WORKING PAPERS

N° 1710

February 2026

# "Self-Revealing Renegotiation"

Andrea Attar, Lorenzo Bozzoli, Roland Strausz

Toulouse School of Economics

# Self-Revealing Renegotiation

Andrea Attar, Lorenzo Bozzoli, Roland Strausz*

February 11, 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). In 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, the construction achieves 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 off-path punishments, which she activates after observing a renegotiation offer. It verifies standard commitment assumptions by only conditioning decisions on public information, without requiring any third-party enforcement. In practical terms, it can be implemented with 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: Università degli Studi di Roma "Tor Vergata" (email: lorenzo.bozzoli@uniroma2.it); 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, 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, embodying the problem of collective opportunism that inherently emerges when dealing with incentive problems. As first pointed out by Dewatripont (1989), this opportunism arises because contracts that optimally resolve incentive problems typically do so by implementing allocations that prove inefficient ex post. Consequently, when contracting parties are unable to credibly commit to refraining from renegotiating away ex post inefficiencies, they find themselves at a disadvantage from an ex-ante perspective.

The inability to prevent renegotiation reflects a conflict 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 through mechanisms that combine 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 over the course of the interaction but becomes verifiable at the contract execution stage. We show that a simple architecture of communication, which only involves a binary message and a coin flip, allows to retrieve the full-commitment allocation without any third-party commitment device. By implementing the second-best allocation via design, we thus overturn the conventional wisdom that the inability to prevent renegotiation in contractual terms fundamentally constrains incentive provision under asymmetric information.

We illustrate our approach in the canonical model of Fudenberg and Tirole (1990) (FT), in which a risk-neutral principal contracts with a risk-averse agent who privately chooses a binary effort level. In this moral hazard setup, the incentive-compatible transfers for high effort entail ex-post inefficient risk sharing, creating scope for Pareto-improving renegotiation. FT show that this renegotiation threat prevents achieving second-best efficiency when the principal is restricted to using revelation mechanisms. We show that self-revealing mechanisms—which incorporate both bidirectional communication and conditional disclosure—fully mitigate this threat.

[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."

Specifically, we construct a self-revealing mechanism which *uniquely* implements the second-best allocation: high effort is chosen deterministically in the unique equilibrium outcome. While FT analyze the renegotiation game using revelation mechanisms, which only require the agent to report her private information, we set up an alternative mode of communication to prevent renegotiation. Our approach draws on the dynamic mechanism design principles of Forges (1986) and Myerson (1986) but uses signals for a distinct purpose: rather than correlating players' behaviors, our mechanism exploits private signals to the agent to generate off-equilibrium punishments that deter renegotiation.

Thus, enriching the structure of communication allows to reconcile the conflict between renegotiation-proofness and economic efficiency: our optimal mechanism successfully prevents renegotiation while implementing the same allocation that would obtain under full commitment.

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 modifies the contract by either improving or worsening her expected utility depending on the coin toss.

Intuitively, reporting *renegotiation* enables the agent to trigger a random automated counter-offer, whose outcome is privately revealed to the agent. Thus, when a renegotiation is proposed, and the agent reports this event, she accepts the new offer only when the randomization implemented by the original mechanism is unfavorable to her. This makes any attempt to renegotiate prohibitively costly to the principal. In equilibrium, the agent's self-enforcing punishment fully prevents renegotiation, leading to a unique implementation of the second-best allocation. The random counter-offer is key to our result: it requires the mechanism to send private signals to the agent, creating crucial informational asymmetries.

While bidirectional communication is necessary, it is not sufficient for eliminating renegotiation threats. Our 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 contract executes. Our mechanism resolves this paradox through its self-revealing property: bidirectional communication remains private during renegotiation but becomes verifiably disclosed if the original contract is executed.

Our mechanism also verifies standard commitment requirements under renegotiation. By only conditioning decisions on public information, it requires no external enforcement

from a third-party. Moreover, the timing of the agent's communication need not be monitored: her self-interest ensures she finds strategically optimal to communicate in accordance with our construction.

In practical terms, our optimal mechanism is directly implementable through existing smart contract technologies. Using off-the-shelf commit-and-reveal cryptographic tools, a contracting party can privately encode a message on a blockchain and later disclose it publicly: this directly mirrors the self-revealing functionality of keeping bidirectional communication private and revealing it verifiably upon contract execution. We provide a proof-of-concept implementation in Solidity (v0.8.0), demonstrating compatibility with current smart-contract toolkits.

Our approach extends beyond the FT framework. The key insight—bidirectional communication with strategically timed information disclosure generates self-enforcing punishments against renegotiation—applies wherever ex-post inefficiencies create renegotiation incentives. To support this view, we extend our approach to other settings of contract renegotiation.

We first consider alternative extensive forms for the renegotiation game, focusing, in particular, on the case of infinite rounds of renegotiation. We hence model, in the spirit of Strulovici (2017), an infinite-horizon setting where renegotiation breaks down with positive probability in each round, in which case the last accepted contract executes. In our moral hazard context, second-best efficiency involves imperfect risk sharing, which leaves in principle room for renegotiation after each round. We instead construct a self-revealing mechanism, offered at the ex-ante stage, which implements the second-best allocation thereby suggesting that backward induction reasoning is *not* key to our approach.

We next argue that self-revealing mechanisms retain their power under alternative specification of the renegotiation process. We consider, in particular, the situations in which renegotiation may also complement, and need not necessarily replace, the original mechanism. This gives rise to a richer set of renegotiation opportunities: a new offer may exploit the observability of the original mechanism's transfers to undo any potential punishment. We show, however, that a modified self-revealing revealing mechanism can be designed to prevent these additional effects, and implement the second-best allocation even under this supplementary view of renegotiation.

The broader implication of our analysis is therefore that mechanism design is 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.** Our work contributes to the literature on contract renegotiation which, starting with Dewatripont (1989), focuses on optimal renegotiation-proof mechanisms. The renegotiation threat has been typically assessed under different informational assumptions: while Fudenberg and Tirole (1990) and Ma (1994) consider a moral hazard scenario, renegotiation under incomplete information is analyzed by Hart and Tirole (1988), Laffont and Tirole (1990), and, more recently, by Maestri (2017). Strulovici (2017) establishes a non-cooperative foundation for Coasian outcomes in an infinite-horizon renegotiation game.

Bolton (1990) points out that optimally preventing renegotiation requires introducing private information at the renegotiation stage. Without mechanisms that send signals to agents, this information can only be generated through agent randomization over reports or efforts at equilibrium. Such randomization implies allocative costs, making second-best efficiency unattainable.

Instead, our self-revealing mechanism generates private information through signals sent to the agent, eliminating these randomization costs. Critically, this uncertainty matters only off-equilibrium: following a renegotiation offer, the mechanism's signals enable the agent to punish renegotiation attempts, uniquely implementing the second-best allocation.

Formally, our approach draws on the dynamic mechanism design tradition initiated by Forges (1986) and Myerson (1986). In contrast with these early works, we explicitly consider an extensive form game in which a player, i.e. 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.

Rahman and Obara (2010) achieve virtual implementation through mediated contracts conditioning on private communication, but do not address contractual enforceability. By contrast, we achieve full implementation and our conditional revelation—disclosing communication only when the original contract executes—provides explicit enforcement power so that they are implementable without a trusted third party (mediator).

Bester and Strausz (2007) are the first to develop the 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.

The renegotiation problem can be rationalized as the 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 and identifies its methodological limitations. 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. Proofs are in Appendix A.

[2]See Bisin and Guaitoli, 2004; Attar and Chassagnon, 2009; Attar et al., 2011; Attar et al., 2019.

# 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 $b$, where $g > b > 0$. The probability distribution over outputs depends on the binary effort $e \in E \equiv \{L, H\}$. 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 $Y_e \equiv p_e g + (1 - p_e)b$.

**Payoffs 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 utility 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, with $c$ represented in utility space unless noted.

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

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

where $U^0$ is her reservation payoff.[4] The principal's expected payoff 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 payoff $U \in \mathbb{R}$. We also define, for each $e \in E$, the function $V_e^{FI} : \mathbb{R} \to \mathbb{R}$ where

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

[3]These assumptions are all directly taken from FT. As we explicitly show in Appendix C, FT's unboundedness assumption is not crucial to our results.

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

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

With observable effort, the principal's optimal contract induces efficient risk-sharing while guaranteeing the agent her reservation payoff $U^0$. We refer to $c^{FB} \equiv c_H^{FI}(U^0)$ 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 payoff $U$ 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 payoff 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) binds at the solution, implying that the second-best contract is $c^{SB} \equiv c^{IC}(U^0)$. The second-best, ex-ante efficient, allocation $(H, c^{SB})$ yields $V^{SB} \equiv V_H^{IC}(U^0)$ to the principal, and $U_H(c^{IC}(U^0)) = U^0$ 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. The impact of this renegotiation threat is assessed in a non-cooperative game between the principal at the contract design stage and his future self at the interim stage. The timing of this game is 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.

(*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:

(*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 $\rho \in \{y, n\}$. Acceptance implies that $c$ is replaced by $c^r$.

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

Stages $(i) - (vi)$ define the *primitive* game $G$, which captures the physical constraints arising under renegotiation. The game embodies the following assumptions:

*A*.1. The renegotiation offer $c^r$ is made only once, i.e., at stage $(iv)$. As already argued by FT, any *finite* number $k$ of renegotiation rounds does not add any strategic effect: all bargaining would occur in the last round, making the analysis equivalent to the single-round case.[6]

*A*.2. The original offer $c$ *cannot* condition on the renegotiation offer $c^r$ or on the agent's decision $\rho$ at stage $(v)$. This captures the legal doctrine that parties cannot prevent renegotiation contractually.

*A*.3. If $c^r = \emptyset$, both parties remain bound to the original contract $c$. This reflects the legal doctrine requiring mutual consent for contract modification.

*A*.4. If $c^r \neq \emptyset$, the contract $c^r$ replaces contract $c$ only if the agent accepts it at stage $(v)$ by declaring $\rho = y$. In this case, $c$ becomes irrelevant. By contrast, if the agent rejects $c^r$ in stage $(v)$ by declaring $\rho = n$, the contract $c^r$ becomes irrelevant.

*A*.5. Contracts $c$ and $c^r \neq \emptyset$ are exclusive; at most one executes at stage $(vi)$.[7]

[6] See Section 6B in Fudenberg and Tirole (1990). In Section 5.1, we explicitly analyze the case with infinite rounds of renegotiation in the spirit of Strulovici (2017).

[7] 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.

**FT's Renegotiation Game.** FT show that, for any probability $x \in (0,1)$ that the agent selects $e = H$ at stage $(iii)$, the renegotiation stage $(iv)$ corresponds to Stiglitz (1977)'s seminal setting of 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 mechanism, 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 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 $\gamma_c$ 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 $e = L$.

When characterizing the equilibria of $G_C$, FT exploit the *renegotiation-proofness principle*, and argue it is without loss to focus on the principal offering a mechanism in stage $(i)$ that is not renegotiated on the equilibrium path. Restricting attention to revelation mechanisms $\gamma_c \in C$, FT then show that, in the unique (perfect Bayesian) equilibrium allocation of $G_C$, $e = H$ is only implemented with probability $x^{FT} < 1$.

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 on-path renegotiation in the latter case).[8] 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. As we shall argue, considering this new channel of communication allows to reconcile renegotiation-proofness of the optimal mechanism with second-best, ex-ante, efficiency of the equilibrium allocation.

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

# 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 explicitly recognize the dynamic nature of this game. Thus, we introduce bidirectional communication, as emphasized in the dynamic mechanism design frameworks of Forges (1986) and Myerson (1986).

Separating the design of communication from that of final transfers, we write a dynamic mechanism $(\mathcal{C}, \tau)$ in terms of two elements. First, a *communication protocol* $\mathcal{C}$ that specifies the (possibly bidirectional) communication exchanged at each stage $t$, including how stage-$t$ signals are generated following any history. Second, a *decision rule* $\tau$ that maps communication into final transfers.

Following dynamic mechanism design, 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 communication history *at the final payout stage*. This defines our notion of *self-revealing* mechanisms, extending canonical dynamic mechanisms by conditioning revelation on execution.

In our dynamic contracting framework, these self-revealing 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 contractually enforceable in standard contract-theoretic terms and do not require third-party mediation.[9] 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. 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 $\sigma(h|m) = \sigma(t|m) = 1/2$ for each $m \in \mathcal{M}$.

[9]Section 4.3 explicitlly 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.

4. The agent sends $m$ and receives $s$ *before* her participation decision $\rho$.[10]

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 $\Gamma$.

In the remainder of this section, we let the principal design mechanisms in the class $\Gamma$ 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 $G_\Gamma$.

## 3.1 The Self-Revealing Renegotiation Game $G_\Gamma$

Allowing the principal to select self-revealing mechanisms from $\Gamma$ modifies the primitive game $G$ into the extensive-form game $G_\Gamma$ 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 $\gamma$. If she rejects, the game ends and 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 $s \in \{h, t\}$. After this bidirectional communication, if $\gamma^r \neq \emptyset$, she publicly accepts or rejects $\gamma^r$ by declaring $\rho \in \{y, n\}$. Acceptance implies that $\gamma$ is replaced by $\gamma^r$.

($vi$) The communication $(m, s)$ from stage ($v$) is publicly revealed if and only if $\gamma$ executes (i.e., either $\gamma^r = \emptyset$ or $\rho = n$) in which case transfers are determined by $\tau(m, s)$. If $\rho = y$, transfers are determined by a report $m^r \in E$ sent by the agent in $\gamma^r$. Nature publicly draws the output realization $g$ or $b$ and conditional transfers are executed.

In $G_\Gamma$, 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 involves no loss of generality.

[10] 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.

To see this, note that a self-revealing mechanism $\gamma$ at stage $(i)$ constrains feasible renegotiation offers at stage $(iv)$ in two ways. First, no offer can contractually condition on the private communication $(m, s)$: if the agent accepts $\gamma^r$, the original mechanism $\gamma$ does not execute, so $(m, s)$ are never publicly revealed. Second, rejected offers are payoff-irrelevant.

Given these constraints, the renegotiation stage becomes a mechanism design problem where the principal faces an agent with private information $(e, m, s)$. Her preferences over contracts within $\gamma^r$, however, depend solely on $e$: an agent with a given $e$ but different $(m, s)$ evaluates any contract $c$ identically via $U_e(c)$. Consequently, for any $(m, s)$, she has the same set of optimal reports in any mechanism $\gamma^r$, regardless of its message space. This implies that the principal cannot screen on $(m, s)$. Although different $(m, s)$ may correspond to different outside options in the original mechanism $\gamma$ for the agent, this heterogeneity only affects *whether* she accepts $\gamma^r$. Thus, the information relative to $(m, s)$ cannot be elicited via screening. This implies that restricting a renegotiation offer to be a revelation mechanism $\gamma^r \in C$ is without loss of generality.

A (pure) strategy for the principal in $G_\Gamma$ 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 $\lambda$ in $G_\Gamma$ has three components. First, it associates with any $\gamma \in \Gamma$ a probability. Second, for any history $(\gamma, e, \gamma^r)$, the strategy $\lambda$ specifies a probability distribution over messages $m \in \mathcal{M}$. Third, for any $(\gamma, e, \gamma^r \neq \emptyset, m, s)$, the strategy $\lambda$ specifies whether to accept or reject by declaring $\rho \in \{y, n\}$. Finally, for $\gamma^r \neq \emptyset$ and $\rho = y$, the strategy $\lambda$ specifies a message $m^r \in E$ in the renegotiated mechanism $\gamma^r$.

We consider the perfect Bayesian equilibria (henceforth equilibria) of $G_\Gamma$.[11] 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 $G_\Gamma$ must induce an equilibrium in each $G_\Gamma(\gamma)$. Therefore, in an equilibrium of $G_\Gamma$, the principal chooses an optimal mechanism $\gamma$ anticipating that 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 $G_\Gamma$ differs from FT's $G_C$ 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 $\Gamma$ add bidirectional communication: next to the agent sending messages,

[11] The principal has only one information set in the game $G_\Gamma$, 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.

she also receives 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 $G_\Gamma$: 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:

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

Intuitively, $\gamma^*$ sets the second-best contract $c^{SB}$ as the "default" one, which the agent may get by sending $m = N$. The mechanism also allows the agent to trigger a random "counter-offer" by sending $m = R$, indicating that the principal made a renegotiation offer. The realization of this lottery may either increase the agent's utility by $\Delta U$ or decrease it by $\Delta U$. For any $e \in E$, the counter-offer yields the agent the same expected utility as $c^{SB}$ but costs the principal more since $\Phi$ is convex and $V_e^{IC}$ is concave in $U$. The principal views the counter-offer as random, whereas the agent observes its realization after sending $m = R$ but before deciding whether to accept the principal's renegotiation offer.

The next lemma guarantees that we can find a $\Delta U$ large enough to induce an agent's behavior that prevents renegotiation.

**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 $\Delta U$ to the agent. Additionally, the principal prefers $c^{SB}$ to a 50-50 lottery between the full-insurance contract leaving $\Delta U$ less to the agent, and the incentive-compatible one leaving the agent an extra utility $\Delta U$. 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 payoff from offers the agent always accepts; the second bounds his payoff 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 $\gamma^r$ and let $\hat{U}_e^r$ denote her corresponding payoff upon acceptance:

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

Consider the following strategy profile $\{\gamma^r(\gamma^*), \lambda(\gamma^*)\}$: The principal does not renegotiate, $\gamma^r(\gamma^*) = \emptyset$. 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 $\gamma^*$ depends on the principal's offer $\gamma^r$ as follows:

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

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

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

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 $\gamma^r$ as defined in (3).

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 $x = 1$, constitute an equilibrium of $G_\Gamma(\gamma^*)$, yielding the claimed allocation $(H, c^{SB})$.

Note first that the only non-trivial information set for the principal in $G_\Gamma(\gamma^*)$ is at the renegotiation stage, when he offers $\gamma^r$. 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 develop the remainder of our arguments in two lemmas whose formal proofs appear in Appendix A.

**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 $\gamma^*$ after an offer $\gamma^r$ 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 dominates in expected terms the payoff associated with the report $m = N$. In addition, the proof 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 payoff $U^0 + \Delta U$ that she gets with probability 1/2 by reporting $m = R$ to $\gamma^*$ 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 payoff $U^0$, Proposition 1 implies that the game $G_\Gamma$ has *an* equilibrium in which the renegotiation threat is fully mitigated. The result stands in stark contrast to FT, who restrict the principal to offer revelation mechanisms at the initial stage.

The proof establishes that $\gamma^*$ makes any renegotiation unprofitable. For offers yielding payoffs in $(U^0 - \Delta U, U^0 + \Delta U]$, the agent sends $m = R$, triggering the punishment mechanism. In FT, renegotiation-proofness and second-best efficiency are incompatible: achieving one precludes the other. We reconcile this conflict through $\gamma^*$, which implements the second-best efficient allocation, and exploits signals to generate off-equilibrium punishments that prevent any renegotiation. To see this, consider the renegotiation offer that undermines $x = 1$ in FT. Against $\gamma^*$, the agent effectively reveals this offer by sending $m = R$ and then rejects it with probability 1/2 (when $s = t$). This random rejection, triggered by the agent's private signal, makes renegotiation unprofitable to the principal given the unfavorable terms associated with it.

Because this punishment hinges on the random signal $s$, it effectively implements a randomized contract: the agent's actual payoff 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 following Chade and Schlee (2012), the optimal renegotiation offer against any distribution over efforts chosen by the agent is deterministic. Anticipating such deterministic renegotiation, the principal cannot benefit from committing ex ante to stochastic mechanisms.[12]

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 $\gamma^*$ 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 payoff 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-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) show 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 C that Proposition 1 extends to cases where the agent's utility is CRRA,

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

accommodating limited liability constraints that bound transfers below.

### 3.3 Unique Implementation of the Second Best

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

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

The proof of Proposition 2 in Appendix A constructs a mechanism $\gamma_\varepsilon$ by perturbing $\gamma^*$ in a way that allows us to break all the agent's indifferences at the root of equilibrium multiplicity. Specifically, $\gamma_\varepsilon$ 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 $\gamma_\varepsilon$ 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 $\gamma_\varepsilon$ at the initial stage, the principal can hence guarantee himself a payoff arbitrarily close to $V^{SB}$, which obtains under full commitment. Since $V^{SB}$ is also an upper bound, uniqueness of the equilibrium allocation follows.

## 4 A New Approach to Renegotiation Proofness

In Section 3, we constructed a renegotiation-proof mechanism that uniquely implements the allocation $(H, c^{SB})$ which obtains under full commitment. This contrasts with the

approach followed by FT: they apply the renegotiation-proofness principle to characterize an equilibrium allocation, which fails to be second-best efficient. This failure arises because FT restrict to revelation mechanisms, which do not incorporate private signals to the agent. By contrast, our construction exploits the interaction between the agent's reports and the signals she receives, to generate a new set of punishments which successfully deter renegotiation.

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 $\gamma^*$: 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.

**A verifiability paradox.** The dynamic 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 $\gamma^*$ resolves this *verifiability paradox* through its self-revealing property. Communication remains private throughout the renegotiation stage but becomes publicly revealed 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, privacy guarantees that any profitable renegotiation offer is rejected with probability 1/2. In such cases, the self-revelation requirement applies, and communication can be made public at no cost. That is, we require no trusted third party to observe communication and execute transfers on behalf of the parties. In particular, there is no need for any multilateral payment system, which would itself be susceptible

to renegotiation by additional parties. The optimal mechanism $\gamma^*$ only requires standard contract enforcement by courts.

**A self-enforcing communication protocol.** The mechanism $\gamma^*$ is structurally simple, featuring a binary message space $\{N, R\}$ and a binary signal $\{h, t\}$ generated by a fair coin toss. This matches the complexity of FT's revelation mechanisms, which rely on binary reports $\{H, L\}$ but do not involve signals.

Our approach differs from FT in the specific protocol we adopt. The construction in FT is, in principle, compatible with several communication protocols, since the stage at which the agent sends her message is immaterial.[13] Our construction, instead, exploits the agent communicating at stage $(v)$, after receiving a potential renegotiation offer but before accepting it. This raises a question: does our mechanism effectively require courts to verify adherence to this communication protocol?

The mechanism itself ensures compliance through the agent's strategic incentives. Even if courts cannot verify when the agent communicates, she finds it optimal to communicate at stage $(v)$ rather than earlier or later. Communicating after observing $\gamma^r$ allows her to condition her message on whether renegotiation was attempted, maximizing the information she can exploit. We formalize this intuition in Appendix C by constructing a protocol that delegates to the agent the choice of her communication timing: in any pure strategy equilibrium, 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 occurred. The protocol operates *under the shadow of the court*: by delegating timing choice to the agent, the mechanism aligns her strategic interests with the required protocol.

## 4.2 The Commitment Requirements of $\gamma^*$

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

The renegotiation game $G_\Gamma$ reveals the specific commitment assumptions we exploit. 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

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

or occurrence of the agent's communication in $\gamma$ is infeasible by construction. At the renegotiation stage, the principal therefore has the same commitment power as in FT.

When designing a mechanism $\gamma$ at stage $(i)$, the principal is bound by the legal doctrine of renegotiation. As in FT, he cannot commit to the features or occurrence of a renegotiation offer. Furthermore, replacing the initial offer requires mutual consent. This mutual consent has an important implication: 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 $\gamma^*$ before responding. The mechanism grants the agent an explicit right to send message $m$ and receive signal $s$ before deciding on any renegotiation offer. Because renegotiation requires mutual consent, the principal cannot unilaterally revoke this right. Moreover, because the agent's communication is private, the principal cannot condition his offer on the agent not having communicated. By committing to a mechanism that makes the agent privately informed of counter-offers, the principal completely offsets the renegotiation threat.

## 4.3 From Theory to Practice: Smart Contracts Implementation

We next show how current smart contract technologies provide the natural tool—the "commit-and-reveal" technique—to implement our optimal mechanism in practice. By design, conditional transfers become enforceable by standard means, once the mechanism self-reveals its private information. Hence, the implementation challenge centers entirely on the communication protocol.

Indeed, current smart contract technologies cannot send private signals to players, which may conflict with the privacy requirements of our mechanism $\gamma^*$.[14] We solve this issue by modifying $\gamma^*$ 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 let the signal $s \in \mathcal{S}^*$ remain representing a fair coin toss: $\sigma^{**}(h|m) = \sigma^{**}(t|m) = 1/2$ for all $m \in \mathcal{M}^{**}$.[15] However, we let the signal $s$ be publicly observable, with the implication that a renegotiating offer can condition on its

[14]Note that if the signal $s$ were public rather than private, the principal could make signal-conditional renegotiation offers that undermine the effectiveness of $\gamma^*$ 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.

[15]In practice, $s$ can be instantiated via a verifiable on-chain randomness source (e.g., a VRF 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.

realizations, and define its decision rule $\tau^{**}$ as

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

Effectively, $\gamma^{**}$ allows the agent the option between two random counter-offers, which only differ by the face of the coin flip that leads to the better or worse contract. Thus $\gamma^{**}$ 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 $\gamma^{**}$ still implements the second-best allocation.[16] 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 (payoff $U^0 + \Delta U$) that make rejecting renegotiation optimal. This random rejection punishes the principal in expectation, deterring renegotiation just as in the original mechanism $\gamma^*$. For instance, in the intermediate region, the principal's expected payoff 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 $\gamma^{**}$ 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, $\gamma^{**}$ can be implemented via smart contracts that are self-executing programs on transparent blockchains.[17] While this may seem paradoxical given that blockchain transactions are publicly recorded, cryptographic techniques allow us to achieve the required privacy within this transparent environment.

[16] This is shown formally in Appendix C.

[17] 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, in general, an enforcement of smart contracts depends on the shadow of the law. To see this in our specific context of $\gamma^{**}$, 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.

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.[18]

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 B a complete Solidity smart contract that implements $\gamma^{**}$ 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,[19] 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., 2024).

By explicitly allowing both contract termination and modification, these adaptability functions reintroduce classic time-consistency concerns in the smart contracts paradigm.[20] We regard our results as relevant in this respect: the finding that blockchain-compatible

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

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

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

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,[21] and *timelocks* between the approval and implementation of upgrades, which give users time to assess and exit the contract before changes take effect.[22]

# 5 The Power of Self-Revealing Mechanisms

Traditional approaches to renegotiation design, as summarized by Bolton (1990), share a common principle: optimal renegotiation-proof mechanisms require private information for the agent at the renegotiation stage, leaving the principal uncertain when he attempts to renegotiate the original mechanism.

In the FT context, this is achieved by having the agent randomize over her efforts at equilibrium. Such randomization requires making the agent indifferent over several alternatives. This imposes an allocative cost for incentive-compatibility reasons. By contrast, our optimal mechanism avoids these incentive-compatibility costs, fully mitigating the renegotiation threat without sacrificing second-best efficiency.

In this section, we show that these insights extend to other settings of contract renegotiation, reinforcing the general relevance of our approach. We first consider extensions to standard renegotiation frameworks with alternative extensive forms, focusing on the case in which infinite rounds of renegotiation are allowed. We then evaluate the implications of richer renegotiation opportunities, thereby discussing the supplementary view of renegotiation.

## 5.1 Alternative Extensive Forms

Since the self-revealing mechanism $\gamma^*$ operates *after* effort is chosen, the results in Propositions 1 and 2 naturally extend to any countable or continuous effort spaces.[23] The same observation guarantees that our approach does not exploit the principal's specific objective function. For instance, to address the government failure emphasized by Netzer and

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

[22] "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).

[23] This parallels the extension in Fudenberg and Tirole (1990, Section 5.A).

Scheuer (2010), a utilitarian planner can rely on a modified version of $\gamma^*$ to implement second-best insurance under renegotiation threats.

Propositions 1 and 2 also extend to situations in which the agent, rather than the principal, initiates renegotiation, as analyzed by Ma (1994). Establishing these extensions requires constructing a mechanism where the principal implements punishments through his own private communication. This approach yields unique implementation even when renegotiation threats originate from the agent, contrasting with the equilibrium multiplicity in Ma (1994).[24]

As noted in Assumption $A.1$, our results are robust to any *finite* number of renegotiation rounds. However, the second-best allocation implemented by Proposition 1 involves inefficient risk sharing, leaving room for further renegotiation after each round. This raises the question: do our results depend on the number of rounds being finite?

To show they do not, 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, breakdown round $T^*$ follows a geometric distribution: $\Pr(T^* = T) = (1-\eta)^{T-1} \cdot \eta$ and $\Pr(T^* = T' \mid T^* \geq T) = (1-\eta)^{T'-T} \cdot \eta$. 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 $\eta$, we denote $G^\eta$ the corresponding primitive game, which extends the game $G$ by allowing for infinite renegotiation rounds.

We construct a self-revealing mechanism $\xi^{0*}$, offered at the ex-ante stage, which implements the second-best allocation in this context. The mechanism $\xi^{0*}$, 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 $G^\eta$ with mechanisms selected from $\Xi$, after $\xi^{0*}$ 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 $\xi^T \in \Xi \cup \{\emptyset\}$.

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

[24] Formal construction available from the authors.

*T.iii*) The agent accepts ($\rho^T = y$) or rejects ($\rho^T = n$) the renegotiation offer $\xi^T$, 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 $\xi^T$. 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 $\xi^{0*}$ requires the agent to submit a report $m^{0*}_T \in \{N, R\}$ in each round $T.ii)$, i.e. after a renegotiation offer $\xi^T$ 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 $\gamma^*$ to an infinite horizon.

The principal may attempt to renegotiate $\xi^{0*}$ at any $T \geq 1$, until $T^*$ realizes. Observe that the set of feasible renegotiation offers at each round $\xi^T$ is taken to be $\Xi$. That is, any renegotiated mechanism, once accepted, features the same communication protocol as that relative to $\xi^{0*}$ after round $T$. 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 C, exploits the idea that any renegotiation offer can be simply characterized in terms of the continuation utility it yields to the agent. Thus, there is no loss of generality in considering that any $\xi^T$ features the same communication protocol associated to $\xi^{0*}$ after round $T$. Any attempt to renegotiate can hence be punished following the same logic as developed in the proof of Proposition 1.

## 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 $A.4$).[25] Under this view, the agent cannot combine $\gamma^*$ with a renegotiation offer—contracts are exclusive (Assumption $A.5$). We now examine *supplementary* renegotiation, where such combinations are possible. This creates new strategic considerations.

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

Upon observing $\gamma_+^r$, the agent finds it optimal to report $m = N$ in $\gamma^*$, triggering the equilibrium transfers (as reporting $m = R$ would trigger the punishment lottery, causing $\gamma_+^r$ to pay nothing since its transfer is contingent on $\gamma^*$ implementing $c^{SB}$). She then accepts $\gamma_+^r$, which offsets these transfers and implements $c_H^{FI}(U^0+\varepsilon)$, guaranteeing her a payoff $U^0 + \varepsilon > U^0$. The principal also gains: full insurance eliminates the risk premium embedded in $c^{SB}$, reducing expected transfers. For $\varepsilon$ sufficiently small, both parties are strictly better off. Thus, $\gamma^*$ is vulnerable to supplementary renegotiation: the principal can profitably deviate by conditioning on $\gamma^*$'s realized transfers.[26]

However, alternative mechanisms can restore the second-best also under supplementary renegotiation. Consider a modified self-revealing mechanism with the following structure: when that pays a flat transfer if 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 from observing both the transfer and the realized output, eliminating the vulnerability demonstrated above.[27]

This modification 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 $\gamma^*$, creating the uncertainty that deters renegotiation. Under supplementary rene-

[25] 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".

[26] While observability of the final transfers could in principle be exploited also under the replacement view of renegotiation, the exclusivity assumption at the root of this view (see A.5) makes any such conditional offer not strategically relevant. The profitability of the above renegotiation offer, indeed, crucially hinges on the possibility to combine it with the original one.

[27] Formal analysis available from the authors shows that Proposition 1 extends to supplementary renegotiation under CRRA preferences for the agent.

gotiation, an additional difficulty arises: while communication remains unverifiable, the principal can observe its consequences through the realized transfers from $\gamma^*$. 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 keeps messages private at the moment of renegotiation yet verifiable at execution. The combination of bidirectional communication and its strategically timed disclosure enables off-equilibrium punishments that restore the full-commitment second-best without distorting on-path incentives.

We establish these results in the canonical renegotiation framework of Fudenberg and Tirole (1990), and we show that they extend to several settings of renegotiation under moral hazard. We do not see any specific argument that limits the applicability of our approach to situations in which the agent holds some private information (e.g., Laffont and Tirole (1990)). However, a comprehensive analysis of the portability of our insights to frameworks of renegotiation under incomplete information is beyond the scope of the present work.

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 informa-

tion is disclosed can substitute for formal commitment, with implications for environments beyond the canonical model and for the governance of digital markets.

# A 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 $\Phi$. 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 $\Phi$ convex:

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

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 for any $(e, \gamma^r \neq \emptyset, m, s, y)$ with $m$ the agent's message $m \in \{N, R\}$ in the original self-revealing mechanism $\gamma$. From comparing her payoff $U_e(\tau^*(m, s))$ of remaining in $\gamma^*$ with her utility $\hat{U}_e^r$ of accepting $\gamma^r$, it follows that, at each history $(e, \gamma^r \neq \emptyset, m, s)$, the agent's participation behavior is optimal. Next, consider any history $(e, \gamma^r \neq \emptyset)$. Because the agent observes $s$ before making her decision $\rho$, 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

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

where $\hat{U}_e^r$ is defined in (3).

Hence, it is optimal for the agent to send $m = N$ in $\gamma^*$ 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 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 $\gamma^*$, followed by $\rho = n$, 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 $m = R$, it is optimal for the agent to choose $\rho = y$ when $s = h$ (as rejection leads to $U^0 - \Delta U < \hat{U}_e^r$), and $\rho = n$ when $s = t$ (as rejection leads to $U^0 + \Delta U \geq \hat{U}_e^r$). We next argue that sending $m = R$ in $\gamma^*$, as prescribed by $\lambda(\gamma^*)$, is optimal. That is, the 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 $\hat{U}_e^r/2 + (U^0 + \Delta U)/2$. Reversing the inequality in (6), we only need to show that

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

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 satisfied by assumption.

(*iii*) If $\hat{U}_e^r \in (U^0 + \Delta U, \infty)$, then we have $U^0 < U^0 + \Delta U < \hat{U}_e^r$ 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 $\gamma^*$, and then accepts $\gamma^r$ 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 an optimal behavior of the agent since she obtains the same payoff $U^0$ under any report in $\gamma^*$. 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. This completes the proof of sequential rationality. ■

**Proof of Lemma 3:** In the subgame $G_\Gamma(\gamma^*)$, the principal's best response is either $\gamma^r = \emptyset$ or a revelation mechanism $\gamma^r \in C$ that maps the renegotiation report $m^r \in E$ to a contract $(u_g, u_b)$.[28] Given the principal's belief $x = 1$ and the agent's risk-aversion, any accepted $\gamma^r$ that maximizes the principal's payoff yields full insurance to the agent of type $e = H$: so, conditional on acceptance, his payoff equals $V_H^{FI}(\hat{U}_H^r)$ for some scalar $\hat{U}_H^r$.

Given belief $x = 1$ and the agent's strategy $\lambda(\gamma^*)$, we verify that the principal's expected payoff does not exceed $V^{SB} = V_H^{IC}(U^0)$ for any $\hat{U}_H^r \in \mathbb{R}$. We distinguish three cases:

(*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.

[28] $\gamma^r$ cannot directly condition on $(m, s)$: acceptance of $\gamma^r$ precludes $\gamma^*$ from disclosing $(m, s)$ publicly, and rejection of $\gamma^r$ makes it payoff-irrelevant.

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

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

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

Thus, the principal cannot gain by offering any $\gamma^r \neq \emptyset$ . ■

**Proof of Proposition 2**. We construct a mechanism $\gamma_\varepsilon$ that uniquely implements $e = H$ and yields a principal's payoff 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 $c_\varepsilon^{SB}$ yields the agent the payoff $U^0$ if she selects $e = H$, and $U^0 - \varepsilon$ if $e = L$.

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

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

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 now characterize all the agent's behavioral strategies $\lambda(\gamma_\varepsilon)$ that are sequentially rational in the subgame $G_\Gamma(\gamma_\varepsilon)$. Note that the agent's sequentially rational behavior depends on $\tau_\varepsilon$ and $\hat{U}_e^r$, but not on the principal's belief $x$. We start from the terminal nodes of $G_\Gamma(\gamma_\varepsilon)$.

Recalling (3), note that in any history $(e, \gamma^r, m, s, y)$ 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 $\gamma^r$ 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 payoff $U_e(\tau_\varepsilon(m,s))$ of staying in $\gamma_\varepsilon$ with the utility $\hat{U}_e^r$ of accepting $\gamma^r$:

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

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

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

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

(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 $\gamma^r$ yields $\hat{U}_e^r$ to the agent if accepted. For $e=H$, $m=N$ is optimal if

$$\max\{U^0,\hat{U}_H^r\}\geq\frac{1}{2}\max\left\{\hat{U}_H^r,U^0-\Delta U-\kappa\varepsilon\right\}+\frac{1}{2}\max\{\hat{U}_H^r,U^0+\Delta U\},\tag{10}$$

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

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

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 $\gamma^*$ 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 $m=N$ strictly dominates $m=R$ for both effort levels.

We now address the agent's optimal effort choice. A full characterization of the agent's best response to every possible principal strategy $\gamma^r(\gamma_\varepsilon)$ is unnecessary. To establish uniqueness, it suffices to show that when the principal plays the equilibrium strategy $\gamma^r(\gamma_\varepsilon)=\emptyset$, the agent's unique optimal effort is $e=H$. To see this, note that given $\gamma^r=\emptyset$, the agent's unique optimal report is $m=N$ regardless of $e$ (as shown above). This leads to the implementation of transfers $c_\varepsilon^{SB}$, which satisfy $U_H(c_\varepsilon^{SB})>U_L(c_\varepsilon^{SB})$ by construction. Hence, $e=H$ is strictly optimal.

We now derive the principal's optimal behavior in the subgame $G_\Gamma(\gamma_\varepsilon)$.

We show that for any effort probability $x\in[0,1]$ and any optimal reporting and participation behavior of the agent (as characterized above), the principal's best response is $\gamma^r=\emptyset$ 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 $x=0$. Moreover, the first argument in the proof of Lemma 3 implies that we can characterize any renegotiated offer that the principal considers optimal by some $\hat{U}_x^r\in(-\infty,+\infty)$, representing the agent's expected utility 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 $e = L$. 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 payoff that the principal himself expects from $\hat{U}_x^r$:

1. For $\hat{U}_1^r < U^0 - \Delta U$, the principal expects payoff $V_H(c_\varepsilon^{SB})$ and for $\hat{U}_0^r < U^0 - \Delta U - 2\varepsilon$, the principal expects payoff $V_L(c_\varepsilon^{SB})$. 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 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 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 (10) 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 (11) 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 (10) 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 (11) 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, (10) and (11) both hold with equality. For any randomization over the agent's decisions, the principal expects a payoff 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 (10) and (11) 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 payoff $\frac{1}{2}V_H^{FI}(\hat{U}_1^r) + \frac{1}{2}V_H^{IC}(U^0 + \Delta U)$ for $x = 1$, and payoff $\frac{1}{2}V_L^{FI}(\hat{U}_0^r) + \frac{1}{2}V_L^{IC}(U^0 + \Delta U)$ 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 $\hat{U}_e^r \in (U^0+\Delta U, \infty)$, 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 $x = 1$, 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:

$$V_H(c_\varepsilon^{SB}) - \frac{1}{2}V_H^{FI}(U^0 - \Delta U) - \frac{1}{2}V_H^{IC}(U^0 + \Delta U) > 0, \tag{12}$$

and

$$V_H(c_\varepsilon^{SB}) - V_H^{FI}(U^0 + \Delta U) > 0. \tag{13}$$

Observe that, if $\varepsilon = 0$, (12) and (13) are strictly satisfied because they coincide with (8) and (9), respectively. Since $V_H(c_\varepsilon^{SB})$ is continuous in $\varepsilon$, there is a $\varepsilon^H > 0$ such that (12) and (13) 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

$$V_L(c_\varepsilon^{SB}) - \frac{1}{2}V_L^{FI}(U^0 - \Delta U - 2\varepsilon) - \frac{1}{2}V_L^{IC}(U^0 + \Delta U) > 0 \tag{14}$$

and

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

Again, since $V_L(c_\varepsilon^{SB})$ is continuous in $\varepsilon$, there is a $\varepsilon^L > 0$ such that (14) and (15) 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 payoff 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 $c_\varepsilon^{SB}$.

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 payoff 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 $x$.

To see why $V_x^* < V_x(c_\varepsilon^{SB})$ extends to intermediate beliefs, observe that under $x \in (0,1)$ the principal could offer a screening menu $\gamma^r \in C$ with $\gamma^r(H) \neq \gamma^r(L)$. However, screening does not improve his payoff. For any menu $\gamma^r$, an agent who chose effort $e$ and accepts renegotiation optimally reports in $\gamma^r$ to obtain $\hat{U}_e^r = \max_{m^r \in E} U_e(\gamma^r(m^r))$. Her optimal reporting and acceptance behavior in $\gamma_\varepsilon$ depends on $(\tau_\varepsilon, \hat{U}_e^r)$, not on the belief $x$. Since we established that, for any sequentially rational reporting and participation of the agent, $V_e(c_\varepsilon^{SB}) > V_e^*(\gamma^r, \lambda(\gamma_\varepsilon))$ for each $e \in E$ separately, the inequality $V_x(c_\varepsilon^{SB}) > V_x^*(\gamma^r, \lambda(\gamma_\varepsilon))$ holds for all $x \in [0,1]$. Thus, the suboptimality of renegotiation extends to intermediate beliefs $x \in (0,1)$.

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 that makes her accept it with positive probability. Therefore, $e = H$ is her only optimal choice as already argued. 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 payoff 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 mutual best responses.

We now turn to the entire game $G_\Gamma$. Note first that,once the principal offers $\gamma_\varepsilon$, 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.[29] Consequently, in any equilibrium of $G_\Gamma$, the principal must obtain at least the payoff $V^{SB}$: any inferior payoff $V' < V^{SB}$ is not sequentially rational since the principal could deviate to some $\gamma_\varepsilon$ and uniquely obtain $V_H(c_\varepsilon^{SB}) \in (V', V^{SB})$. The existence of an appropriate $\gamma_\varepsilon$ 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 $G_\Gamma$ yields the principal a payoff 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

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

contract as $c^{SB} = c^{IC}(U^0)$. This characterization extends to $G_\Gamma$: any mechanism $\gamma \in \Gamma$ achieving principal payoff $V^{SB}$ 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 considering the transfers implemented by $\xi^{0*}$. They are defined by the sequence of decision rules $\tau^{0*} = (\tau_{T'}^{0*})_{T' \geq 1}$. Each function $\tau_{T'}^{0*}$ associates any history of communication between the agent and $\xi^{0*}$ 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 $\Delta U$ is such that:[30]

$$\Delta U > d \quad \text{and} \quad \frac{1}{2} V_H^{FI}\left(U^0 - \Delta U - d\right) + \frac{1}{2} V_H^{FI}(U^0 + \Delta U - d) < V^{SB}. \tag{16}$$

In $\xi^{0*}$, the agent can send in each round $T$ either a status quo report $N$ or an *irreversible* punishment report $R$. If $R$ is *not* reported, then the second-best transfers $c^{SB}$ are implemented. As soon as $R$ is sent, all future reports become payoff-irrelevant and the punishment outcome characterized in (16) is implemented. 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.* At $T = 0$, the agent observes $\xi^{0*}$ and 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 $\xi^{0*}$) or $m_{T'}^{T} \in \{N, R\}$ (if this is $\xi^{T}$ 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.* The principal may attempt to renegotiate the mechanism $\xi^{0*}$ at any round $T \geq 1$, until $T^*$ realizes. A renegotiated mechanism $\xi^T$ offered at round $T$ requires the agent to submit a report $\hat{e}^T \in \{H, L\}$ at $T.iv)$. Further, at any $T' > T$, it

[30]Existence of $\Delta U$ satisfying (16) follows from continuity and the fact that, for large $\Delta U$, the RHS diverges to $-\infty$ while $V^{SB}$ is finite. The condition $\Delta U > d$ is then satisfiable for $\Delta U$ in an appropriate range.

also requires her 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.[31] The mechanism $\xi^T$ specifies the sequence of decision rules $\tau^T = (\tau^T_{T'})_{T' \geq T}$, with $\tau^T_{T'}$ being the rule for round $T' \geq T$. It 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 $\xi^T$ 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 let $\Xi$ be the set of all such $\xi^T$ mechanisms.

We denote $\mathcal{H}^1_P \equiv \xi^{0*}$ the relevant history for the principal at $T = 1$, which only contains the offer $\xi^{0*}$. We then let $\mathcal{H}^T_P \equiv (\xi^{0*}, \xi^1, \rho^1, \ldots, \xi^{T-1}, \rho^{T-1})$ be a principal's history at the end of stage $T$. Thus, a (pure) behavioral strategy for the principal in $G^\eta_\Xi(\xi^{0*})$ associates to each $T \geq 1$ and to each $\mathcal{H}^T_P$ a renegotiated mechanism $\xi^T \in \Xi$.

To start with, denote $\mathcal{P}^{0*}_T$ the set of principal's histories, and $\mathcal{A}^{0*}_T$ the set of agent's histories, such that $\rho^{T'} = n$ for all $T' : 1 \leq T' < T$. At any such history, $\xi^{0*}$ is still in place in round $T$. At any history $\mathcal{H}^T_P \in \mathcal{P}^{0*}_T$ the principal may either offer a mechanism $\xi^T$, or decide not to renegotiate. At any history $(\mathcal{H}^T_A, \xi^T)$, with $\mathcal{H}^T_A \in \mathcal{A}^{0*}_T$, the agent reports $m^{0*}_T \in \{N, R\}$ in $\xi^{0*}$ and privately observes the signal $s^{0*}_T \in \{h, t\}$; then, at any $(\mathcal{H}^T_A, \xi^T, m^{0*}_T, s^{0*}_T)$ she selects $\rho^T \in \{y, n\}$ and, at any $(\mathcal{H}^T_A, \xi^T, m^{0*}_T, s^{0*}_T, y)$, she selects $\hat{e}^T \in \{H, L\}$.

We next construct players' equilibrium strategies, and show that there are no profitable unilateral deviations. The proof is developed in three steps.

*Step 1. Strategies and beliefs.* We first describe the agent's equilibrium behavior in $G^\eta_\Xi(\xi^{0*})$. At $T = 0$ she takes $e = H$ with probability one. Then, we explicitly characterize her reporting and participation behavior only at the histories $(\mathcal{H}^T_A, \xi^T)$ such that $e = H$ and $\mathcal{H}^T_A \in \mathcal{A}^{0*}_T$. Relative to all other histories, we only require that the agent behaves in a sequentially rationally way given the principal's equilibrium behavior. Consider, in particular, any history $(\mathcal{H}^T_A, \xi^T, m^{0*}_T, s^{0*}_T, y)$, in which $(m^{0*}_T, s^{0*}_T) \in \{N, R\} \times \{h, t\}$ is the communication entertained by the agent with the mechanism $\xi^{0*}$. Since the renegotiation offer $\xi^T$ cannot condition on such private communication, and its acceptance effectively replaces any previous mechanism, the agent's continuation payoff corresponding to $(\mathcal{H}^T_A, \xi^T, m^{0*}_T, s^{0*}_T, y)$ is indeed independent of $(m^{0*}_T, s^{0*}_T)$. We denote it $\hat{U}^{0*}_H(\xi^T)$.

To construct the agent's reporting and participation behavior at any $(\mathcal{H}^T_A, \xi^T)$, we distinguish two mutually exclusive situations, according to the relevant round $T$ and the history of the communication $Z^{0*}_{T-1}$ between the agent and the original mechanism $\xi^{0*}$ up to this round.

[31] The restriction to $\xi^T \in \Xi$ is without loss of generality. Upon accepting $\xi^T$, the agent's prior communication $(m^{0*}_{T'}, s^{0*}_{T'})_{T' < T}$ within $\xi^{0*}$ becomes payoff-irrelevant: agents with the same $e$ but different communication histories evaluate any contract identically. Hence, as in Section 3, the principal cannot screen on this information, and restricting the message space of $\xi^T$ to effort reports involves no loss.

*1. $T > 1$ and $R \in Z^{0*}_{T-1}$.* That is, the history of communication within $\xi^{0*}$ contains at least a report $R$. In any such case, we let the agent report $m^{0*}_T = R$ in $\xi^{0*}$. In addition, her participation decision depends on the signal $s^{0*}_T$ received from $\xi^{0*}$. Specifically:

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

*2. Either $T = 1$, or $R \notin Z^{0*}_{T-1}$.* In any such case, the agent's report in $\xi^{0*}$, which we denote $m^{0*}_T$, is determined as follows:

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

Finally, at any $(\mathcal{H}^T_A, \xi^T, m^{0*}_T, s^{0*}_T)$, the agent participation within $\xi^T$ is such that:

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

We next specify the principal's equilibrium behavior in $G^{\eta}_{\Xi}(\xi^{0*})$. We let the principal choose $\xi^T = \emptyset$ at any history $\mathcal{H}^T_P \in \mathcal{P}^{0*}_T$ on the equilibrium path. 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. We assume that, at any on-the-equilibrium-path history $\mathcal{H}^T_P \in \mathcal{P}^{0*}_T$, the principal believes that $e = H$ with probability one, and that $m^{0*}_{T'} = N$ for all $T' < T$, while he assigns probability one-half to each $s^{0*}_{T'} \in \{h, t\}$. Thus, the principal's on-path beliefs are Bayes-consistent given the agent's behavior. We also require that, at any history $\mathcal{H}^T_P \in \mathcal{P}^{0*}_T$ which is *off* the equilibrium path, the principal still believes that $e = H$ with probability one, while holding a degenerate belief on $m^{0*}_{T'} = R$ for each $T' < T$ starting with $T' = 1$. In particular, he believes that $s^R = t$ has been disclosed by $\xi^{0*}$ to her at the initial round.

*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 $m^{0*}_T = N$ at any $T \geq 1$ yields the agent her reservation payoff $U^0$. Suppose, instead, that she takes $e = L$ at $T = 0$. Then, any subsequent reporting strategy yields her again $U^0$. Indeed, reporting $m^{0*}_T = N$ in $\xi^{0*}$ at any $T \geq 1$ yields the second-best transfers $c^{SB}$. By reporting $R$ in any round $T \geq 1$, the agent triggers the punishment lottery yielding her

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

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 $\xi^{0*}$ up to this round.

1. $T > 1$ and $R \in Z_{T-1}^{0*}$. That is, the history of communication within $\xi^{0*}$ contains at least a report $R$. In any such case, given $\tau^{0*}$, any agent's report from round $T$ onwards in $\xi^{0*}$ is payoff-irrelevant, guaranteeing the optimality of our constructed behavior. Concerning participation, rejecting a renegotiated offer $\xi^T$ secures the agent a continuation payoff 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 $R \notin Z_{T-1}^{0*}$. Consider first on-path histories, that is, any agent's history $(\mathcal{H}_A^T, \emptyset)$ such that $\mathcal{H}_A^T \in \mathcal{A}_T^{0*}$ and $e = H$. By reporting $m_T^{0*} = N$ in $\xi^{0*}$ at any $T \geq 1$ the agent obtains the second-best transfers $c^{SB}$, which yields her the reservation payoff $U^0$. By reporting $R$ in any round $T \geq 1$, the agent triggers the punishment above, and gets the expected payoff $U^0 - d < U^0$ regardless of her subsequent communication behavior. Thus, $m_T^{0*} = N$ is the unique optimal report.

Consider next any off-path agent's history $(\mathcal{H}_A^T, \xi^T)$ such that $\mathcal{H}_A^T \in \mathcal{A}_T^{0*}$ and $e = H$. The agent's constructed participation behavior can be straightforwardly verified to be sequentially rational by comparing, for each $(m_T^{0*}, s_T^{0*}) \in \{N, R\} \times \{h, t\}$, her payoff of accepting $\xi^T$ to that of holding $\xi^{0*}$.

$$\max\{U^0, \hat{U}_H^{0*}(\xi^T)\} \geq \frac{1}{2}\max\{U^0 - \Delta U - d, \hat{U}_H^{0*}(\xi^T)\} + \frac{1}{2}\max\{U^0 + \Delta U - d, \hat{U}_H^{0*}(\xi^T)\}. \quad (17)$$

Indeed, the LHS of (17) is the agent's payoff from reporting $N$ and following the constructed participation behavior, and the RHS represents the expected payoff from reporting $R$ and following again the constructed participation behavior. It is then easy to check that the threshold reporting strategy constructed above is consistent with (17).

*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 $\mathcal{H}_P^T \in \mathcal{P}_T^{0*}$ 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 $m_{T'}^{0*} = N$ in $\xi^{0*}$ 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 $\hat{U}_H^{0*}(\xi^T)$.[32]

[32]Since the non-profitability of a deviation, given the agent's equilibrium behavior, does not depend

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 $\xi^{0*}$ after observing $\xi^T$, and to reject it for any $s_T^{0*} \in \{h, t\}$. Hence, any such $\xi^T$ 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 $\xi^{0*}$. This guarantees her the payoff $U^0 - \Delta U - d$ (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 payoff 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 (16). 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, by construction, the principal believes that $m_{T'}^{0*} = R$ has been reported by the agent in $\xi^{0*}$ at any $T' < T$, and that $s^R = t$ has been disclosed by $\xi^{0*}$ to her at $T = 1$. Such a communication guarantees that $\xi^{0*}$ implements a first-best allocation, which involves full insurance, and yields the utility $U^0 + \Delta U - d$ to the agent in any future rounds. Thus, it cannot be profitable for the principal to renegotiate under this belief, and, at any $T'' \geq T$, it is optimal for him to stick to his equilibrium offer $\xi^{T''} = \emptyset$.

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. ■

on the principal's continuation play after offering $\xi^T$, the analysis below guarantees that *all* deviations starting at $\mathcal{H}_T^P$ are unprofitable for the principal, not only the one-shot deviations.

# B 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 $\gamma^{**}$ 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 $\gamma^{**}$ over the Ethereum blockchain using the commit-and-reveal technique.[33] 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 $\sigma$, 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 $\sigma$ to the smart contract by which the smart contract can recover the original message $m$ so that it can make the transfers according to $\tau^{**}$.

We set up the contract such that if the agent does not reveal the seed $\sigma$ 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.

[33] 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 $\gamma^{**}$ with a reveal-and-commit technique based on the keccak-256 hash function in Solidity.

# C Additional Results

This appendix develops several extensions.

## Irrelevance of Random Mechanisms in Fudenberg and Tirole (1990)

We here formalize the claim that random mechanism 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 $\tilde{\gamma} : E \to \Delta(\mathbb{R}^2)$.

**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 payoff 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 payoff attainable by the principal by a renegotiation offer $\tilde{\gamma}^r \in \tilde{C}$ is the value of the program $P(x, \tilde{U}_H, \tilde{U}_L)$:

$$\begin{aligned} V^*(x, \tilde{U}_H, \tilde{U}_L) = \max_{\tilde{\gamma} \in \tilde{C}} \quad & Y(x) - x[p_H \mathbb{E}(\Phi(u_g)|\tilde{c}^r_H) + (1 - p_H)\mathbb{E}(\Phi(u_b)|\tilde{c}^r_H)] \\ & - (1 - x)[p_L \mathbb{E}(\Phi(u_g)|\tilde{c}^r_L) + (1 - p_L)\mathbb{E}(\Phi(u_b)|\tilde{c}^r_L)] \end{aligned} \tag{18}$$

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

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 the 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.[34] 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 $G_{\tilde{C}}(\tilde{\gamma})$ if and only if it is 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 payoffs 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 payoff by offering any of the mechanisms characterized in Fudenberg and Tirole (1990, Proposition 3.4). Thus, the unique equilibrium's payoff 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 $\gamma^*$ 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 $\gamma^*$ 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

$$U^0 > U^\ell \equiv \frac{p_L}{\Delta p} d, \tag{19}$$

[34] 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.[35] 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* (19) *holds, then there is a* $\pi \in (0,1)$ *such that, for all* $e \in E$*:*

$$V^{SB} > \max\left\{V_e^{FI}\left(\frac{U^0 - U^\ell}{\pi} + U^\ell\right),\, (1-\pi)V_e^{FI}(U^\ell) + \pi V_e^{IC}\left(\frac{U^0 - U^\ell}{\pi} + U^\ell\right)\right\}. \quad (20)$$

**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:

$$\lim_{\pi\to 0}\hat{V}_e(\pi) = Y_e - \Phi(U^\ell + D(e)) - \lim_{\pi\to 0}\frac{\Phi\left(\frac{U^0-U^\ell}{\pi} + U^\ell + D(e)\right)}{\frac{1}{\pi}} = -\infty. \quad (21)$$

To see why (21) 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 (19), 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 (21). 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

$$V_e^{IC}(U^0) > \hat{V}_e(\pi) \quad \forall e \in E. \quad (22)$$

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

$$\hat{V}_e(\pi) > \max\left\{V_e^{FI}\left(\frac{U^0 - U^\ell}{\pi} + U^\ell\right), (1-\pi)V_e^{FI}(U^\ell) + \pi V_e^{IC}\left(\frac{U^0 - U^\ell}{\pi} + U^\ell\right)\right\}. \quad (23)$$

[35] If (19) 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 (22) and (23) together yield $V_e^{IC}(U^0) >$ RHS of (20) for all $e \in E$. Since $V^{SB} = V_H^{IC}(U^0) > V_L^{IC}(U^0)$ by the maintained assumption that high effort is optimal in the second-best, it follows that (20) holds for all $e \in E$. ■

The proof of the lemma 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. The result obtains by appropriately designing the transfers' distribution $(\pi, 1-\pi)$.

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 $\gamma^*$ the idea that the message $m = R$ activates a (random) counter-offer, which activates the relevant punishment. By sending $m = R$ in $\gamma^b$ the agent receives a "low" transfer with probability $1-\pi$ and a "high" one with probability $\pi$. 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 $\gamma^b$ implements the second-best allocation.

## Self-Enforced Timing of Communication

We demonstrate that Proposition 1 still holds if the agent can strategically select the timing of her report in the original mechanism, and the associated disclosure.

To achieve this goal, we introduce a new class of self-revealing mechanisms $\Gamma^\mu$, with $\gamma^\mu = \{\mathcal{M}^\mu, \mathcal{S}^\mu, \sigma^\mu, \tau^\mu\} \in \Gamma^\mu$. A mechanism in this class modifies the primitive game by allowing the agent to report at stages $(iii)$ $(iv)$, $(v)$ and $(vi)$.

The space of the agent's reports in $\gamma^\mu$ is hence $\mathcal{M}^\mu = \{N, R, \emptyset\}^4$, which extends $\mathcal{M}$ along two directions. First, four stages of communication are allowed. Second, in each stage, the agent can send the *empty* message $\emptyset$, which represents her choice of not reporting to the mechanism in that stage. Following the same intuition, we let $\mathcal{S}^\mu = \{h, t, \emptyset\}^4$ be the set of signals. For any stage $t \in \{iii, iv, v, vi\}$, we denote by $m_t \in \{N, R, \emptyset\}$ an agent's report to the mechanism $\gamma^\mu$, and by $s_t \in \{h, t, \emptyset\}$ a signal sent by the mechanism to the agent.

We shall construct $\gamma^\mu \in \Gamma^\mu$ to be such that the agent effectively selects the timing of her relevant communication. This guarantees that no external enforcement is required to verify the adherence on a specific communication protocol. In particular, we let $\gamma^\mu$ be such that:

1. The decision rule $\tau^\mu : \mathcal{M}^\mu \times \mathcal{S}^\mu \to \mathbb{R}^2$ is flat over permutations of $(m_t, s_t)_{t=iii}^{vi}$. Furthermore, it forces the agent to report only one non-empty message in the mechanism, by inflicting a very large punishment to her for any $(m_t, s_t)_{t=iii}^{vi}$ such that $|\{m_t = \emptyset\}| \neq 3$.

2. Each stage-$t$ disclosure rule $\sigma_t : \{N, R, \emptyset\}^t \times \{h, t, \emptyset\}^{t-1} \to \Delta(\{h, t, \emptyset\})$ shares the following features:

   - If there is a $s_{t'} \neq \emptyset$, with $t' < t$, then $\sigma_t$ is degenerate on $s_t = \emptyset$. That is, a mechanism $\gamma^\mu$ discloses at most one non-empty signal.
   - If $s_{t'} = \emptyset$ for any $t' < t$, then two cases may occur. First $m_{t'} = \emptyset$ for all $t' \leq t$, in which case $\sigma_t$ is degenerate on $s_t = \emptyset$ ($\gamma^\mu$ does not send any meaningful signal). Second, there exists $m_{t'} \neq \emptyset$ with $t' \leq t$ but $s_{t'} = \emptyset$ for all $t' < t$, in which case $\sigma_t$ extracts the outcome $s_t \in \{h, t\}$ of a fair coin toss. That is, a mechanism $\gamma^\mu$ discloses a coin toss outcome to the agent when she reports a non-empty message.

We denote by $\Gamma^\mu$ the set of all such mechanisms. Thus, any optimal report of the agent in a given $\gamma^\mu \in \Gamma^\mu$ must involve exactly one non-empty message, thereby inducing the disclosure of only one non-empty signal. Let $(m_t, s_t)_{t=iii}^{vi}$ be any array of messages and signals which exhibits this feature, and denote by $(m_j, s_j) \in \{N, R\} \times \{h, t\}$ its only non-empty element. We let $\tau^\mu(m_j, s_j)$ be the corresponding decision implemented by $\tau^\mu$, which, given (1), does not condition on the time index of the non-empty message and signal but only on their content. Thus, a mechanism $\gamma^\mu \in \Gamma^\mu$ is completely identified by a tuple of eight transfers $(\tau^\mu(m_j, s_j))_{(m_j,s_j)\in\{N,R\}\times\{h,t\}} \in \mathbb{R}^8$.

This construction guarantees that a court need *not* verify the exact sequence of the communication taking place in $\gamma^\mu$ to enforce its transfers, but only their effective content. In the same vein, a court does not need to determine whether the signal is sent after an offer is made and before its acceptance. The specific timing of disclosure is ultimately determined by the agent, through the non-empty report she makes in $\gamma^\mu$.

We now consider the overall game $G_{\Gamma^\mu}$ where the principal selects a mechanism in $\Gamma^\mu$ at the ex-ante stage. The following holds.

**Lemma 6** *The game $G_{\Gamma^\mu}$ 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 $(iii)$ if the agent has accepted $\gamma^\mu$:[36]

$(iii)$ The agent sends a private message $m_{iii} \in \{N, R, \emptyset\}$ in $\gamma^\mu$. If $m_{iii} \neq \emptyset$ the agent receives a private random signal $s_{iii} \in \{h, t\}$ distributed as $\left(\frac{1}{2}, \frac{1}{2}\right)$, otherwise she receives the private signal $s_{iii} = \emptyset$ with probability one. After this communication phase, the agent privately chooses $e \in E$.

$(iv)$ If $m_{iii} \neq \emptyset$, the agent sends the private message $m_{iv} = \emptyset$ to $\gamma^\mu$ and receives the private signal $s_{iv} = \emptyset$. If $m_{iii} = \emptyset$, the agent sends the private message $m_{iv} \in \{N, R, \emptyset\}$ to $\gamma^\mu$: then, if $m_{iv} \neq \emptyset$ the agent receives a private random signal $s_{iv} \in \{h, t\}$ distributed as $\left(\frac{1}{2}, \frac{1}{2}\right)$, otherwise she receives the private signal $s_{iv} = \emptyset$. After this communication phase, without observing $e$ nor $(m_{iii}, m_{iv})$, the principal makes a public renegotiation offer $\gamma^r \in C \cup \{\emptyset\}$, where $\emptyset$ represents the principal's decision not to renegotiate.

$(v)$ If $(m_{iii}, m_{iv}) \neq (\emptyset, \emptyset)$, the agent sends the private message $m_v = \emptyset$ to $\gamma^\mu$, and she privately receives $s_v = \emptyset$. If $(m_{iii}, m_{iv}) = (\emptyset, \emptyset)$, the agent sends a private message $m_v \in \{N, R, \emptyset\}$ to $\gamma^\mu$: then, if $m_v \neq \emptyset$ the agent receives a private random signal $s_v \in \{h, t\}$ distributed as $\left(\frac{1}{2}, \frac{1}{2}\right)$, otherwise she receives the private signal $s_v = \emptyset$. After this communication phase, if $\gamma^r \neq \emptyset$, she publicly accepts or rejects $\gamma^r$ by declaring $\rho \in \{y, n\}$. Acceptance implies that $\gamma^\mu$ is replaced by $\gamma^r$.

$(vi)$ If $\gamma^r = \emptyset$ or $\rho = n$, and $(m_{iii}, m_{iv}, m_v) \neq (\emptyset, \emptyset, \emptyset)$, the agent sends the private message $m_{vi} = \emptyset$ to $\gamma^\mu$, she privately receives the private signal $s_{vi} = \emptyset$, the array $(m_j, s_j)$ is publicly revealed and transfers occur according to $\tau^\mu(m_j, s_j)$. If $\gamma^r = \emptyset$ or $\rho = n$, and $(m_{iii}, m_{iv}, m_v) = (\emptyset, \emptyset, \emptyset)$, the agent sends the private message $m_{vi} \in \{N, R\}$ to $\gamma^\mu$, she privately receives a random signal $s_{vi} \in \{h, t\}$ distributed as $\left(\frac{1}{2}, \frac{1}{2}\right)$, the array $(m_j, s_j) = (m_{vi}, s_{vi})$ is publicly revealed and transfers occur according to $\tau^\mu(m_j, s_j)$. If $\rho = y$, the relevant transfers are determined by a report $m^r \in E$ sent by the agent in $\gamma^r$. Nature publicly draws the output realization $g$ or $b$, and final transfers occur.

A behavioral strategy of the principal in the subgame is a distribution over the set of the renegotiated offers $C$. Since communication is private, the principal *cannot* strategically nor contractually condition his offer on the agent's report's timing, nor on its content.[37] A behavioral strategy of the agent specifies a distribution over

[36] To streamline exposition, we incorporate in the description of $G_{\Gamma^\mu}$ the (optimal) agent's behavior of sending only one non-empty message in $\gamma^\mu$.

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

$m_{iii} \in \{N, R, \emptyset\}$ at the initial node and an effort probability $x \in [0,1]$ at any history $(m_{iii}, s)$. If $m_{iii} \neq \emptyset$, it features a distribution over participation decisions $\rho \in \{y, n\}$ at each history $(m_{iii}, s_{iii}, e, \emptyset, \emptyset, \gamma^r, \emptyset, \emptyset)$ and a distribution over $m^r \in E$ at the continuation where $\rho = y$. If $m_{iii} = \emptyset$, it features a distribution over $m_{iv} \in \{N, R, \emptyset\}$ at each history $(\emptyset, \emptyset, e)$. Then, one must distinguish two cases. If $m_{iv} \neq \emptyset$, the agent's behavior features a distribution over participation decisions $\rho \in \{y, n\}$ at each history $(\emptyset, \emptyset, e, m_{iv}, s_{iv}, \gamma^r, \emptyset, \emptyset)$ and a distribution over $m^r \in E$ at the continuation where $\rho = y$. If, instead, $m_{iv} = \emptyset$, the agent's behavior features a distribution over $m_v \in \{N, R, \emptyset\}$ at each history $(\emptyset, \emptyset, e, \emptyset, \emptyset, \gamma^r)$ and a distribution over participation decisions $\rho \in \{y, n\}$ at each history $(\emptyset, \emptyset, e, \emptyset, \emptyset, \gamma^r, m_v, s_v)$, followed by a distribution over $m^r \in E$ if $\rho = y$ and a distribution over $m_{vi} \in \{N, R, \emptyset\}$ if $\rho = n$ and $(m_v, s_v) = (\emptyset, \emptyset)$.

We now show that $(H, c^{SB})$ is indeed a pure-strategy equilibrium allocation of $G_{\Gamma^\mu}$.

Consider in fact the mechanism $\gamma^{\mu*} \in \Gamma^\mu$ that executes the same transfers as $\gamma^*$: i.e., $\tau^{\mu*}(m_j, s_j) = \tau^*(m_j, s_j)$ for all $(m_j, s_j) \in \{N, R\} \times \{h, t\}$. This mechanism implements the second-best allocation $(H, c^{SB})$ in the subgame $G_{\Gamma^\mu}(\gamma^{\mu*})$.

To get the result, we construct a continuation equilibrium of $G_\Gamma(\gamma^{\mu*})$ where on the equilibrium 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 the equilibrium path, if the principal offers $\gamma^r \neq \emptyset$, the agent always selects $m_v \neq \emptyset$ and takes her participation decisions $\rho$ following the rules established in Proposition 1. The arguments developed in the proof of Proposition 1 guarantee that these strategies constitute an equilibrium. In particular, since $\gamma^r = \emptyset$ at equilibrium, the option to report $m_{iii}$ or $m_{iv}$ early is strategically irrelevant for the agent.

Thus, given this continuation equilibrium, the principal obtains $V^{SB}$ by offering $\gamma^{\mu*}$ at the ex ante stage. Since the principal cannot obtain more with any other offer, this is an equilibrium payoff that she obtains in $G_{\Gamma^\mu}$.

Following the logic developed in the proof of Proposition 2, we now argue that this is the only principal's continuation payoff at $G_{\Gamma^\mu}(\gamma^{\mu*})$ compatible with a pure strategy equilibrium of $G_{\Gamma^\mu}$. To establish this, we construct a perturbed version $\gamma^\mu_\varepsilon$ of $\gamma^{\mu*}$ to which the principal can deviate and obtain a unique continuation payoff arbitrarily close to $V^{SB}$, under the restriction to pure strategies. Indeed, $\gamma^\mu_\varepsilon$ induces the same transfers as the tie-breaking mechanism $\gamma_\varepsilon$ in the proof of Proposition 2: that is, $\tau^\mu_\varepsilon(m_j, s_j) = \tau_\varepsilon(m_j, s_j)$ for all $(m_j, s_j) \in \{N, R\} \times \{h, t\}$. Observe in particular that $\tau^\mu_\varepsilon(m_j, h) = c^{IC}(U^0 - \Delta U - \kappa\varepsilon)$ with $\kappa > 2$. We introduce the additional requirement in the construction of $\gamma^\mu_\varepsilon$ that $\kappa$ is

large enough to verify for both $e \in E$:[38]

$$\frac{1}{2}V_e^{FI}(U^0 - \Delta U - \kappa\varepsilon) + \frac{1}{2}V_e^{IC}(U^0 + \Delta U) \geq V_e^{FI}(U^0 + \Delta U). \tag{24}$$

We now argue that every equilibrium in pure strategies of $G_{\Gamma^\mu}(\gamma_\varepsilon^\mu)$ yields exactly $V_H(c_\varepsilon^{SB})$ to the principal. The proof is in five steps.

—*Step 1.* We show that there is no pure-strategy equilibrium of $G_{\Gamma^\mu}(\gamma_\varepsilon^\mu)$ where either $m_{iii} \neq \emptyset$, or $m_{iii} = \emptyset$ but $m_{iv} \neq \emptyset$ on the equilibrium path.

Suppose first that either $m_{iii} = N$, or $m_{iii} = \emptyset$ but $m_{iv} = N$ on path. For Bayes-consistency, $(m_{iii} = N, m_{iv} = \emptyset)$ or $(m_{iii} = \emptyset, m_{iv} = N)$ must also be the principal's equilibrium belief on the agent's on-path stage-$(iii)$ and stage-$(iv)$ reports. Under any of such beliefs, the principal's optimal renegotiation offer must be degenerate on $c_H^{FI}(U^0)$ or $c_L^{FI}(U^0 - \varepsilon)$, according to the agent's equilibrium effort decision $e \in E$. These transfers, in fact, yield the agent her reservation payoff $U^0$ (if $e = H$) or $U^0 - \varepsilon$ (if $e = L$) from $\tau_\varepsilon^\mu(N, s) = c_\varepsilon^{SB}$, while yielding the full-insurance payoff $V_H^{FI}(U^0)$ or $V_L^{FI}(U^0 - \varepsilon)$ to the principal. Then, by the principal's sequential rationality, an offer as such must be featured in any equilibrium as described. However, the agent, anticipating this offer, could deviate by sending, for example, $m_{iii} = \emptyset$ followed by the same effort $e \in E$ featured in her original behavior and $m_{iv} = R$. This would yield to her the expected payoff $U^0 + \frac{\Delta U}{2} > U^0$ if $e = H$, or, $U^0 + \frac{\Delta U - \varepsilon}{2} > U^0 - \varepsilon$ if $e = L$, which constitutes a contradiction.

Suppose instead that $m_{iii} = R$ or $m_{iii} = \emptyset$ but $m_{iv} = R$ on the equilibrium path. Again, the principal must hold the degenerate equilibrium belief that the agent's on-path stage-$(iii)$ and stage-$(iv)$ reports are $(m_{iii} = R, m_{iv} = \emptyset)$ or $(m_{iii} = \emptyset, m_{iv} = R)$ in the two cases. In this scenario, the principal's optimal renegotiation has to be either $c_e^{FI}(U^0 - \Delta U - \kappa\varepsilon)$, or $c_e^{FI}(U^0 + \Delta U)$, where $e \in E$ is the agent's equilibrium effort level. Indeed, the proof of Proposition 2 establishes that the principal's optimal offer must be a full-insurance contract, as long as he believes that the agent's effort behavior is degenerate. Also, the utility left to the agent by such an offer must be either $(U^0 + \Delta U)$, the lowest utility level that she may accept for both $s_j \in \{h, t\}$, or $U^0 - \Delta U - \kappa\varepsilon$, the lowest utility she may accept when $s_j = h$. Comparing the principal's payoffs under the two deviations gives the terms of (24). Hence, the inequality in (24) guarantees that inducing $c_e^{FI}(U^0 - \Delta U - \kappa\varepsilon)$ is the best option for the renegotiating principal, for each $e \in E$.

Then, any pure-strategy equilibrium as described must feature this offer of the principal. Since, for any equilibrium $e \in E$, this offer yields no more than $U_e(\tau_\varepsilon^\mu(R, h))$ to the agent, her payoff at any equilibrium as such would be $U^0 - \frac{\kappa}{2}$, which she obtains under

[38] Existence of such a $\kappa$ obtains by observing that $\lim_{\kappa \to \infty} V_e^{FI}(U^0 - \Delta U - \kappa\varepsilon) = \infty$, and that all other terms in (24) are finite for every $\kappa > 2$.

any optimal participation behavior, and for each level of effort, after sending $m = R$. Hence, the agent can profitably deviate: for instance, she can send $m_{iii} = N$, obtaining $U^0$ in the continuation play by selecting the same $e \in E$ as in the original behavior, and rejecting $\gamma^r$ for all $s_{iii} \in \{h, t\}$. This implies that the reports $(m_{iii} = R, m_{iv} = \emptyset)$, as well as $m_{iii} = \emptyset$ followed by $m_{iv} = R$ on the equilibrium path, are incompatible with a pure-strategy equilibrium.[39]

—*Step 2.* This step shows that the agent's option to delay her report until after the participation decision (i.e., choosing $m_v = \emptyset$) produces no strategic effects.

To show this, we argue that all pure-strategy equilibrium allocations are also supported in a pure-strategy equilibrium where the agent sends $m_v \neq \emptyset$ after every offer $\gamma^r$ on or off the equilibrium path. In particular, any pure-strategy equilibrium of $G_{\Gamma^\mu}(\gamma_\varepsilon^\mu)$ where $m_v = \emptyset$ is taken at some history $(\emptyset, \emptyset, e, \emptyset, \emptyset, \gamma^r)$ has a corresponding equilibrium where $m_v = N$ at every such history, supporting the same equilibrium allocation.

To see this note first that, in $\gamma_\varepsilon^\mu$, the agent obtains $U^0$ (if $e = H$) or $U^0 - \varepsilon$ (if $e = L$) from $m_j = N$. Instead, from $m_j = R$, she obtains in expectation $U^0 - \frac{\kappa}{2}\varepsilon$ regardless of her effort decision. Since $\varepsilon > 0$ and $\kappa > 2$, the unique optimal report of the agent at every history $(e, \emptyset, \emptyset, \gamma^r, \emptyset, \emptyset, n)$ is thus $m_{vi} = N$.

Furthermore, observe that, by construction of $\gamma_\varepsilon^\mu$, we have $\tau_\varepsilon^\mu(N, s) = c_\varepsilon^{SB}$ for all $s_j \in \{h, t\}$. Hence, the realization of $s_j$ is payoff-irrelevant when $m_j = N$. Starting from an equilibrium behavior where $m_v = \emptyset$ in some history off the equilibrium path, the agent can therefore 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 taken at $(e, \emptyset, \emptyset, \gamma^r, \emptyset, \emptyset)$ in the original equilibrium. Since the rejection payoff $U_e(c_\varepsilon^{SB})$ and acceptance payoff $\hat{U}_e^r$ are both independent of the signal, this participation decision remains optimal. Hence, the newly constructed strategy is featured in an equilibrium supporting the same allocation.

—*Step 3.* We show that any pure-strategy equilibrium of $G_{\Gamma^\mu}(\gamma_\varepsilon^\mu)$ involves either $\gamma^r = \emptyset$ or any alternative offer that the agent rejects. Observe that, as established in Step 1, the principal believes in any equilibrium that $m_{iii} = m_{iv} = \emptyset$ with probability one. Also, fix without loss of generality an equilibrium as constructed in Step 2: in any equilibrium as

[39] In Step 1 we cover explicitly the case that the agent, after reporting $m_{iii} \neq \emptyset$, selects the same effort at each history $(m_{iii}, s_{iii})$, thereby *not* exploiting the random signal $s_{iii} \in \{h, t\}$ to introduce stochasticity in her choice of effort. Although we do not include the full argument for parsimony, the reasoning in Step 1 extends to such case. The argument extends immediately to the case that $m_{iii} = N$ since this report renders the signal $s_{iii}$ payoff-irrelevant in $\gamma_\varepsilon^\mu$. In case the agent reports $m_{iii} = R$ and selects different effort levels $\hat{e}(h), \hat{e}(t)$ as $s_{iii} \in \{h, t\}$, for $\Delta U$ large enough, a condition analogue to (24) guarantees that $c_{\hat{e}(h)}^{FI}(U^0 - \Delta U - \kappa\varepsilon)$ is still the principal's optimal offer for any combination of $(\hat{e}(h), \hat{e}(t)) \in E^2$. Once established that the principal makes this offer in any equilibrium as such, the remainder of the argument in Step 1 follows directly.

such, the agent always reports in $\gamma_\varepsilon^\mu$ only *after* observing an offer $\gamma^r$ indexed by a utility level $\hat{U}_e^r$, but *before* taking her participation decision. Thus, the principal anticipates that her optimal reporting and participation behavior will coincide with that characterized in the proof of Proposition 2 where the agent's reports exhibit this timing by construction of $\gamma_\varepsilon$. But then, as shown there, any offer accepted with positive probability by the agent yields to the principal a payoff strictly below $V_e(c_\varepsilon^{SB})$ for all $e \in E$. This guarantees that only $\gamma^r = \emptyset$ or any offer not accepted by agent are compatible with the principal's sequential rationality.

—*Step 4.* Given Step 3, the agent anticipates at her initial decision node that the principal will offer $\gamma^r = \emptyset$. Consequently, she also anticipates her unique optimal report to be $m_j = N$, which leads with probability one to the execution of the strictly incentive-compatible transfers $c_\varepsilon^{SB}$. Therefore, she is strictly better off choosing $e = H$.

Taken together, Steps 1–4 imply that every pure-strategy equilibrium of $G_{\Gamma^\mu}(\gamma_\varepsilon^\mu)$ induces the allocation $(H, c_\varepsilon^{SB})$, yielding to 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 mutual best responses. Since $\lim_{\varepsilon \to 0} V_H(c_\varepsilon^{SB}) = V^{SB}$, the logic developed in the proof of Proposition 2 guarantees that $V^{SB}$ is the unique equilibrium payoff for the principal in $G_{\Gamma^\mu}$. Thus, following again Proposition 2, $(H, c^{SB})$ is the unique equilibrium allocation. ■

Importantly, Lemma 6 shows that, when the agent is delegated the enforcement of the communication protocol, the principal can neither push the agent to accept an offer without reporting in $\gamma^*$ (i.e., induce $m_{iii} = m_{iv} = m_v = \emptyset$), nor strategically wait until a report is sent by the agent before making an offer (i.e., induce $m_{iii} \neq \emptyset$ or $m_{iv} \neq \emptyset$). While the proof of Lemma 6 focuses on pure strategies for parsimony, the argument naturally extends to mixed strategies. In particular, following again the proof of Proposition 2, any renegotiation that is unprofitable when the agent plays a pure strategy, is *a fortiori* unprofitable when effort is mixed. Thus, $\gamma^r = \emptyset$ is still optimal for the principal in this richer scenario.

## 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 $\gamma^{**}$ 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 $\gamma^{**}$ 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 $\gamma^r = \emptyset$, where $\mathcal{M}^r = E$ and $\tau^r : \mathcal{M}^r \times \mathcal{S} \to \Delta C$, allowing 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. After this communication phase, if $\gamma^r \neq \emptyset$, the agent publicly accepts or rejects $\gamma^r$ by declaring $\rho \in \{y, n\}$. Acceptance implies that $\gamma^{**}$ is replaced by $\gamma^r$.

(*vi*) The message $m$ is publicly revealed if and only if $\gamma^{**}$ executes (i.e. either $\gamma^r = \emptyset$ or $\rho = n$) in which case transfers are determined by $\tau^{**}(m, s)$. If $\rho = y$, transfers are determined by a report $m^r \in \mathcal{M}^r$ sent by the agent in $\gamma^r$ and the previous realization of $s$. Nature publicly draws the output realization $g$ or $b$, and conditional transfers are executed.

A pure behavior for the principal in $G_{Pub}(\gamma^{**})$ is a signal-contingent renegotiated offer $\gamma^r$.[40] An agent's behavioral strategy $\lambda$ consists of a randomization $(1-x, x)$ over $e \in E$ at her initial history, a randomization over messages in $\mathcal{M}^{**}$ at each history $(e, \gamma^r)$, a randomization over participation decisions $\rho \in \{y, n\}$ at each history $(e, \gamma^r, m, s)$ 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 $\gamma^{**}$, let $\hat{m}_e^r(s) = \arg\max_{m^r \in \mathcal{M}^r} U_e(\tau^r(\hat{m}_e^r(s), s))$ be an optimal message that the agent may send after accepting $\gamma^r$, 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 payoff $\hat{U}_e^r(s) \equiv U_e(\tau^r(\hat{m}_e^r(s), s))$.

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 $\gamma^r$, 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 $\hat{U}_e^r(s) \geq U^0 - \Delta U$;

[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.

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

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

$$\begin{aligned}\frac{1}{2}\max\{U^0, \hat{U}_e^r(h)\} + \frac{1}{2}\max\{U^0, \hat{U}_e^r(t)\} \geq \\ \max\Big\{\frac{1}{2}\max\{U^0 - \Delta U, \hat{U}_e^r(h)\} + \frac{1}{2}\max\{U^0 + \Delta U, \hat{U}_e^r(t)\}, \\ \frac{1}{2}\max\{U^0 + \Delta U, \hat{U}_e^r(h)\} + \frac{1}{2}\max\{U^0 - \Delta U, \hat{U}_e^r(t)\}\Big\},\end{aligned} \tag{25}$$

the agent sends $m = N$ in $\gamma^{**}$. Observe that the LHS of (25) corresponds to the agent's expected payoff of reporting $m = N$ in $\gamma^{**}$, followed by her signal-contingent participation decisions. The RHS of (25) characterizes the payoff corresponding to the best alternative report.

(*ii*) For any $e \in E$, and for any $\gamma^r \neq \{\emptyset\}$ such that (25) is *not* satisfied, the agent sends $m = R_1$ in $\gamma^{**}$ whenever

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

(*iii*) For any $e \in E$, and for any $\gamma^r \neq \{\emptyset\}$ such that (25) and (26) are *not* satisfied, the agent sends $m = R_2$ in $\gamma^{**}$

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 sets, 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 payoff of accepting $\gamma^r$ versus retaining $\gamma^{**}$; the reporting behavior is also described by comparing the agent's continuation payoff 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 payoff 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 $\gamma^{**}$.

Observe first that, for any $\gamma^r$ such that the agent reports $m = R_1$ in $\gamma^{**}$, the principal's payoff 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 payoff providing full insurance to the agent conditional on each realized signal. In this case, Lemma 1 guarantees that $V^{SB} > V^R$. Thus, the principal prefers not to renegotiate than renegotiating an offer which induces the report $m = R_1$.

A symmetric argument applies to any $\gamma^r$ such that the agent reports $m = R_2$ in $\gamma^{**}$. In any such case, one can also check that the principal cannot achieve a payoff greater than $V^R$.

Thus, any profitable renegotiation $\gamma^r$ must be such that the agent's equilibrium strategy prescribes to report $m = N$ in $\gamma^{**}$. That is, given (25), and since $e = H$, one should have:

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

We now argue that (27) is satisfied only if one of the following two conditions is met:

$$\hat{U}_H^r(s) < U^0, \ \forall s \in \mathcal{S}^{**} \text{ or } \hat{U}_H^r(s) \geq U^0 + \Delta U, \ \forall s \in \mathcal{S}^{**}. \tag{28}$$

To see this, suppose that (28) does not hold, which leads to consider three cases.

(*i*) If $\hat{U}_H^r(t) < U^0$ and $\hat{U}_H^r(h) \geq U^0$, then the LHS of (27) is $\frac{1}{2}\hat{U}_H^r(h) + \frac{1}{2}U^0$ 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 (27).

(*ii*) If $U^0 \leq \hat{U}_H^r(t) < U^0 + \Delta U$, then the LHS of (27) 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 (27). 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 (27) 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 (27) 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 (27)

Thus, following a renegotiation $\gamma^r$, $\lambda(\gamma^{**})$ prescribes $m = N$ and only if (28) holds. Two cases must then be considered:

(i) If $\hat{U}_H^r(s) < U^0 \ \forall s \in \mathcal{S}^{**}$, then (27) rewrites $U^0 \geq U^0$, and is thus satisifed with equality. Thus, $\lambda(\gamma^{**})$ prescribes to report $m = N$ in $\gamma^{**}$ 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 (27) 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 $\gamma^{**}$. In addition, for any such $\gamma^r$, the agent is guaranteed the payoff $U^0 + \Delta U$ in the continuation play, which implies that the principal's payoff 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. ■

# 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.

**_ , 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 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.

**Myerson, Roger B.**, "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.

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

**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.

**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.