.

—*Reporting and participation decisions.* Consider any agent's history $(\mathcal{H}_A^T, \xi^T)$ such that $\mathcal{H}_A^T \in \mathcal{A}_T^{0*}$ and $e = H$. Once again, we distinguish two mutually exclusive situations, according to the relevant round T and the history of the communication Z_{T-1}^{0*} between the agent and the original mechanism ξ^{0*} up to this round.

1. $T > 1$ and $R \in Z_{T-1}^{0*}$. That is, the history of communication within ξ^{0*} contains at least a report R. In any such case, given τ^{0*}, any agent's report from round T onwards in ξ^{0*} is payoff-irrelevant, guaranteeing the optimality of our constructed behavior. Concerning participation, rejecting a renegotiated offer ξ^T secures the agent a continuation utility of

either $U^0 - \Delta U - d$ (if $s^R = h$) or $U^0 + \Delta U - d$ (if $s^R = t$) given the principal's equilibrium behavior. This guarantees the optimality of our constructed participation behavior.

2. Either $T = 1$, or $R \notin Z^{0*}_{T-1}$. Consider first on-path histories, that is, any agent's history $(\mathcal{H}^T_A, \emptyset)$ such that $\mathcal{H}^T_A \in \mathcal{A}^{0*}_T$ and $e = H$. Given the principal's equilibrium behavior, by reporting $m^{0*}_T = N$ in ξ^{0*} and $m^{0*}_{T'} = N$ at every $T' > T$ the agent obtains the second-best transfers c^{SB}, which yield her the reservation utility U^0. By reporting $m^{0*}_T = R$, the agent triggers the punishment in ξ^{0*}, and gets the expected utility $U^0 - d < U^0$ regardless of her subsequent communication behavior. Thus, $m^{0*}_T = N$ is the unique optimal report.

Consider next any off-path agent's history $(\mathcal{H}^T_A, \xi^T)$ such that $\mathcal{H}^T_A \in \mathcal{A}^{0*}_T$ and $e = H$. The agent's constructed participation behavior can be straightforwardly verified to be sequentially rational by comparing, for each $(m^{0*}_T, s^{0*}_T) \in \{N, R\} \times \{h, t\}$, her utility of accepting ξ^T to that of holding ξ^{0*}. Concerning her reporting behavior, the following inequality guarantees that reporting $m^{0*}_T = N$ is optimal:

$$\max\{U^0, \hat{U}^{0*}_H(\xi^T)\} \geq \frac{1}{2}\max\{U^0 - \Delta U - d, \hat{U}^{0*}_H(\xi^T)\} + \frac{1}{2}\max\{U^0 + \Delta U - d, \hat{U}^{0*}_H(\xi^T)\}. \quad (18)$$

As the principal does not renegotiate at all $T' > T$, the LHS of (18) is the agent's utility from reporting N and following the constructed continuation behavior, and the RHS represents the expected utility from reporting R and following the constructed continuation behavior. One can then check that (18) holds whenever

$$\hat{U}^{0*}_H(\xi^T) \leq U^0 - (\Delta U - d),$$

so that reporting N is optimal in the region where the constructed strategy prescribes N. If instead

$$\hat{U}^{0*}_H(\xi^T) > U^0 - (\Delta U - d),$$

then reporting R is optimal: it is strictly optimal when

$$\hat{U}^{0*}_H(\xi^T) \in \big(U^0 - (\Delta U - d),\, U^0 + \Delta U - d\big),$$

and weakly optimal when

$$\hat{U}^{0*}_H(\xi^T) \geq U^0 + \Delta U - d.$$

Hence the constructed reporting strategy is sequentially rational.

Step 3. The principal's sequential rationality. We now verify the optimality of the principal's behavior where explicitly characterized. We start from the principal's histories on-the-equilibrium-path. That is, we take any history $\mathcal{H}^T_P \in \mathcal{P}^{0*}_T$ such that $\xi^{T'} = \emptyset$ for any $T' : 1 \leq T' < T$. In any such history, the principal holds the Bayes-consistent belief that the agent has reported $m^{0*}_{T'} = N$ in ξ^{0*} in any $T' < T$.

To verify that the principal does not gain by offering $\xi^T \neq \emptyset$ at round T, we distinguish two cases according to the value $\hat{U}_H^{0*}(\xi^T)$.[33]

1. $\hat{U}_H^{0*}(\xi^T) \leq U^0 - (\Delta U - d)$. The agent's equilibrium behavior prescribes to report $m_T^{0*} = N$ in ξ^{0*} after observing ξ^T, and to reject it for any $s_T^{0*} \in \{h, t\}$. Hence, any such ξ^T offer is payoff-equivalent to $\xi^T = \emptyset$ for the principal.

2. $\hat{U}_H^{0*}(\xi^T) > U^0 - (\Delta U - d)$. The agent's equilibrium behavior prescribes to report $m_T^{0*} = R$ in ξ^{0*}. This guarantees her the utility $U^0 - \Delta U - d$ (if $s_T^{0*} = h$) or $U^0 + \Delta U - d$ (if $s_T^{0*} = t$), which can be achieved by selecting $\rho^T = n$ in any history $(\mathcal{H}_T^A, \xi^T, R, s_T^{0*})$. Thus, the principal's continuation utility is at most

$$\frac{1}{2}V_H^{FI}(U^0 - \Delta U - d) + \frac{1}{2}V_H^{FI}(U^0 + \Delta U - d) < V^{SB},$$

where the inequality follows from (17). Hence, any such deviation is unprofitable to the principal.

Finally, consider any off-the-equilibrium path history $\mathcal{H}_P^T \in \mathcal{P}_T^{0*}$. In any such situation, two cases can be distinguished. First, the principal believes that the agent has always reported $m_{T'}^{0*} = N$ in all past $T' < T$. Then, regardless of all past signals realizations, the reasoning guaranteeing that $\xi^T = \emptyset$ is sequentially optimal on the equilibrium path extends readily, as the two situations are strategically identical from the principal's perspective given the agent's equilibrium behavior. Second, the principal believes that $m_{T'}^{0*} = R$ has been reported by the agent in ξ^{0*} at some $T' < T$, and the corresponding signal is $s^R \in \{h, t\}$. In that case, ξ^{0*} already implements a full-insurance first-best contract: $c_H^{FI}(U^0 - \Delta U - d)$ if $s^R = h$, and $c_H^{FI}(U^0 + \Delta U - d)$ if $s^R = t$. Since any accepted renegotiation must give the agent at least the continuation utility secured by the existing contract, and since the existing contract is already first-best efficient conditional on that utility, renegotiation cannot strictly improve the principal's payoff. Hence offering $\xi^{T''} = \emptyset$ is sequentially rational at any $T'' \geq T$.

Therefore, deviations starting at any $\mathcal{H}_P^T \in \mathcal{P}_T^{0*}$, on and off the equilibrium path, yield weakly less than V^{SB} to the principal in the continuation of $(\mathcal{H}_P^T, \xi^T)$, given his beliefs and the agent's equilibrium behavior. This guarantees that the principal's equilibrium strategy of offering $\xi^T = \emptyset$ at any such history is sequentially rational. ■

[33]Since the non-profitability of a deviation ξ^T, given the agent's equilibrium behavior, does not depend on the principal's continuation play, the analysis below guarantees that *all* deviations starting at $\mathcal{H}_T^P$ are unprofitable for the principal, not only the one-shot deviations.

8 Smart Contract Implementation

We present, as a proof-of-concept, a fully specified example of a smart contract for a parameterized version of our framework using the commit-and-reveal technique. In particular, let the normalized CRRA utility function $u(w) = \sqrt{w}$ describe the agent's preferences over transfers, implying that the monetary equivalent is $\Phi(u) = u^2$. Let $U^0 = 10$ be the agent's reservation utility. The cost of high effort is $d = 2$ with success probability $p_H = 3/4$, while for low effort the probability is $p_L = 1/4$, i.e., $\Delta p = 1/2$. The good output is $g = 1300$, while the bad output is $b = 100$. Hence, $y_H = 1000$ and $y_L = 400$.

It is easy to check that $\Delta U = 2$ together with the parameterized example satisfies (2), yields the self-revealing mechanism γ^{**} with transfers (in monetary terms, i.e., $w = \Phi(u) = u^2$)

$$\begin{array}{ll} \tau^{**}(N,h) = (169,81); & \tau^{**}(N,t) = (169,81) \\ \tau^{**}(R_1,h) = (121,49); & \tau^{**}(R_1,t) = (225,121) \\ \tau^{**}(R_2,h) = (225,121); & \tau^{**}(R_2,t) = (121,49). \end{array}$$

Figure 1 presents the smart contract that implements γ^{**} over the Ethereum blockchain using the commit-and-reveal technique.[34] The smart contract is written in Solidity, the most common language for Ethereum smart contracts.

To allow the agent to send a secret (hashed) message $m \in \{N, R_1, R_2\}$ with a random seed σ, the smart contract implements the commit-and-reveal technique as previously discussed, based on the public keccak-256 hash function.

After sending the hashed message, the agent waits for the principal to report the realized output level $Y \in \{g, b\}$, at which point the smart contract generates the signal $s \in \{h, t\}$ in a random fashion by recording the realized signal publicly on the blockchain. Finally, the agent is to report the seed σ to the smart contract by which the smart contract can recover the original message m so that it can make the transfers according to τ^{**}.

We set up the contract such that if the agent does not reveal the seed σ honestly, this is interpreted as tearing up the original contract and accepting a renegotiated one, $(\rho = y)$, so that the smart contract stops in that no transfers flow and message m stays hidden. This "waiting indefinitely" behavior faithfully implements the paper's framework, where accepted renegotiation causes the original mechanism to simply stop executing, with transfers flowing instead through the renegotiated contract.

[34] The contract is a minimal proof-of-concept only. It is intentionally not security-hardened. Concretely, it uses a placeholder public coin S (not a verifiable randomness source), does not gate reveal on a recorded renegotiation outcome, does not escrow funds or enforce deadlines/liveness, and accepts Y from the principal without authenticated reporting (relying instead on off-chain legal enforceability). The numeric transfer constants represent wages $w = u^2$ consistent with the utility table, expressed in Ether units. These simplifications are deliberate and solely for illustrating the interface and timing pattern (commit privately; reveal only at enforcement). A production deployment would replace each placeholder with its standard counterpart (verifiable randomness or two-party coin-toss, renegotiation-gated reveal/state machine, escrow with deadlines and fallbacks, authenticated Y reporter or explicit legal backstop).

```
pragma solidity ^0.8.0;
contract CommitRevealTransfer {
  address constant AddressP = 0x362CbcC7a9955332e61d47c107543398C3D25261;
  address constant AddressA = 0x818CbcC8de183AED16f850B17c300DB40a4544Eb;
  uint256 constant TG=169; uint256 constant TGH=121; uint256 constant TGT=225;
  uint256 constant TB=81; uint256 constant TBH=49; uint256 constant TBT=121;
  bytes32 public HASHCOMMIT;  string public S;  string public Y;
  bool public isCommitted;  bool public isRevealed;  bool public isYSent;
  constructor() {
    require(msg.sender==AddressP, "Only AddressP can deploy");}
  function commit(bytes32 _hashCommit) external {
    require(msg.sender==AddressA, "Only AddressA can commit");
    require(!isCommitted, "Already committed");
    HASHCOMMIT = _hashCommit;  isCommitted = true; }
  function generateS() internal {
    require(isCommitted, "Waiting for commit");
    S = block.timestamp % 2==0?"Head" : "Tail"; }
  function sendY(string calldata _Y) external {
    require(msg.sender==AddressP, "Only AddressP can send Y");
    require(isCommitted, "Waiting for commit");
      require(keccak256(abi.encodePacked(_Y))==keccak256(abi.encodePacked("G")) ||
          keccak256(abi.encodePacked(_Y))==keccak256(abi.encodePacked("B")),"Only G/B");
    Y = _Y;   isYSent = true;   generateS();
    }
  function reveal(string calldata _message, string calldata _salt) external {
    require(msg.sender==AddressA, "Only AddressA can reveal");
    require(isYSent, "Waiting for Y");
    require(!isRevealed, "Already revealed");
    require(keccak256(abi.encodePacked(_message, _salt))==HASHCOMMIT, "Invalid");
      require(keccak256(abi.encodePacked(_message))==keccak256(abi.encodePacked("N")) ||
          keccak256(abi.encodePacked(_message))==keccak256(abi.encodePacked("R1")) ||
          keccak256(abi.encodePacked(_message))==keccak256(abi.encodePacked("R2")), "
          Invalid message");
    isRevealed = true;  uint256 transferAmount = determineTransferAmount(_message);
    payable(AddressA).transfer(transferAmount);  }
  function determineTransferAmount(string memory _message) internal view returns (uint256
      ) {
    if (keccak256(abi.encodePacked(_message))==keccak256(abi.encodePacked("N"))) {
      return keccak256(abi.encodePacked(Y))==keccak256(abi.encodePacked("G"))?TG : TB;
    } else if (keccak256(abi.encodePacked(_message))==keccak256(abi.encodePacked("R1")))
        {
      if (keccak256(abi.encodePacked(Y))==keccak256(abi.encodePacked("G"))) {
        return keccak256(abi.encodePacked(S))==keccak256(abi.encodePacked("Head")) ? TGH
            : TGT;
      } else {
        return keccak256(abi.encodePacked(S))==keccak256(abi.encodePacked("Head")) ? TBH
            : TBT;
      }
    } else {
      if (keccak256(abi.encodePacked(Y))==keccak256(abi.encodePacked("G"))) {
        return keccak256(abi.encodePacked(S))==keccak256(abi.encodePacked("Head")) ? TGT
            : TGH;
      } else {
        return keccak256(abi.encodePacked(S))==keccak256(abi.encodePacked("Head")) ? TBT
            : TBH;
      }
    }
  }
  receive() external payable {require(msg.sender==AddressP, "Only AddressP can send");}
}
```

Figure 1: The smart contract implementing the self-revealing mechanism γ^{**} with a reveal-and-commit technique based on the keccak-256 hash function in Solidity.

9 Additional Results

This appendix develops several extensions.

Irrelevance of Random Mechanisms in Fudenberg and Tirole (1990)

We here formalize the claim that random mechanisms play no role in the FT construction. To achieve this task, we let $G_{\tilde{C}}$ be a game that enlarges the set of available mechanisms C to $\tilde{C}$ to include all stochastic mechanisms $\tilde{\gamma}^r : E \to \Delta(\mathbb{R}^2)$.

Lemma 4 *$G_{\tilde{C}}$ has only one equilibrium allocation, which coincides with that in G_C.*

Proof. For any $\tilde{\gamma} \in \tilde{C}$, define $\tilde{\gamma}(e) = \tilde{c}_e$ and let

$$\tilde{U}_e \equiv p_e \mathbb{E}[u_g|\tilde{c}_e] + (1-p_e)\mathbb{E}[u_b|\tilde{c}_e]$$

be the agent's expected utility after taking the effort $e \in E$, and truthfully reporting it in $\tilde{\gamma}$. Consider the subgame $G_{\tilde{C}}(\tilde{\gamma})$, and suppose that $e = H$ is chosen with probability $x \in [0,1]$. The revelation principle guarantees that the maximal utility attainable by the principal from a renegotiation offer $\tilde{\gamma}^r \in \tilde{C}$ is the value of the program $P(x, \tilde{U}_H, \tilde{U}_L)$:

$$\begin{aligned} V^*(x, \tilde{U}_H, \tilde{U}_L) = \max_{\tilde{\gamma} \in \tilde{C}} \quad & Y(x) - x[p_H \mathbb{E}(\Phi(u_g)|\tilde{c}^r_H) + (1-p_H)\mathbb{E}(\Phi(u_b)|\tilde{c}^r_H)] \\ & - (1-x)[p_L \mathbb{E}(\Phi(u_g)|\tilde{c}^r_L) + (1-p_L)\mathbb{E}(\Phi(u_b)|\tilde{c}^r_L)] \end{aligned} \tag{19}$$

$$\begin{aligned} \text{s.t.:} \quad & p_H \mathbb{E}(u_g|\tilde{c}^r_H) + (1-p_H)\mathbb{E}(u_b|\tilde{c}^r_H) \geq \tilde{U}_H && (IRC_H) \\ & p_L \mathbb{E}(u_g|\tilde{c}^r_L) + (1-p_L)\mathbb{E}(u_b|\tilde{c}^r_L) \geq \tilde{U}_L && (IRC_L) \\ & p_H \mathbb{E}(u_g|\tilde{c}^r_H) + (1-p_H)\mathbb{E}(u_b|\tilde{c}^r_H) \geq p_H \mathbb{E}(u_g|\tilde{c}^r_L) + (1-p_H)\mathbb{E}(u_b|\tilde{c}^r_L) && (ICC_H) \\ & p_L \mathbb{E}(u_g|\tilde{c}^r_L) + (1-p_L)\mathbb{E}(u_b|\tilde{c}^r_L) \geq p_L \mathbb{E}(u_g|\tilde{c}^r_H) + (1-p_L)\mathbb{E}(u_b|\tilde{c}^r_H) && (ICC_L) \end{aligned}$$

where $Y(x) = xY_H + (1-x)Y_L$. The following two results hold:

Claim 1 *$P(x, \tilde{U}_H, \tilde{U}_L)$ admits a unique solution, which is deterministic.*

Proof. See Chade and Schlee (2012, Proposition 1). ■

Denote $\gamma^r(\tilde{\gamma}, x)$ the unique solution of $P(x, \tilde{U}_H, \tilde{U}_L)$.

Claim 2 *For any $\tilde{\gamma} \in \tilde{C}$ and $x \in [0,1]$ there is a $\gamma_{\tilde{\gamma}} \in C$ such that $\gamma^r(\tilde{\gamma}, x) = \gamma^r(\gamma_{\tilde{\gamma}}, x)$.*

Proof. Given $\tilde{\gamma} \in \tilde{C}$, we take the mechanisms $\gamma_{\tilde{\gamma}} \in C$ yielding the transfers $U^e_\omega = \mathbb{E}(u_\omega|\tilde{c}_e)$ for each $(e, \omega) \in E \times \{g, b\}$. Thus, for any $x \in [0,1]$, the optimal renegotiation offer in $G_C(\gamma_{\tilde{\gamma}})$ obtains again from solving $P(x, \tilde{U}_H, \tilde{U}_L)$. ■

Given $\gamma_{\tilde{\gamma}}$, the following holds:

Claim 3 *The subgames $G_{\tilde{C}}(\tilde{\gamma})$ and $G_C(\gamma_{\tilde{\gamma}})$ have the same equilibrium allocations.*

Proof. Consider $G_{\tilde{C}}(\tilde{\gamma})$, and let $x \in [0,1]$ be an equilibrium effort distribution. Given Claim 2, the optimal renegotiation offer is $\gamma^r(\tilde{\gamma}, x) = \gamma^r(\gamma_{\tilde{\gamma}}, x)$, which is accepted by the agent, who truthfully reports her effort.[35] Furthermore, the transfers corresponding to the unique solution of $P(x, \tilde{U}_H, \tilde{U}_L)$ are implemented. Thus, playing $e = H$ with probability $x \in [0,1]$ is sequentially rational for the agent in $G_{\tilde{C}}(\tilde{\gamma})$, together with the principal's optimal renegotiation, if and only if they are also sequentially rational in $G_C(\gamma_{\tilde{\gamma}})$. This guarantees that the two subgames have the same equilibrium allocations. ■

To conclude the proof, denote x^{FT} the equilibrium probability of $e = H$ characterized by FT, and U^{FT} the equilibrium rent of the agent. Claim 3 implies that the upper bound $V^{FT} = V^*(x^{FT}, U^{FT}, U^{FT})$ of the principal's utilities characterized by FT in G_C is also an upper bound in $G_{\tilde{C}}$. In the latter game, the principal can achieve V^{FT} as the unique continuation utility by offering any of the mechanisms characterized in Fudenberg and Tirole (1990, Proposition 3.4). Thus, the unique equilibrium's utility of the principal in $G_{\tilde{C}}$ is V^{FT}, and the same distributions over efforts and transfers are implemented. ■

The Case of Bounded Transfers

Let the agent's utility over monetary transfers exhibit constant relative risk aversion (CRRA) structure:

$$u(w) = \frac{w^\alpha}{\alpha},$$

with CRRA parameter $\alpha \in (0,1)$. The function u has domain $[0,\infty)$ and range $[0,\infty)$; hence, its inverse $\Phi(u) = (\alpha u)^{\frac{1}{\alpha}}$ has domain $[0,\infty)$ coinciding with the range of u. The requirement that monetary transfers be non-negative imposes a form of limited liability for the agent. At the same time, this assumption renders unfeasible those mechanisms that rely on "extreme" transfers to punish the principal's attempts to renegotiate, as it may be the case for the mechanism γ^* constructed in Section 3.

We now show that our implementation result also obtains in this context. Specifically, we first establish an analogue of Lemma 1 for CRRA preferences, and then exploit it to argue that a slightly modified version of the mechanism γ^* allows to implement the second-best allocation. In developing our analysis, we focus on situations in which the restriction on transfers does not affect the agent's incentives to undertake her efficient level of effort. That is, we let

$$U^0 > U^\ell \equiv \frac{p_L}{\Delta p} d, \tag{20}$$

[35] See Fudenberg and Tirole (1990, p. 1295).

which is necessary and sufficient to guarantee that the second-best allocation involves strictly positive transfers in each state.[36] Given (1), we therefore have $c^{SB} = \left(U^0 + \frac{1-p_L}{\Delta p}d,\, U^0 - \frac{p_L}{\Delta p}d\right)$.

We can now establish the following:

Lemma 5 *If the agent's preferences are such that* $\Phi(u) = (\alpha u)^{\frac{1}{\alpha}}$ *with* $\alpha \in (0,1)$ *and* (20) *holds, then there is a* $\pi \in (0,1)$ *such that, for each* $e \in E$*:*

$$V_e^{IC}(U^0) > \max\left\{V_e^{FI}\left(\frac{U^0 - U^\ell}{\pi} + U^\ell\right),\, (1-\pi)V_e^{FI}(U^\ell) + \pi V_e^{IC}\left(\frac{U^0 - U^\ell}{\pi} + U^\ell\right)\right\}. \tag{21}$$

Proof. For a given $e \in E$, define the function $\hat{V}_e : (0,1) \to \mathbb{R}$ as

$$\hat{V}_e(\pi) \equiv (1-\pi)V_e^{FI}(U^\ell) + \pi V_e^{FI}\left(\frac{U^0 - U^\ell}{\pi} + U^\ell\right).$$

Note that $\hat{V}_e$ is defined and continuous for all $\pi \in (0,1)$. We now argue that:

$$\lim_{\pi\to 0}\hat{V}_e(\pi) = Y_e - \Phi(U^\ell + D(e)) - \lim_{\pi\to 0}\frac{\Phi\left(\frac{U^0-U^\ell}{\pi} + U^\ell + D(e)\right)}{\frac{1}{\pi}} = -\infty. \tag{22}$$

To see why (22) holds, simplify the last term as:

$$\lim_{\pi\to 0}\frac{\Phi\left(\frac{U^0-(1-\pi)U^\ell}{\pi} + D(e)\right)}{\frac{1}{\pi}} = \lim_{\pi\to 0}\frac{\Phi\left(\frac{U^0-U^\ell}{\pi} + U^\ell + D(e)\right)}{\frac{U^0-U^\ell}{\pi} + U^\ell + D(e)}\cdot\frac{\frac{U^0-U^\ell}{\pi} + U^\ell + D(e)}{\frac{1}{\pi}} = \lim_{u'\to\infty}\frac{\Phi(u')}{u'}\cdot(U^0-U^\ell)$$

under the change of variable $u' \equiv \frac{U^0-U^\ell}{\pi} + U^\ell + D(e)$. Since $U^0 > U^\ell$ by (20), and since $\alpha \in (0,1)$,

$$\lim_{u'\to\infty}\frac{\Phi(u')}{u'}\cdot(U^0-U^\ell) = \lim_{u'\to\infty}(u')^{\frac{1-\alpha}{\alpha}}\cdot(U^0-U^\ell) = \infty,$$

which implies (22). Thus, for each $e \in E$ and each constant $\kappa \in \mathbb{R}$, there exist $\delta_e(\kappa) \in (0,1)$ such that $\hat{V}_e(\pi) < \kappa$ for all $\pi \in (0, \delta_e(\kappa))$. Let $\overline{\pi}_e \equiv \delta_e(V_e^{IC}(U^0))$ for all $e \in E$. Then,

$$V_e^{IC}(U^0) > \hat{V}_e(\pi) \quad \forall \pi \in (0, \overline{\pi}_e).$$

It follows that for any choice of $\pi \in (0, \min\{\overline{\pi}_H, \overline{\pi}_L\})$, we have

$$V_e^{IC}(U^0) > \hat{V}_e(\pi) \quad \forall e \in E. \tag{23}$$

From $U^0 > U^\ell$, V_e^{FI} strictly decreasing and Φ strictly convex, it also holds that:

$$\hat{V}_e(\pi) > \max\left\{V_e^{FI}\left(\frac{U^0 - U^\ell}{\pi} + U^\ell\right),\, (1-\pi)V_e^{FI}(U^\ell) + \pi V_e^{IC}\left(\frac{U^0 - U^\ell}{\pi} + U^\ell\right)\right\}. \tag{24}$$

[36] If (20) is violated, then there is no pair of nonnegative transfers such that both (IC) and (PC) simultaneously bind in the second-best problem, and corner solutions emerge.

Inequalities (23) and (24) together yield (21) for all $e \in E$. ■

The proof of Lemma 5 shows how to construct a set of punishments against renegotiation when the monetary transfers received by the agent in each state are constrained to be nonnegative.

Indeed, the distribution $(\pi, 1-\pi)$ characterized in the proof is key to define the mechanism $\gamma^b = \{\mathcal{M}^b, \mathcal{S}^b, \sigma^b, \tau^b\}$, with $\mathcal{M}^b = \mathcal{M}^*$ and $\mathcal{S}^b = \mathcal{S}^*$, $\sigma^b(h) = 1-\pi$ and $\sigma^b(t) = \pi$, and transfers

$$\tau^b(N,h) = \tau^b(N,t) = c^{SB};\ \tau^b(R,h) = c^{IC}(U^\ell);\ \tau^b(R,t) = c^{IC}\left(\frac{U^0 - U^\ell}{\pi} + U^\ell\right).$$

This mechanism shares with γ^* the idea that the message $m = R$ activates a (random) counter-offer, which inflicts the relevant punishment. By sending $m = R$ in γ^b the agent receives a "low" transfer with probability $1-\pi$ and a "high" one with probability π. At the same time, the distribution is designed to guarantee the agent an expected utility of U^0:

$$(1-\pi)U^\ell + \pi\left(\frac{U^0 - U^\ell}{\pi} + U^\ell\right) = U^0,$$

which makes incentive-compatible to report $m = N$ on path. The same logic developed in the proof of Proposition 1 then guarantees that γ^b implements the second-best allocation.

Self-Enforced Timing of Communication

We show that Proposition 1 extends to a setting in which the agent chooses when to report and, consequently, when the associated disclosure occurs.

We introduce a new class of self-revealing mechanisms Γ^μ with $\gamma^\mu = (\mathcal{M}^\mu, \mathcal{S}^\mu, \sigma^\mu, \tau^\mu) \in \Gamma^\mu$, in which the agent can send a message at each of the stages (iii), (iv), (v) and (vi). The message space is $\mathcal{M}^\mu = \{N, R, \emptyset\}^4$, extending $\mathcal{M}$ along two directions: four stages of communication are available and, at each stage, the agent can send the *empty* message $\emptyset$, representing her choice not to communicate. Analogously, the signal space is $\mathcal{S}^\mu = \{h, t, \emptyset\}^4$. For any stage $k \in \{iii, iv, v, vi\}$, we denote by $m_k \in \{N, R, \emptyset\}$ the agent's private report and by $s_k \in \{h, t, \emptyset\}$ the private signal sent by the mechanism.

The mechanism γ^μ is constructed so that the agent herself determines the timing of her report and the coin toss. No external enforcement of a specific communication protocol is required. Specifically, γ^μ satisfies:

1. The decision rule $\tau^\mu : \mathcal{M}^\mu \times \mathcal{S}^\mu \to \mathbb{R}^2$ is invariant to the timing of the non-empty message: it depends only on the content (m_k, s_k) of the unique stage k at which $m_k \neq \emptyset$. If the agent sends more than one non-empty message ($|\{k : m_k \neq \emptyset\}| \neq 1$), the mechanism imposes a sufficiently large penalty to make this strictly dominated.

2. The disclosure rule $\sigma = (\sigma_k)_{k=iii}^{vi}$ sends at most one non-empty signal: σ_k privately reveals to the agent the outcome $s_k \in \{h, t\}$ of a fair coin toss at the first stage k at which $m_k \neq \emptyset$, and $s_k = \emptyset$ at all other stages. If $m_k = \emptyset$ for all k, then $s_k = \emptyset$ for all k.

That is, whenever the agent sends a non-empty report, the mechanism responds with a single coin-toss signal at that stage and remains silent otherwise.

Any optimal reporting of the agent in a given $\gamma^\mu \in \Gamma^\mu$ involves exactly one non-empty message by property (1), inducing the disclosure of exactly one non-empty signal by property (2). Let $(m_k, s_k)_{k=iii}^{vi}$ be any message-signal array with this property, and let $(m_j, s_j) \in \{N, R\} \times \{h, t\}$ denote its unique non-empty element, where $j \in \{iii, iv, v, vi\}$ is the stage at which the agent reports. By property (1), τ^μ depends only on the content (m_j, s_j), not on j. A mechanism $\gamma^\mu \in \Gamma^\mu$ is therefore identified by four transfer pairs $(\tau^\mu(m_j, s_j))_{(m_j,s_j)\in\{N,R\}\times\{h,t\}} \in \mathbb{R}^{2\times 4}$.

This construction ensures that a court need *not* verify the timing of communication in γ^μ, only its content. In particular, the court need not determine whether a report is sent before or after a renegotiation offer. Also, it need not verify the timing of the coin toss, which is determined ~~selected~~ by the agent herself, through her choice of when to send a non-empty report.

We now consider the overall game G_{Γ^μ} in which the principal selects a mechanism in Γ^μ at the ex-ante stage.

Lemma 6 *The game G_{Γ^μ} has a unique pure-strategy equilibrium allocation, which coincides with the second-best one (H, c^{SB}).*

Proof. We start by considering the following subgame $G_{\Gamma^\mu}(\gamma^\mu)$, which starts as of stage (iii) after the agent has accepted γ^μ. Since sending more than one non-empty message is strictly dominated, we restrict attention to strategies in which the agent sends exactly one non-empty message across stages (iii)–(vi). Thus, at each stage $k \in \{iii, iv, v, vi\}$, the agent's communication in γ^μ follows a common rule: if she has already privately sent a non-empty message at some earlier stage, she privately sends $m_k = \emptyset$ and receives $s_k = \emptyset$. Otherwise, she privately sends $m_k \in \{N, R, \emptyset\}$; if $m_k \neq \emptyset$, she receives a private signal $s_k \in \{h, t\}$ drawn from a fair coin toss, and if $m_k = \emptyset$, she receives $s_k = \emptyset$. Note that, unlike in G_Γ, the agent may communicate before choosing effort at stage (iii). Given this rule, the game proceeds as follows:

(iii) The agent communicates in γ^μ (per the rule above). She then privately chooses $e \in E$.

(*iv*) The agent communicates in γ^μ. Without observing e or prior communication, the principal makes a public renegotiation offer $\gamma^r \in C \cup \{\emptyset\}$.

(*v*) The agent communicates in γ^μ. If $\gamma^r \neq \emptyset$, the agent then publicly accepts or rejects γ^r by declaring $\rho \in \{y, n\}$.

(*vi*) If $\gamma^r \neq \emptyset$ and $\rho = y$, the agent sends a private message $m^r \in E$ in γ^r. Nature publicly draws the output realization g or b, and transfers are determined by $\gamma^r(m^r)$. If, instead, either $\gamma^r = \emptyset$ or $\rho = n$, then γ^μ executes. If no non-empty message has been sent at stages (*iii*)–(*v*), the agent sends $m_{vi} \in \{N, R\}$ and receives $s_{vi} \in \{h, t\}$; otherwise $m_{vi} = s_{vi} = \emptyset$. Nature publicly draws the output realization g or b, the non-empty communication (m_j, s_j) in γ^μ is publicly revealed and transfers are determined by $\tau^\mu(m_j, s_j)$.

A pure strategy of the principal in the subgame specifies a renegotiation offer $\gamma^r \in C \cup \{\emptyset\}$. Since communication is private, the principal can condition his offer neither on the timing nor on the content of the agent's report.[37]

A behavioral strategy of the agent specifies, at each stage $k \in \{iii, iv, v, vi\}$, a distribution over $m_k \in \{N, R, \emptyset\}$ conditional on the history of play; as established above, we focus on strategies in which exactly one m_k is non-empty. It further specifies an effort probability $x \in [0, 1]$ after any stage-(*iii*) communication, a participation decision $\rho \in \{y, n\}$ at each history following $\gamma^r \neq \emptyset$, and a report $m^r \in E$ in γ^r whenever $\rho = y$.

We now show that (H, c^{SB}) is indeed a pure-strategy equilibrium allocation of G_{Γ^μ}. Consider the mechanism $\gamma^{\mu*} \in \Gamma^\mu$ with $\tau^{\mu*}(m_j, s_j) = \tau^*(m_j, s_j)$ for all $(m_j, s_j) \in \{N, R\} \times \{h, t\}$. We construct a continuation equilibrium of $G_{\Gamma^\mu}(\gamma^{\mu*})$ in which, on path, the agent chooses high effort $e = H$; the principal makes no renegotiation offer, $\gamma^r = \emptyset$; the agent reports $m_{iii} = m_{iv} = \emptyset$ and $m_v = N$. Off path, if the principal offers $\gamma^r \neq \emptyset$, the agent reports at stage (*v*) and takes her participation decision ρ following the rules in Proposition 1. The proof of Proposition 1 guarantees that these strategies constitute an equilibrium. Indeed, since $\gamma^r = \emptyset$ is offered on-path, the option to report at stage (*iii*) or (*iv*) is strategically irrelevant.

Under this continuation equilibrium, the principal obtains V^{SB} by offering $\gamma^{\mu*}$ at stage (*i*). Since he cannot obtain more with any other mechanism, V^{SB} is an equilibrium utility of the principal in G_{Γ^μ}, supported by the initial offer of $\gamma^{\mu*}$ and the continuation just described.

We now argue that V^{SB} is the unique continuation utility of the principal in $G_{\Gamma^\mu}(\gamma^{\mu*})$ compatible with a pure strategy equilibrium, following the logic of Proposition 2. We

[37] As in Section 3.1, revelation mechanisms without signals entail no loss of generality at the renegotiation stage.

construct a perturbed version γ_ε^μ of $\gamma^{\mu*}$ with $\tau_\varepsilon^\mu(m_j, s_j) = \tau_\varepsilon(m_j, s_j)$ for all $(m_j, s_j) \in \{N, R\} \times \{h, t\}$, where τ_ε is the tie-breaking mechanism from Proposition 2. In particular, let $\tau_\varepsilon^\mu(m_j, h) = c^{IC}(U^0 - \Delta U - \kappa\varepsilon)$ with $\kappa > 2$. We require κ to be large enough that, for both $e \in E$:[38]

$$\frac{1}{2}V_e^{FI}(U^0 - \Delta U - \kappa\varepsilon) + \frac{1}{2}V_e^{IC}(U^0 + \Delta U) \geq V_e^{FI}(U^0 + \Delta U). \tag{25}$$

We show that every pure-strategy equilibrium of $G_{\Gamma^\mu}(\gamma_\varepsilon^\mu)$ yields exactly $V_H(c_\varepsilon^{SB})$ to the principal. The argument rests on two claims. The first establishes that the additional timing flexibility in Γ^μ is strategically irrelevant: in any equilibrium, the agent's reporting behavior can be taken to coincide with the timing in $G_\Gamma(\gamma_\varepsilon)$. The second applies the logic of the proof of Proposition 2 to pin down the equilibrium allocation.

Claim 1 (Timing Reduction). Any pure-strategy equilibrium allocation of $G_{\Gamma^\mu}(\gamma_\varepsilon^\mu)$ can be supported by a strategy profile in which the agent reports only at stage (v).

Proof. The proof is divided into two parts.

Part (a): early reporting is ruled out. Since communication is private, the principal cannot distinguish reporting at stage (iii) from that at stage (iv). It suffices to show that on-path reporting of $m_j = N$ or $m_j = R$ before stage (v) is incompatible with equilibrium requirements.

Suppose first that $m_j = N$ on path at stage (iii) or (iv). For Bayes-consistency, the principal's equilibrium belief must assign probability one to the agent having reported N before stage (v). Under this belief, his optimal renegotiation offer is $c_H^{FI}(U^0)$ or $c_L^{FI}(U^0-\varepsilon)$, depending on the agent's equilibrium effort $e \in E$. These transfers yield the agent her reservation utility U^0 (if $e = H$) or $U^0 - \varepsilon$ (if $e = L$) from $\tau_\varepsilon^\mu(N, s) = c_\varepsilon^{SB}$, while yielding the principal $V_H^{FI}(U^0)$ or $V_L^{FI}(U^0 - \varepsilon)$. By sequential rationality, these are the only two offers that may arise in any such equilibrium. However, in both cases, the agent can profitably deviate by delaying her report and sending $m_j = R$ at a later stage with the same effort e. This yields her $U^0 + \frac{\Delta U}{2} > U^0$ if $e = H$, or $U^0 + \frac{\Delta U - \varepsilon}{2} > U^0 - \varepsilon$ if $e = L$, a contradiction.

Suppose instead that $m_j = R$ on path at stage (iii) or (iv). The principal must hold the degenerate belief that the agent reported R before stage (v). As established in the proof of Proposition 2, the principal's optimal offer must be a full-insurance contract (given his belief that effort is degenerate). The utility left to the agent must be either $U^0 + \Delta U$, the lowest level she accepts for both $s_j \in \{h, t\}$, or $U^0 - \Delta U - \kappa\varepsilon$, the lowest she accepts when $s_j = h$. Comparing the principal's utility under these two options gives the terms of (25). Hence, (25) guarantees that offering $c_e^{FI}(U^0 - \Delta U - \kappa\varepsilon)$ is optimal

[38] Such a κ exists since $\lim_{\kappa\to\infty} V_e^{FI}(U^0 - \Delta U - \kappa\varepsilon) = \infty$, while all other terms in (25) are finite for every $\kappa > 2$.

for each $e \in E$. Any such equilibrium must therefore feature this offer. Since, for any equilibrium effort $e \in E$, this offer yields the agent no more than $U_e(\tau_\varepsilon^\mu(R,h))$, her utility is $U^0 - \frac{\kappa}{2}\varepsilon$, obtained under any optimal participation behavior and for each effort. The agent can profitably deviate: she sends $m_j = N$ at any stage, selects the same $e \in E$, and rejects γ^r, obtaining U^0, a contradiction.[39]

Part (b): late reporting is allocation-equivalent to stage-(v) reporting. We argue that every pure-strategy equilibrium allocation is also supported in an equilibrium where the agent sends $m_v \neq \emptyset$ after every offer γ^r, on or off path. In particular, any equilibrium where $m_v = \emptyset$ at some history with $m_{iii} = m_{iv} = \emptyset$ has a corresponding equilibrium where $m_v = N$ at every such history, supporting the same allocation.

In γ_ε^μ, the agent obtains U^0 (if $e = H$) or $U^0 - \varepsilon$ (if $e = L$) from $m_j = N$, and $U^0 - \frac{\kappa}{2}\varepsilon$ from $m_j = R$ regardless of effort. Since $\varepsilon > 0$ and $\kappa > 2$, the unique optimal report at any history where $\rho = n$ and $m_{iii} = m_{iv} = m_v = \emptyset$ is $m_{vi} = N$.

Furthermore, by construction of γ_ε^μ, $\tau_\varepsilon^\mu(N,s) = c_\varepsilon^{SB}$ for all $s_j \in \{h,t\}$. Hence the realization of s_j is payoff-irrelevant when $m_j = N$. Starting from an equilibrium where $m_v = \emptyset$ at some off-path history, the agent can thus adopt the following equivalent behavior: send $m_v = N$ rather than $m_v = \emptyset$ at every such history and select, for each payoff-irrelevant realization of $s_v \in \{h,t\}$, the same participation decision as in the original equilibrium. Since the rejection utility $U_e(c_\varepsilon^{SB})$ and the acceptance utility $\hat{U}_e^r$ are both independent of the signal, this participation decision remains optimal. The newly constructed strategy therefore supports the same allocation in equilibrium. This establishes Claim 1. ■

Claim 2 (Application of Proposition 2). The unique pure-strategy equilibrium allocation of $G_{\Gamma^\mu}(\gamma_\varepsilon^\mu)$ is (H, c_ε^{SB}).

Proof. By Claim 1, in any pure-strategy equilibrium, the agent reports at stage (v): after observing the principal's offer γ^r but before her participation decision. The principal believes $m_{iii} = m_{iv} = \emptyset$ with probability one. Since $\tau_\varepsilon^\mu = \tau_\varepsilon$, the agent's reporting and participation behavior in $G_{\Gamma^\mu}(\gamma_\varepsilon^\mu)$ coincides with that characterized in the proof of Proposition 2, where the agent's reports exhibit this timing by construction of γ_ε.

Two consequences follow. First, as shown in Proposition 2, any renegotiation offer accepted with positive probability yields the principal a utility strictly below $V_e(c_\varepsilon^{SB})$ for all $e \in E$. Hence only $\gamma^r = \emptyset$ or offers rejected with probability one by the agent are

[39]Part *(a)* covers the case in which the agent, after reporting early, selects the same effort at each history, thereby *not* exploiting the signal to introduce stochasticity in her effort choice. If the agent reports $m_{iii} = R$ and conditions her effort on s_{iii}, selecting $\hat{e}(s_{iii})$ contingent on the signal realization, $s_{iii} \in \{h,t\}$, then for ΔU large enough a condition analogous to (25) guarantees that $c^{FI}_{\hat{e}(h)}(U^0 - \Delta U - \kappa\varepsilon)$ remains the principal's optimal offer for any $(\hat{e}(h), \hat{e}(t)) \in E^2$. The remainder of the argument then follows directly. The extension to $m_{iii} = N$ is immediate, since this report renders s_j payoff-irrelevant in γ_ε^μ.

compatible with the principal's sequential rationality. Second, given $\gamma^r = \emptyset$, the agent anticipates her unique optimal report to be $m_j = N$, which leads to the execution of the strictly incentive-compatible transfers c_ε^{SB}. She is therefore strictly better off choosing $e = H$. This establishes Claim 2. ■

Taken together, Claims 1 and 2 imply that every pure-strategy equilibrium of $G_{\Gamma^\mu}(\gamma_\varepsilon^\mu)$ induces the allocation (H, c_ε^{SB}), yielding the principal $V_H(c_\varepsilon^{SB})$. Equilibrium existence in the subgame is ensured by the fact that $e = H$, $m_{iii} = m_{iv} = m_{vi} = \emptyset$, $m_v = N$, and $\gamma^r = \emptyset$ are optimal behaviors on the equilibrium path. Since $\lim_{\varepsilon \to 0} V_H(c_\varepsilon^{SB}) = V^{SB}$, as in Proposition 2, V^{SB} is the unique equilibrium utility of the principal in G_{Γ^μ}, and (H, c^{SB}) is the unique equilibrium allocation. ■

Lemma 6 shows that delegating the choice of communication protocol to the agent does not generate any new incentive for renegotiation: the principal cannot induce the agent to accept a renegotiation offer without first reporting in γ^μ, nor can he benefit from the agent reporting before the renegotiation stage (i.e., induce $m_{iii} \neq \emptyset$ or $m_{iv} \neq \emptyset$). While the proof restricts attention to pure strategies for simplicity, the argument extends to behavioral strategies. As in Proposition 2, any renegotiation that is unprofitable against any pure effort choice is *a fortiori* unprofitable against a mixture, since the principal's utility after renegotiation is linear in the agent's effort distribution. Hence $\gamma^r = \emptyset$ remains optimal for the principal under behavioral strategies.

Renegotiation with Public Signals

We here show that privacy of the signals is not needed to achieve our efficiency result. Specifically, we show that the mechanism γ^{**} as defined in Section 4.3 supports the second-best allocation (H, c^{SB}) at equilibrium. To argue this, first consider the subgame $G_{Pub}(\gamma^{**})$, which starts after γ^{**} is offered and accepted:

(*iii*) The agent privately chooses $e \in E$.

(*iv*) Without observing e, the principal makes a public renegotiation offer $\gamma^r = (\mathcal{M}^r, \tau^r)$ or does not renegotiate ($\gamma^r = \emptyset$), where $\mathcal{M}^r = E$ and $\tau^r : \mathcal{M}^r \times \mathcal{S} \to \mathbb{R}^2$, allowing the renegotiating principal to condition on the realization of $s \in \mathcal{S}$.

(*v*) The agent sends a private message $m \in \mathcal{M}^{**} = \{N, R_1, R_2\}$. The signal $s \in \mathcal{S}^* = \{h, t\}$ distributed as $\sigma^{**} = \left(\frac{1}{2}, \frac{1}{2}\right)$ is realized and publicly revealed. If $\gamma^r \neq \emptyset$, the agent publicly accepts or rejects γ^r by declaring $\rho \in \{y, n\}$.

(*vi*) If $\gamma^r \neq \emptyset$ and $\rho = y$, the agent sends a private message $m^r \in E$ in γ^r. Nature publicly draws the output realization g or b, and transfers are determined by $\tau^r(m^r, s)$. If,

instead, either $\gamma^r = \emptyset$ or $\rho = n$, then γ^{**} executes. Nature publicly draws the output realization g or b, the message m from stage (v) is publicly revealed, and transfers are determined by $\tau^{**}(m, s)$.

A pure strategy of the principal in $G_{Pub}(\gamma^{**})$ is a signal-contingent renegotiated offer γ^r.[40] An agent's behavioral strategy λ consists of a randomization $(x, 1-x)$ over $e \in E$ at her initial history, a randomization over messages in $\mathcal{M}^{**}$ at each history (e, γ^r), a randomization over participation decisions $\rho \in \{y, n\}$ at each history (e, γ^r, m, s) where $\gamma^r \neq \emptyset$ and a randomization over messages in $\mathcal{M}^r$ at the continuation history where $\rho = y$. The following holds:

Lemma 7 *The allocation (H, c^{SB}) is supported in an equilibrium of $G_{Pub}(\gamma^{**})$.*

Proof. For any signal $s \in \mathcal{S}^{**}$ extracted in γ^{**}, let $\hat{m}_e^r(s) \in \arg\max_{m^r \in \mathcal{M}^r} U_e(\tau^r(\hat{m}_e^r(s), s))$ be an optimal message that the agent may send after accepting γ^r, having chosen the effort $e \in E$ and observed the public realization of $s \in \mathcal{S}^*$. Following (3), we denote $\hat{U}_e^r(s)$ the agent's corresponding optimal utility $\hat{U}_e^r(s) \equiv U_e(\tau^r(\hat{m}_e^r(s), s))$.

We now construct a PBE of $G_{Pub}(\gamma^{**})$ which implements the allocation (H, c^{SB}).

The principal's equilibrium behavior prescribes not to renegotiate, i.e., $\gamma^r = \emptyset$. We now construct the agent's equilibrium behavior starting from the terminal histories. At each history $(e, \gamma^r \neq \emptyset, m, s, y)$, she sends an optimal message $\hat{m}_e^r(s)$ to γ^r, which she has accepted. At each history $(e, \gamma^r \neq \emptyset, m, s)$ the agent's participation decisions are the following:

(i) If $m = N$, for all $s \in \{h, t\}$, the agent selects $\rho = y$ iff $\hat{U}_e^r(s) \geq U^0$;

(ii) If $(m = R_1, s = h)$ or $(m = R_2, s = t)$, the agent selects $\rho = y$ iff $\hat{U}_e^r(s) \geq U^0 - \Delta U$;

(iii) If $(m = R_1, s = t)$ or $(m = R_2, s = h)$, the agent selects $\rho = y$ iff $\hat{U}_e^r(s) \geq U^0 + \Delta U$.

At each history $(e, \gamma^r = \emptyset)$, the agent sends $m = N$ to γ^{**}. At each history $(e, \gamma^r \neq \emptyset)$ the agent's messages in γ^{**} look as follows:

(i) For any $e \in E$ and for any γ^r such that

$$\begin{aligned} &\frac{1}{2}\max\{U^0, \hat{U}_e^r(h)\} + \frac{1}{2}\max\{U^0, \hat{U}_e^r(t)\} \geq \\ &\quad \max\Big\{\frac{1}{2}\max\{U^0 - \Delta U, \hat{U}_e^r(h)\} + \frac{1}{2}\max\{U^0 + \Delta U, \hat{U}_e^r(t)\}, \\ &\qquad \frac{1}{2}\max\{U^0 + \Delta U, \hat{U}_e^r(h)\} + \frac{1}{2}\max\{U^0 - \Delta U, \hat{U}_e^r(t)\}\Big\}, \end{aligned} \tag{26}$$

[40] Similar arguments to Section 3.1 guarantee that revelation mechanisms not featuring disclosures of signals are without loss of generality at the renegotiation stage.

the agent sends $m = N$ in γ^{**}. Observe that the LHS of (26) corresponds to the agent's expected utility of reporting $m = N$ in γ^{**}, followed by her signal-contingent participation decisions. The RHS of (26) characterizes the utility corresponding to the best alternative report.

(*ii*) For any $e \in E$, and for any $\gamma^r \neq \emptyset$ such that (26) is *not* satisfied, the agent sends $m = R_1$ in γ^{**} whenever

$$\begin{aligned} \frac{1}{2}\max\{U^0 - \Delta U, \hat{U}_e^r(h)\} + \frac{1}{2}\max\{U^0 + \Delta U, \hat{U}_e^r(t)\} \geq \\ \frac{1}{2}\max\{U^0 + \Delta U, \hat{U}_e^r(h)\} + \frac{1}{2}\max\{U^0 - \Delta U, \hat{U}_e^r(t)\}. \end{aligned} \tag{27}$$

(*iii*) For any $e \in E$, and for any $\gamma^r \neq \emptyset$ such that (26) and (27) are *not* satisfied, the agent sends $m = R_2$ in γ^{**}

To complete the description of the agent's behavior, at her initial history she takes the effort decision $e = H$ with probability $x = 1$. Finally, the principal belief attributes probability one to $e = H$ at his only information set, consistently with the agent's behavior.

We next verify the sequential rationality of our construction. It is immediate to check that the agent's strategy is sequentially rational. In particular, the threshold participation behavior simply compares the agent's continuation utility of accepting γ^r versus retaining γ^{**}; the reporting behavior is also described by comparing the agent's continuation utility after sending each report, without further elaboration. The effort choice $e = H$ is optimal since, on the equilibrium path, the incentive-compatible transfers $c^{SB} = c^{IC}(U^0)$ are executed.

To conclude the proof, it remains to check that there is no renegotiated offer $\gamma^r \neq \emptyset$ yielding the principal a strictly higher utility than V^{SB}, which he obtains in equilibrium. To verify it, we partition the set of available renegotiated offers according to the reports that $\lambda(\gamma^{**})$ induce in the mechanism γ^{**}.

Observe first that, for any γ^r such that the agent reports $m \in \{R_1, R_2\}$ in γ^{**}, the principal's utility cannot exceed

$$V^R \equiv \frac{1}{2}V_H^{FI}(U^0 - \Delta U) + \frac{1}{2}V_H^{FI}(U^0 + \Delta U),$$

that is, the utility of providing full insurance to the agent conditional on each realized signal. In this case, (4) in the proof of Lemma 1 guarantees that $V^{SB} > V^R$. Thus, the principal prefers not to renegotiate to ~~than~~ renegotiating an offer which induces the report $m \in \{R_1, R_2\}$.

Thus, any profitable renegotiation γ^r must be such that the agent's equilibrium strategy prescribes to report $m = N$ in γ^{**}. That is, given (26), and since $e = H$, one should

have:

$$\frac{1}{2}\max\{U^0,\hat{U}_H^r(h)\}+\frac{1}{2}\max\{U^0,\hat{U}_H^r(t)\}\geq$$
$$\max\Big\{\frac{1}{2}\max\{U^0-\Delta U,\hat{U}_H^r(h)\}+\frac{1}{2}\max\{U^0+\Delta U,\hat{U}_H^r(t)\},\qquad(28)$$
$$\frac{1}{2}\max\{U^0+\Delta U,\hat{U}_H^r(h)\}+\frac{1}{2}\max\{U^0-\Delta U,\hat{U}_H^r(t)\}\Big\}.$$

We now argue that (28) is satisfied only if one of the following two conditions is met:

$$\hat{U}_H^r(s)\leq U^0-\Delta U,\ \forall s\in\mathcal{S}^*\text{ or }\hat{U}_H^r(s)\geq U^0+\Delta U,\ \forall s\in\mathcal{S}^*.\qquad(29)$$

To see this, suppose that (29) does not hold, which leads to consider three cases.

(*i*) If $\hat{U}_H^r(t)\leq U^0-\Delta U$ and $\hat{U}_H^r(h)>U^0-\Delta U$, then the LHS of (28) is $\frac{1}{2}\max\{U^0,\hat{U}_H^r(h)\}+\frac{1}{2}U^0$ and its RHS is at least $\frac{1}{2}\hat{U}_H^r(h)+\frac{1}{2}(U^0+\Delta U)$, which obtains for $m=R_1$. The latter is strictly greater than the former, which violates (28).

(*ii*) If $U^0-\Delta U\leq\hat{U}_H^r(t)<U^0+\Delta U$, then the LHS of (28) is $\frac{1}{2}\max\{U^0,\hat{U}_H^r(h)\}+\frac{1}{2}\hat{U}_H^r(t)$. Suppose now that $\hat{U}_H^r(h)<U^0+\Delta U$: the value of the RHS is at least $\frac{1}{2}(U^0+\Delta U)+\frac{1}{2}\hat{U}_H^r(t)$, which obtains for $m=R_2$. The latter is strictly greater than the former, which violates (28). In the mutually exclusive case $\hat{U}_H^r(h)\geq U^0+\Delta U$, the value of the RHS is at least $\frac{1}{2}\hat{U}_H^r(h)+\frac{1}{2}(U^0+\Delta U)$, which obtains for $m=R_1$, which leads to violate (28) again.

(*iii*) If $\hat{U}_H^r(t)\geq U^0+\Delta U$, and $\hat{U}_H^r(h)<U^0+\Delta U$, the LHS of (28) is $\frac{1}{2}\max\{U^0,\hat{U}_H^r(h)\}+\frac{1}{2}\hat{U}_H^r(t)$, and the RHS is at least $\frac{1}{2}(U^0+\Delta U)+\frac{1}{2}\hat{U}_H^r(t)$, which obtains for $m=R_2$. The latter is strictly greater than the former, which violates (28)

Thus, following a renegotiation γ^r, $\lambda(\gamma^{**})$ prescribes $m=N$ and only if (29) holds. Two cases must then be considered:

(i) If $\hat{U}_H^r(s)\leq U^0-\Delta U\ \forall s\in\mathcal{S}^*$, then (28) rewrites $U^0\geq U^0$, and is thus satisfied with equality. Thus, $\lambda(\gamma^{**})$ prescribes to report $m=N$ in γ^{**} and to choose $\rho=n$, which yields the principal the same profit V^{SB} obtained without renegotiation.

(ii) If $\hat{U}_H^r(s)\geq U^0+\Delta U\ \forall s\in\mathcal{S}^*$, then (28) rewrites $U^0+\Delta U\geq U^0+\Delta U$, and is thus satisfied with equality. Thus, $\lambda(\gamma^{**})$ prescribes to report $m=N$ in γ^{**}. In addition, for any such γ^r, the agent is guaranteed the utility $U^0+\Delta U$ in the continuation play, which implies that the principal's utility cannot exceed $V_H^{FI}(U^0+\Delta U)$, which is strictly less than V^{SB} as shown in Lemma 1.

Thus, the principal's strategy $\gamma^r=\emptyset$ is sequentially rational. ■

Supplementary Renegotiation

We extend our analysis to the case in which a renegotiated offer may *supplement*, and not necessarily replace, the original one. As discussed in Section 5.2, the optimal mechanism γ^*, which implements the second-best allocation, is vulnerable to a supplementary offer which conditions on its final transfers. Such an offer provides a supplementary transfer to the agent only if the transfers executed by γ^* reveal that she has reported $m = N$. This allows the principal to simultaneously improve risk-sharing and circumvent the punishments incorporated in γ^*.

However, the vulnerability of γ^* to supplementary renegotiation does not reflect a general weakness of self-revealing mechanisms. We formally establish this point by constructing a self-revealing mechanism $\gamma^o \in \Gamma$ that successfully implements the second-best allocation under a large class of CRRA preferences for the agent.

Specifically, we take $\gamma^o = \{\mathcal{M}^o, \mathcal{S}^o, \sigma^o, \tau^o\}$, with $\mathcal{M}^o = \{N, R\}$, $\mathcal{S}^o = \{h, t\}$ and $\sigma^o = (1-\pi, \pi)$, where $\pi < \bar{\pi} \equiv \Phi'(u_b^{SB})/\Phi'(u_g^{SB})$ is such that:[41].

$$(1-\pi)V_H^{FI}((1-p_H)u_b^{SB} - d) + \pi V_H^{FI}\left(p_H \frac{u_g^{SB}}{\pi} + (1-p_H)u_b^{SB} - d\right) < V^{SB}. \tag{30}$$

The upper bound $\bar{\pi}$ on π ensures that, at c^{SB}, the marginal cost to the principal of increasing the agent's bad-state transfer exceeds the marginal benefit of reducing her good-state transfer, ~~which is key to render~~ rendering supplementary renegotiation ~~marginally~~ unprofitable ~~at c^{SB}~~.

The decision rule τ^o is:

$$\tau^o(N, s) = c^{SB} \;\; \forall s \in \mathcal{S}^o; \quad \tau^o(R, h) = (0, u_b^{SB}); \quad \tau^o(R, t) = \left(\frac{u_g^{SB}}{\pi}, u_b^{SB}\right).$$

The mechanism γ^o yields the same expected payoffs as γ^* to the agent for each $m \in \{N, R\}$. However, it pays a flat transfer in state $\omega = b$ and a transfer conditional on (m, s) in state $\omega = g$. This feature is key to deter any supplementary renegotiation. Because γ^o's transfer in state $\omega = b$ is uninformative of the agent's report, a renegotiation offer cannot condition the provision of a supplementary transfer on the agent reporting $m = N$. This allows her to combine any additional transfer in $\omega = b$ with the lottery induced by $m = R$, and undermines the power of supplementary renegotiation.

Using the subscript "+" to indicate supplementarity, ~~To develop our argument,~~ we begin by describing the set $\Gamma_+^r(\gamma^o)$ of *supplementary* renegotiation offers available to the principal once γ^o has been accepted. A supplementary offer $\gamma_+^r = \{\mathcal{M}_+^r, \tau_+^r\} \in \Gamma_+^r(\gamma^o)$

[41] Existence of such a π is guaranteed by arguments analogous to those in the proof of Lemma 5

involves a finite message space $\mathcal{M}_+^r$ and a decision rule $\tau_+^r : \mathcal{M}_+^r \times \left\{u_b^{SB}, 0, u_g^{SB}, \frac{u_g^{SB}}{\pi}\right\} \to \mathbb{R}^2$. The rule τ_+^r can condition on both the agent's report $m_+^r \in \mathcal{M}_+^r$ in γ_+^r and the final transfer $u_\omega^o \in \left\{u_b^{SB}, 0, u_g^{SB}, \frac{u_g^{SB}}{\pi}\right\}$ executed by γ^o.

The (sub)game $G_\Gamma^+(\gamma^o)$ of supplementary renegotiation induced by γ^o modifies the stages (iii)-(vi) of G_Γ as follows:

(iii) The agent privately chooses $e \in E$.

(iv) Without observing e, the principal makes a public renegotiation offer $\gamma_+^r \in \Gamma_+^r(\gamma^o) \cup \{\emptyset\}$, where $\emptyset$ represents the principal's decision not to renegotiate.

(v) The agent sends a private message $m \in \{N, R\}$ in γ^o and receives a private random signal $s \in \{h, t\}$. If $\gamma_+^r \neq \emptyset$, the agent publicly accepts or rejects γ_+^r by declaring $\rho \in \{y, n\}$.

(vi) If $\gamma_+^r \neq \emptyset$ and $\rho = y$, the agent sends a private message $m_+^r \in \mathcal{M}_+^r$ in γ_+^r, after which nature draws the output realization g or b. In this case, *both* γ^o and γ_+^r execute. If, instead, either $\gamma_+^r = \emptyset$ or $\rho = n$, then *only* γ^o executes after nature draws g or b.

The game $G_\Gamma^+(\gamma^o)$ embeds the following assumptions:

—At stage (iv), we restrict the principal to make renegotiation offers in the set $\Gamma_+^r(\gamma^o)$. These mechanisms allow the agent to report her private information (e, m, s) on her choice of effort and the communication she entertained within γ^o. At the same time, they do not make use of signals, since, upon accepting γ_+^r, the agent no longer reports ~~is *not* any longer reporting~~ in γ^o.

—At stage (v), the agent may either accept γ_+^r in addition to γ^o, or remain bound only by ~~γ^o *and* γ_+^r, or participate only in~~ γ^o. That is, in contrast with Assumption $A.3$, accepting the new offer does *not* cancel the original one. There is, in particular, no need to explicitly consider offers which lead the agent to "replace" the original mechanism γ^o.[42]

—At stage (vi), final output realizes and transfers are implemented. The simultaneous execution of γ^o and γ_+^r ~~The possibility of having both mechanisms γ^o and γ_+^r to execute simultaneously~~ requires a careful specification of the disclosure rule incorporated in the original mechanism γ^o. Since any renegotiation offer supplements the original one, the enforceability of the relevant punishment requires that the communication entertained within γ^o also be publicly revealed off-path.~~is only possible if the communication~~

[42]For any offer $\gamma_+^r = \{\mathcal{M}_+^r, \tau_+^r\} \in \Gamma_+^r(\gamma^o)$ inducing the agent to replace γ^o, thereby implementing the final transfers $(u_g^r(m_+^r), u_b^r(m_+^r))$, with $m_+^r \in \mathcal{M}_+^r$, there is another offer $\gamma_+' = \{\mathcal{M}_+^r, \tau_+'\} \in \Gamma_+^r(\gamma^o)$ ~~$\gamma_+' = \{\mathcal{M}_+', \tau_+'\} \in \Gamma_+^r(\gamma^o)$~~ inducing the agent to accept both γ^o and γ_+' and implementing the same final transfers. This obtains by letting $\tau_+'(m_+^r, u_\omega^o) = u_\omega^r(m_+^r) - u_\omega^o$ ~~$(\tilde{u}_g'(m_+^r, u_g^o), \tilde{u}_b'(m_+^r, u_b^o)) = (u_g^r(m_+^r) - u_g^o, u_b^r(m_+^r) - u_b^o)$~~ for every $m_+^r \in \mathcal{M}_+^r$, every $\omega \in \{g, b\}$, and every $u_\omega^o \in \{u_b^{SB}, 0, u_g^{SB}, u_g^{SB}/\pi\}$ ~~and (u_g^o, u_b^o) implementable by γ^o~~.

~~entertained within γ^o is also publicly revealed off the path~~. In our construction, to guarantee enforceability of γ^o it is sufficient to condition the disclosure of its communications on the realization of the state $\omega = g$, since γ^o's transfers in state $\omega = b$ are not contingent on communication. At the same time, this disclosure does not generate new renegotiation opportunities since γ^r_+ can already directly condition on γ^o's transfers in the state $\omega = g$.

The following result establishes that γ^o is renegotiation-proof to supplementary offers for a large class of preferences for the agent.

Proposition 4 *If $\Phi(u) = (\alpha u)^{\frac{1}{\alpha}}$, with $\alpha \in (0,1)$, then the second-best allocation (H, c^{SB}) is supported in an equilibrium of $G^+_\Gamma(\gamma^o)$.*

Proof. For any $\gamma^r_+ \in \Gamma^r_+(\gamma^o)$ and for any $m^r_+ \in \mathcal{M}^r_+$, we write $\tau^r_+(m^r_+, u^o_\omega) = (u^r_g(m^r_+, u^o_g), u^r_b(m^r_+, u^o_b))$. We construct an equilibrium where, on path, the agent chooses $e = H$ and $m = N$, and the principal offers $\gamma^r_+ = \emptyset$.

Off-path, we only require that, at each history $(H, \gamma^r_+, N, s, y)_{s\in\{h,t\}}$, the agent reports in γ^r_+ the same message, denoted:

$$m^N_{\gamma^r_+} \in \arg\max_{m^r_+ \in \mathcal{M}^r_+} p_H(u^{SB}_g + u^r_g(m^r_+, u^{SB}_g)) + (1-p_H)(u^{SB}_b + u^r_b(m^r_+, u^{SB}_b)).$$

We also let the principal believe that $e = H$ with probability one at his only information set, consistently with the agent's equilibrium play.

On path, the agent's sequential rationality can be checked observing that γ^o yields her U^0 for any $(e, m) \in \{H, L\} \times \{N, R\}$ when $\gamma^r_+ = \emptyset$. Choosing $e = H$ and $m = N$ is thus optimal. Off-path, $m^r_+ = m^N_{\gamma^r_+}$ is optimal by construction at each (H, γ^r_+, N, s, y).

We verify instead the principal's sequential rationality in two steps: first, we establish a lower bound on the agent's continuation utility at each history $(H, \gamma^r_+)_{\gamma^r_+ \in \Gamma^r_+(\gamma^o)}$; second, we show that all offers inducing either $m = N$ or $m = R$ are unprofitable for the principal given his belief that $e = H$ with probability one.

Step 1. At each history (H, γ^r_+), the agent's continuation utility is at least U^0, which she obtains by choosing $m = N$ and $\rho = n$ for all s. We now identify another lower bound.

Consider the following deviation: the agent reports $m = R$, accepts γ^r_+ when $s = h$ reporting $m^r_+ = m^N_{\gamma^r_+}$ in it, and rejects it when $s = t$. This yields her:

$$U^N_{\gamma^r_+} \equiv (1-\pi)[p_H u^r_g(m^N_{\gamma^r_+}, 0) + (1-p_H)(u^{SB}_b + u^r_b(m^N_{\gamma^r_+}, u^{SB}_b))] + \pi\left[p_H \frac{u^{SB}_g}{\pi} + (1-p_H)u^{SB}_b\right] - d.$$

Since $\alpha \in (0,1)$, only positive utility transfers are defined. To guarantee that final transfers satisfy this requirement for every $\gamma^r_+ \in \Gamma^r_+(\gamma^o)$, it must be that

$$u^o_\omega + u^r_\omega \geq 0 \ \ \forall \omega \in \{g, b\} \text{ and } \forall u^o_\omega \in \left\{u^{SB}_b, 0, u^{SB}_g, \frac{u^{SB}_g}{\pi}\right\}.$$

This implies in particular that $u_g^r(m_{\gamma_+^r}^N, 0) \geq 0$, and thus, that $U_{\gamma_+^r}^N$ is bounded below by $U^0 + (1-\pi)(1-p_H)u_b^r(m_{\gamma_+^r}^N, u_b^{SB})$.[43] Thus, at equilibrium, it must be that $U_{\gamma_+^r}^N \geq \max\{U^0, U^0 + (1-\pi)(1-p_H)u_b^r(m_{\gamma_+^r}^N, u_b^{SB})\}$.

Step 2a. Fix any renegotiated offer $\gamma_+^r \in \Gamma_+^r(\gamma^o)$. At each history $(H, \gamma_+^r, N, s, y)_{s\in\{h,t\}}$, the principal's continuation payoff is:

$$V_{\gamma_+^r}^N \equiv y_H - p_H\Phi(u_g^{SB} + u_g^r(m_{\gamma_+^r}^N, u_g^{SB})) - (1-p_H)\Phi(u_b^{SB} + u_b^r(m_{\gamma_+^r}^N, u_b^{SB})). \tag{31}$$

We exploit the result in Step 1 to show that $V_{\gamma_+^r}^N \leq V^{SB}$. We consider two cases.

Suppose first that $u_b^r(m_{\gamma_+^r}^N, u_b^{SB}) < 0$. Then, the state-contingent final transfers to the agent when she accepts γ_+^r provide her with less insurance than the second-best transfers c^{SB}, while yielding an expected utility of at least U^0. Hence, by convexity of Φ, they must yield less than V^{SB} to the principal.

Suppose then that $u_b^r(m_{\gamma_+^r}^N, u_b^{SB}) \geq 0$. In this case, the agent gets

$$U_{\gamma_+^r}^N \geq U^0 + (1-\pi)(1-p_H)u_b^r(m_{\gamma_+^r}^N, u_b^{SB}) \geq U^0. \tag{32}$$

We now characterize the supplementary transfers $(u_g^r, u_b^r) \geq 0$, which maximize $V_{\gamma_+^r}^N$ given $U_{\gamma_+^r}^N$. It is easy to check that (32) binds at any principal's optimal choice. At the solution, in particular, we have $u_g^r = -\pi\frac{1-p_H}{p_H}u_b^r$.[44] Plugging this in (31) yields the following upper bound for $V_{\gamma_+^r}^N$:

$$\max_{u_b^r \geq 0} \quad \nu(u_b^r) \equiv y_H - p_H\Phi\left(u_g^{SB} - \pi\frac{1-p_H}{p_H}u_b^r\right) - (1-p_H)\Phi(u_b^{SB} + u_b^r).$$

Observe that ν is twice differentiable and strictly concave in u_b^r. Also, $\nu'(0) = \pi(1-p_H)\Phi'(u_g^{SB}) - (1-p_H)\Phi'(u_b^{SB}) < 0$ since $0 < \pi < \frac{\Phi'(u_b^{SB})}{\Phi'(u_g^{SB})}$ by construction of γ^o. As a consequence, ν is maximized at $u_b^r = 0$ where $\nu(0) = V^{SB}$. Thus $V_{\gamma_+^r}^N \leq V^{SB}$ holds for any accepted offer γ_+^r inducing $m = N$. Also, since any rejected offer yields V^{SB} to the principal given $(e = H, m = N)$, the principal's continuation payoff at (H, γ_+^r, N) cannot exceed V^{SB} for any $\rho \in \{y, n\}$.

Step 2b. At each history (H, γ_+^r, R), the agent receives a signal $s \in \{h, t\}$, with probabilities $(1-\pi, \pi)$. When $s = h$, she can get the expected utility $(1-p_H)u_b^{SB} - d$ by choosing $\rho = n$, which yields the principal at most $V_H^{FI}((1-p_H)u_b^{SB} - d)$. Similarly, when $s = t$, the principal obtains at most $V_H^{FI}\left(p_H\frac{u_g^{SB}}{\pi} + (1-p_H)u_b^{SB} - d\right)$. Thus, at any (H, γ_+^r, R), the principal obtains at most $(1-\pi)V_H^{FI}((1-p_H)u_b^{SB} - d) + \pi V_H^{FI}\left(p_H\frac{u_g^{SB}}{\pi} + (1-p_H)u_b^{SB} - d\right) < V^{SB}$ where the inequality is guaranteed by (30).

[43] This follows from the fact that $(1-\pi)[p_H \cdot 0 + (1-p_H)u_b^{SB}] + \pi\left[p_H u_g^{SB}/\pi + (1-p_H)u_b^{SB}\right] - d = U^0$.

[44] This obtains from manipulating the expression $U^0 + p_H u_g^r + (1-p_H)u_b^r = U^0 + (1-\pi)(1-p_H)u_b^r$.

Hence, any $\gamma_+^r \neq \emptyset$ yields at most V^{SB} to the principal regardless of the report it induces, fixing his equilibrium belief that $e = H$ with probability one. ■

The proof of Proposition 4 exploits the fact that the original mechanism γ^o pays a flat (non-contingent) transfer when $\omega = b$ realizes, and an (m, s)-conditional transfer when $\omega = g$ realizes. At the renegotiation stage, the principal therefore cannot ~~hence~~ infer the agent's communication in state $\omega = b$, and so cannot condition his supplementary transfer in that state on the agent's report m in γ^o. This restricts his ability to construct a profitable supplementary offer: any offer that induces the agent to report $m = R$ in γ^o activates the punishment lottery and yields strictly less than V^{SB} (Step 2b), while any offer that induces $m = N$ must respect the non-negativity of final transfers, which under the assumed preferences forces the supplementary contract to provide weakly more insurance than c^{SB} at no improvement in the principal's payoff (Step 2a). ~~This prevents him from threatening to renegotiate by conditioning his supplementary transfer to the agent reporting $m = N$ in γ^o. As a consequence, any renegotiation offer must either induce the agent to report $m = R$ in γ^o, or sustain a large incentive-compatibility cost of inducing $m = N$, which makes it unprofitable.~~

References

Akbarpour, Mohammad and Shengwu Li, "Credible Auctions: A Trilemma," *Econometrica*, March 2020, *88* (2), 425–467.

Attar, Andrea and Arnold Chassagnon, "On Moral Hazard and Nonexclusive Contracts," *Journal of Mathematical Economics*, 2009, *45(9-10)*, 511–525.

_ **, Catherine Casamatta, Arnold Chassagnon, and Jean-Paul Decamps**, "Multiple Lenders, Strategic Default, and Covenants," *American Economic Journal: Microeconomics*, May 2019, *11* (2), 98–130.

_ **, Eloisa Campioni, Thomas Mariotti, and Alessandro Pavan**, "Keeping Agents in the Dark: Competing Mechanisms, Private Disclosures and the Revelation Principle," *TSE Working Paper*, June 2025, *21* (1227).

_ **, Thomas Mariotti, and François Salanié**, "Nonexclusive Competition in the Market for Lemons," *Econometrica*, 2011, *79(6)*, 1869–1918.

Bester, Helmut and Roland Strausz, "Contracting with imperfect commitment and noisy communication," *Journal of Economic Theory*, 2007, *136*, 236–259.

Bisin, Alberto and Danilo Guaitoli, "Moral Hazard and Nonexclusive Contracts," *RAND Journal of Economics*, 2004, *35(2)*, 306–328.

Bolton, Patrick, "Renegotiation and the dynamics of contract design," *European Economic Review*, May 1990, *34* (2-3), 303–310.

Brzustowski, Thomas, Alkis Georgiadis-Harris, and Balasz Szentes, "Smart Contracts and the Coase Conjecture," *American Economic Review*, 2023, *113(5)*, 1334–1359.

Catalini, Christian and Joshua S. Gans, "Some simple economics of the blockchain," *Communications of the ACM*, 2020, *63* (7), 80–90.

Chade, Hector and Edward Schlee, "Optimal insurance with adverse selection," *Theoretical Economics*, 2012, *7* (3), 571–607.

Davis, Kevin E., "The Demand For Immutable Contracts: Another Look At The Law And Economics Of Contract Modifications," *New York University Law Review*, May 2006, *81*, 487–549.

Dewatripont, Mathias, "Renegotiation and Information Revelation Over Time: The Case of Optimal Labor Contracts," *The Quarterly Journal of Economics*, 1989, *104* (3), 589–619.

Doval, Laura and Vasiliki Skreta, "Mechanism design with limited commitment," *Econometrica*, 2022, *90* (4), 1463–1500.

_ **and** _ , "Optimal mechanism for the sale of a durable good," *Theoretical Economics*, 2024, *19* (2).

Ebrahimi, Amir M, Bram Adams, Gustavo A Oliva, and Ahmed E Hassan, "A large-scale exploratory study on the proxy pattern in ethereum," *Empirical Software Engineering*, 2024, *29* (4), 81.

Forges, Francoise, "An approach to communication equilibria," *Econometrica: Journal of the Econometric Society*, 1986, pp. 1375–1385.

Fudenberg, Drew and Jean Tirole, "Moral hazard and renegotiation in agency contracts," *Econometrica*, 1990, *58* (6), 1279–1319.

Hart, Oliver and John Moore, "Foundations of incomplete contracts," *The Review of Economic Studies*, 1999, *66* (1), 115–138.

Hart, Oliver D. and Jean Tirole, "Contract Renegotiation and Coasian Dynamics," *The Review of Economic Studies*, 1988, *55* (4), 509–540.

Jolls, Christine, "Contracts as Bilateral Commitments: A new Perspective on Contract Modification," *Journal of Legal Studies*, 1997, *26*, 203–237.

Laffont, Jean-Jacques and Jean Tirole, "Adverse Selection and Renegotiation in Procurement," *The Review of Economic Studies*, 1990, *57* (4), 597–625.

Lomys, Niccolò and Takuro Yamashita, "A mediator approach to mechanism design with limited commitment," *Available at SSRN 4116543*, 2022.

Ma, Ching-To Albert, "Renegotiation and optimality in agency contracts," *The Review of Economic Studies*, 1994, *61* (1), 109–129.

Maestri, Lucas, "Dynamic contracting under adverse selection and renegotiation," *Journal of Economic Theory*, 2017, *171*, 136–173.

Maskin, Eric and Jean Tirole, "Unforeseen contingencies and incomplete contracts," *The Review of Economic Studies*, 1999, *66* (1), 83–114.

Micali, Silvio, Michael Rabin, and Salil Vadhan, "Verifiable Random Functions," in "40th Annual Symposium on Foundations of Computer Science" IEEE 1999, pp. 120–130.

Myerson, Roger B., "Optimal coordination mechanisms in generalized principal-agent problems," *Journal of mathematical economics*, 1982, *10* (1), 67–81.

_ , "Multistage Games with Communication," *Econometrica*, March 1986, *54* (2), 323–358.

Narayanan, Arvind, Joseph Bonneau, Edward Felten, Andrew Miller, and Steven Goldfeder, *Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction*, Princeton University Press, 2016.

Netzer, Nick and Florian Scheuer, "Competitive markets without commitment," *Journal of political economy*, 2010, *118* (6), 1079–1109.

Qi, Xuyuanda, Giulio Trigilia, and Kostas Koufopoulos, "Dynamic monopoly power and the Coase conjecture," *Available at SSRN*, 2024.

Rahman, David and Ichiro Obara, "Mediated partnerships," *Econometrica*, 2010, *78* (1), 285–308.

Rochet, Jean-Charles and Lars A Stole, "Nonlinear pricing with random participation," *The Review of Economic Studies*, 2002, *69* (1), 277–311.

Roughgarden, Tim, "Transaction Fee Mechanism Design," Papers, arXiv.org June 2021.

Salehi, Mehdi, Jeremy Clark, and Mohammad Mannan, "Not so immutable: Upgradeability of smart contracts on ethereum," in "International Conference on Financial Cryptography and Data Security" Springer 2022, pp. 539–554.

Stiglitz, Joseph E, "Monopoly, non-linear pricing and imperfect information: the insurance market," *The Review of Economic Studies*, 1977, *44* (3), 407–430.

Strulovici, Bruno, "Contract Negotiation and the Coase Conjecture: a Strategic Foundation for Renegotiation Proof Contracts," *Econometrica*, March 2017, *85*, 585–616.

Szabo, Nick, "Smart Contracts: Building Blocks for Digital Markets," 1996. Accessed on October 3, 2024.

Townsend, Robert M., *Distributed Ledgers: Design and Regulation of Financial Infrastructure and Payment Systems*, MIT Press, 2020.

Wang, Dingding, Jianting He, Siwei Wu, Yajin Zhou, Lei Wu, and Cong Wang, "The Dark Side of Upgrades: Uncovering Security Risks in Smart Contract Upgrades," *arXiv preprint arXiv:2508.02145*, 2025.