Trembling mechanisms

João Correia-da-Silva[*†]

2017-02-16

Abstract. A mechanism design problem where the outside option of privately informed agents is a default game is considered. It is shown that participation constraints can be relaxed by designing the (otherwise off the equilibrium path) beliefs following rejection using a mediation device that, with a small probability, manipulates the acceptance messages received from the agents to originate spurious rejections that are correlated with their announced types. Participation constraints can be further relaxed if the mediation device is also used as a joint randomization device designed to punish a rejector. Applications to collusion under private information are provided.

Keywords: Learning from disagreement, Beliefs off-path, Mediated mechanism design.

JEL Classification Numbers: D82.

1 Introduction

Suppose that you want to establish a binding agreement with a privately informed rival, in an environment in which no commitment is possible regarding subsequent behavior in case of disagreement. Suppose also that you can use a mediator who is able to mimic rejection of the agreement by your rival, so that you never distinguish a genuine rejection from a spurious one. The purpose of this paper is to show how, and to what extent, such a mediator generates additional incentives for your rival to accept the agreement by credibly threatening to induce in you the disagreement beliefs and subsequent behavior that are the most adverse for him.

In the kind of economic relationship that is investigated, one of the parties, the principal, has the power to design the structure of the interaction, while the other party, the agent, can only accept to participate under the rules defined by the principal or reject to participate and

I am grateful to Bruno Jullien, Daniel Garrett, Renato Gomes and, especially, Takuro Yamashita for very useful conversations and suggestions. I also thank Ariel Rubinstein, Dirk Bergemann, Françoise Forges, Jacques Crémer, Johannes Schneider, Mikhail Drugov, Patrick Rey, and audiences at Toulouse, Porto, PEJ 2016, ESEM 2016 and EARIE 2016. The usual disclaimer applies. This research has been funded by the European Commission through the Marie Skłodowska Curie Fellowship H2020-MSCA-IF-2014-657283; and by FEDER (through COMPETE) and FCT in the framework of projects PTDC/IIM-ECO/5294/2012 and PEst-OE/EGE/UI4105/2014.

[*]Toulouse School of Economics. 21 Allée de Brienne, 31000 Toulouse, France. E-mail: joao.correia@tse-fr.eu.

[†]CEF.UP and Faculdade de Economia, Universidade do Porto.

take an outside option. In many environments where the agent has private information about his type, the form of interaction that is optimal for the principal consists in making a "take-it-or-leave-it" proposal of a menu of contracts that is incentive compatible and individually rational: each contract in the menu is designed for one possible type of agent, in a way that makes it optimal for the agent to pick the contract designed for his actual type rather than choose another contract or reject the proposal.

We will focus on situations in which the outside option of the agent involves subsequent strategic interaction with the principal. For example: when a firm proposes a collusive agreement to a rival that has private information about its own efficiency or about demand; when a bidder proposes a pre-auction arrangement to a rival that has private information about the value of the good or procurement contract that will be auctioned; when a contender offers a bribe to an opponent that has private information about his skills; when a plaintiff proposes a settlement to a defendant that has private information about his degree of negligence, or when a defendant proposes a settlement to a plaintiff that has private information about the magnitude of the damages.

In such settings, the value of the outside option for the agent depends on the information inferred by the principal from the agent's rejection of the menu of contracts. This information may harm or benefit the agent, depending on whether it encourages or discourages aggressive behavior by the principal. For example, if rejection of a collusive agreement conveys the information that the agent is very efficient, then, in subsequent interaction where firms simultaneously choose quantities to produce, the principal will expect a high output from the agent and will thus choose a low output. This increases the value of the outside option for the agent, which means that the information content of rejection tightens the participation constraints.

Unable to commit to behave aggressively in the outside game, the principal would like to at least be able to commit to interpret a rejection in the worst possible way for the agent, that is, in the way that induces the principal's subsequent behavior to be as aggressive as possible.[1] This would decrease the value of the outside option for the agent, and thus relax the participation constraints. However, it is hard to conceive how can the principal commit to some future beliefs, or how can the principal control the information that is released from the rejection of a proposal. Perhaps surprisingly, this kind of commitment or control is shown to be viable if the principal can use a mediator who is able to mimic a rejection by the agent.

Before bringing the mediator to the picture, let us look more closely at the game that results from a given proposal: knowing his type, the agent picks a contract from the menu

[1] It is possible that the worst interpretation for the agent depends on his type. In that case, since the principal cannot condition her interpretation on the (unobserved) type of the agent, the *ex ante* worst interpretation optimally trades-off punishing the various types of the agent taking into account the shadow values of the respective participation constraints.

or rejects to participate; in case of rejection, principal and agent simultaneously choose an action.[2] In a sequential equilibrium in which the proposal is always accepted, we can interpret the strategy of the principal as a punishment threat whose credibility is sustained by the beliefs following rejection that are prescribed by the equilibrium. The fact that rejection never occurs on the equilibrium path implies that, if we adopt sequential equilibrium as the solution concept, any beliefs following rejection are admissible as long as they are common knowledge. This is good for the principal in the sense that there always exists a sequential equilibrium in which beliefs following disagreement are those that minimize the agent's payoff in the outside game. However, an appropriate refinement may rule out particularly incredible beliefs and the corresponding punishment strategies.[3]

For example, suppose that the outside game is a quantity-setting duopoly in which the agent has private information about his efficiency, and consider a sequential equilibrium in which the principal believes, following an off-the-equilibrium-path rejection of a menu of collusive agreements, that the agent has the lowest possible efficiency. Unfortunately for the principal, those beliefs will typically be untenable in the sense of not being neologism-proof. If there is any type of agent, other than the least efficient type, that is indifferent between accepting or rejecting the proposal, this type of agent will belong to a credible set of rejectors (set of types that are better off rejecting if the principal believes that this is the set of types that reject). Therefore, it would be unreasonable to believe that the agent has the lowest possible efficiency level. Whenever, as in this example, appropriate refinements rule out the beliefs following rejection that are the worst for the agent, the corresponding punishment loses its credibility, rendering the participation constraints tighter.

Now assume that the principal can offer a menu of contracts to the agent through an incentiveless mediator (a consultant, lawyer, machine or computer software) that is able to mimic a rejection by the agent. After the agent chooses a contract, with a commonly known probability that depends on the chosen contract, the mediator sends to the principal a message of rejection and the deal is off. Exactly the same message is sent if the agent rejects the menu of contracts, thus the principal is not able to distinguish a genuine rejection from a spurious rejection. As we will see, the agent would always like the principal to believe that the rejection was genuine. It is crucial for the principal that the agent cannot prove that a rejection was genuine.

With recourse to such a mediator, the principal can design the (otherwise off-path) beliefs that ensue if the agent rejects the menu of contracts. The only restriction is that beliefs must be common knowledge. By turning disagreement into an event that occurs on the equilibrium

[2]The consideration of an outside game with multiple stages would obscure the analysis.

[3]For example, the intuitive criterion (Cho and Kreps, 1987), some version of divinity (Banks and Sobel, 1987; Cho and Sobel, 1990), neologism-proofness (Grossman and Perry, 1986; Farrell, 1993), stability (Kohlberg and Mertens, 1986), forward induction (van Damme, 1989), or undefeatedness (Mailath et al., 1993).

path, the mediator allows the principal to bypass all refinements based on restrictions of beliefs formed off-path, which could constrain her by ruling out convenient, but implausible, beliefs following disagreement.

If the outside game is a quantity-setting oligopoly, the principal would like to commit to believe that an agent that rejects her proposal has the lowest possible efficiency. Therefore, she should instruct the mediator to mimic rejection (with a small probability) if and only if the agent picks the contract designed for the agent with the lowest possible efficiency. As a result, rejection – by the mediator, through misrepresentation of the agent's decision – occurs in equilibrium, and thus beliefs following disagreement are no longer formed off-path. A genuine rejection is misinterpreted by the principal as a spurious rejection produced by the mediator through the manipulation of the choice of an agent with the lowest possible efficiency, as this is the only kind of rejection that occurs on-path. Hence, the principal will believe that the agent has the lowest possible efficiency.

The introduction of a mediator perturbs the incentive compatibility constraints whenever the choice of contract influences the probability of spurious rejection.[4] Fortunately, this perturbation can be made arbitrarily small because any strictly positive probability of spurious rejection, no matter how small, is infinitely greater than the probability of a genuine rejection (which is zero in equilibrium). Hence, any menu of contracts that is strictly incentive compatible without manipulation remains strictly incentive compatible after the introduction of a mediator that generates the desired beliefs. Strictness of incentive compatibility is not a severe requirement in the sense that a menu of contracts that is incentive compatible can be approximated by one that is strictly incentive compatible (as long as there exists a strictly incentive compatible menu of contracts).

The participation constraints of the agent can be further relaxed if the mediator is able to send private signals to the principal and the agent accompanying the (spurious or genuine) announcement of a rejection. The ability to use the mediator as a randomization device, when playing a mixed strategy in the outside game, may benefit the principal even if the agent has no private information. The principal should instruct the mediator to act as a trustworthy randomizer when the rejection is spurious but not when the rejection is genuine. After a genuine rejection, the mediator should recommend to the principal the worst action for the agent among those in the support of the mixed strategy that is played after a spurious rejection. The principal obeys the recommendation because she believes that the rejection has been spurious and that the mediator has drawn the recommendation according to the mixed

[4]The exception is when the probability of spurious rejection is independent of the agent's choice of contract. In that case, disagreement reveals no information about the agent and thus players have *passive beliefs*.

strategy equilibrium distribution.[5]

For example, suppose that the outside game is a complete information all-pay auction with a cap on bids in which only principal and agent participate.[6] Under the assumption that the valuation for winning of the agent is greater than that of the principal, and that the cap on bids is greater than half of the principal's valuation, the unique equilibrium of this game is in mixed strategies with mass points at the cap, such that the agent wins the auction with a probability that is greater than 50% and makes an expected payment that is lower than the cap. In such an environment, the principal can commit to the harshest possible punishment by instructing the mediator to draw a recommendation in accordance with the mixed strategy equilibrium if the outside game is reached after a spurious rejection, but recommend a bid equal to the cap with 100% probability whenever the rejection is genuine (the principal obeys the recommendation because she presumes that the rejection has been spurious, and that the mixed strategy equilibrium of the outside game is being played). Facing such a punishment, the agent may either bid the cap and win with 50% probability or bid zero. Such an extreme punishment would not be credible without the mediator.

More generally, the ability of the mediator to send private signals to principal and agent together with the announcement of a rejection allow the principal to use the mediator as an informed joint randomization device that is designed to punish a genuine rejector (off the equilibrium path). In the presence of such a mediator, the outside game that follows a spurious rejection becomes an extended outside game in which the principal and the agent first receive private signals from the mediator, who has become informed about the type of the agent through the agent's choice of contract, and then choose a possibly mixed action.[7] The private signals sent by the mediator after a spurious rejection can be assumed, w.l.o.g., to constitute a stochastic profile of recommendations of actions to be chosen in the outside game that is incentive compatible, i.e., that principal and agent have interest in obeying. Such a stochastic action profile can be designated as an Informed Mediator Bayesian Correlated Equilibrium (BCE^{I}) of the outside game, which is, by definition, a Bayesian Nash Equilibrium (BNE) of the extended outside game.[8]

Keep in mind that, after a spurious rejection, the set of incentive compatible distributions over profiles of recommendations is the set of BCE^{I} of the outside game for given beliefs

[5] A similar mechanism has been used by Kandori (1991) and Mailath et al. (2002) in the context of repeated games with private monitoring and imperfect public monitoring, respectively.

[6] This game was analyzed by Che and Gale (1998b) in their study on political lobbying. See also Che and Gale (1998a), Gavious et al. (2002), Szech (2015) and Olszewski and Siegel (2016).

[7] Such an extended game was the subject of recent research by Bergemann and Morris (2013, 2016).

[8] Bergemann and Morris (2013, 2016) designated this notion of correlated equilibrium where an omniscient mediator is employed as *Bayes Correlated Equilibrium*. In the taxonomy of Forges (1993), where the mediator knows everything that agents know but not more than that, this notion is designated as *Bayesian solution*.

(which can be designed by the principal). After a genuine rejection (which does not occur in equilibrium), the principal presumes that the rejection has been spurious and obeys any recommendation from the mediator that she would obey after a spurious rejection. This means that any action in the support of a BCE^{I} of the outside game for some beliefs is a credible punishment threat. Furthermore, since the recommendations and the induced beliefs can be concealed from the agent, any mixture over such actions is also a credible punishment threat.

When the principal interacts with multiple agents, she can use the mediator to design the disagreement beliefs of the principal and the acceptors about the types of all agents. For example, suppose that the disagreement beliefs that are worst for a rejector consist in each of the other players believing that everyone else has the lowest possible efficiency. To generate such beliefs (approximately), the principal should instruct the mediator to generate a spurious rejection with a small probability, $\epsilon_0 > 0$, if all agents choose the contract designed for the agent with the lowest possible efficiency (in short, announce the lowest efficiency); and with an even smaller probability, say $\epsilon_1 > 0$, if all agents except one announce the lowest efficiency. Following a rejection, an acceptor that did not announce the lowest efficiency will believe that all the other agents have the lowest efficiency. An acceptor that has announced the lowest efficiency will believe that, with a probability that converges to one as $\epsilon_1 \to 0$, all other agents have the lowest efficiency.

This means that, when there is more than one privately informed agent, the ability of the mediator to mimic a rejection allows the principal to generate disagreement beliefs that violate the "*no signaling what you don't know*" condition (Fudenberg and Tirole, 1991). Such beliefs would not be consistent in the absence of a mediator.[9]

With multiple agents, as in the single-agent case, the ability of the mediator to send a profile of private signals with a distribution that depends on the input messages received from the agents is very powerful. It allows the mediator to induce principal and acceptors to obey any distribution over action profiles such that any action that is recommended to a player is also recommended to the same player in some BCE^{I} (each action can belong to the support of a different BCE^{I} and even BCE^{I} with different underlying disagreement beliefs).

Trembling mechanisms do not allow the principal to design the beliefs of the rejector. Moreover, since the anticipated behavior of a genuine rejector does not influence the behavior of acceptors (because a rejector is genuine with zero probability), any information that the mediator transmits to a genuine rejector can only increase his payoff, by allowing him to condition his best-response on that information. Hence, a genuine rejector should not receive any relevant signal from the mediator, and thus retain his prior belief and best-respond to the punishment.

[9] In a sequential equilibrium, beliefs formed off-path must satisfy "*no signaling what you don't know*". If types are independently distributed, this implies that a deviation by one agent does not influence beliefs about the other agents. See Fudenberg and Tirole (1991).

Two caveats are worthwhile remarking. The first is that limited commitment is crucial: if the acceptors and the principal could be constrained by the mechanism in case of a rejection by some other agent, the harshest possible penalty could be imposed on the rejector. The design of disagreement beliefs is only relevant in scenarios in which the principal cannot constrain neither himself nor the acceptors to follow the rules of the mechanism unless it is unanimously accepted. The second caveat is that we focus exclusively on equilibria in which agents always accept the principal's proposal. There may exist equilibria without full participation in which agents have higher expected payoffs. In such equilibria, spurious rejections that occur with a small probability would be insufficient to determine disagreement beliefs.

The remainder of the paper is structured as follows: the relation with the literature is described (Section 2); the model is introduced (Section 3); the single-agent case is analyzed (Section 4); the analysis is extended to the multiple-agent case (Section 5); illustrative examples and possible applications are provided (Section 6); concluding remarks are made (Section 7).

2 Literature review

2.1 Theory: mediated mechanism design

Myerson (1982) established the revelation principle for generalized principal-agent problems, where agents (with private information about their characteristics) choose actions that cannot be contracted upon. This means that a principal can restrict her choice of a coordination mechanism to those that are direct and incentive compatible: each agent is asked to report his information and receives a private recommendation of an action to choose; the profile of private recommendations is drawn from a distribution that depends on the profile of reported types, which is such that each agent is better off reporting truthfully and obeying the recommendation if all other agents report truthfully and obey the recommendations made to them.

To reconcile our setting with the environment of Myerson (1982), we can treat our mediator as the principal and our principal as an additional agent.[10] The mediator asks each agent to report his type, and recommends to each agent a pure strategy in the subsequent game. Each agent is recommended to choose the contract designed for his type, and, in case some other agent rejects the proposal, choose the action in the outside game that minimizes the payoff of the rejector. This relaxes participation constraints as much as possible. The problem is that obeying this extreme punishment may be *ex ante* optimal when the proposal is expected to be accepted by all agents but become suboptimal after a zero-probability rejection is observed (Selten, 1975; Kreps and Wilson, 1982).

[10] Limited commitment by the principal is transformed, through mediation, in moral hazard.

Since we are concerned with situations in which the threat of minimaxing a rejector is not credible, our setting must be framed as a multistage game with communication (Myerson, 1986; Forges, 1986).[11] In this framework, the revelation principle still holds, but the fact that the concept of Nash equilibrium is insufficiently restrictive becomes transparent. Myerson (1986) imposed an additional requirement of sequential rationality in the spirit of Kreps and Wilson (1982), and showed that it is satisfied if and only if *codominated* actions are never recommended to any player in any event.[12] Transposed to our setting, this means that any punishment that consists of actions that are not codominated can be induced in a sequential equilibrium.

Acknowledging that the exogenous trembles that generate beliefs off-path in a sequential equilibrium may involve trembling to codominated actions, Myerson (1986) proposed the concept of *predominant* equilibria by iteratively eliminating codominated actions from the game. This is in the spirit of the critiques by Cho and Kreps (1987) and others, who proposed refinements of sequential equilibrium based on restrictions of beliefs off-path.[13]. Myerson (1986) concluded that the harshest punishment that can be enacted in a predominant equilibrium may be much weaker than one that can be constructed using actions that are not codominated.

We also impose a stronger refinement than sequential rationality by requiring that all information sets are reached on-path. As a result, the punishments that we construct trivially satisfy any restriction of beliefs off-path (because there are none), and thus our solution concept is at least as restrictive as predominant equilibrium. Nevertheless, we are able to enforce any punishment that uses actions that are not codominated.

For all information sets to be reached on-path, we require a mediator who is able to generate any signal that agents can receive as a result of a deviation by the other agents. In our setting, this means that even if we allow an agent to transmit a public signal rejecting the proposal, we assume that the mediator can mimic this public signal. It is this signal jamming by the mediator that allows all information sets that result from a genuine rejection to be reached on-path.[14] As a result, the credibility of all punishments that do not rely on codominated actions is restored (equivalently, the credibility of all punishments that only use actions in the

[11] In our setting, a non-trembling mechanism defines the following multistage game with communication. In the first stage: each agent reports a type and is recommended to choose the contract designed for his type. If all agents choose a contract, the resulting grand contract is enforced; if some agent rejects the proposal, his rejection is announced. In the second stage (which only takes place if there is a rejection), each agent is recommended an action to carry out in the outside game. A trembling mechanism can be seen as a mediated stochastic mechanism in which rejection deterministically implies reversion to the outside game, while acceptance by all agents may either lead to the enforcement of the chosen grand contract or to reversion to the outside game.

[12] Loosely speaking, an action is *codominated* if and only if its recommendation implies that some player will gain from deviating, independently of the communication mechanism.

[13] See, for example, Grossman and Perry (1986), Kohlberg and Mertens (1986), Banks and Sobel (1987), van Damme (1989), Mertens (1989, 1991), Cho and Sobel (1990), Blume et al. (1991), Farrell (1993), and Mailath et al. (1993). See also Govindan and Wilson (2005) and the references therein.

[14] The term *signal jamming* was first used in game theory by Fudenberg and Tirole (1986).

support of the BCE^{I} with maximal support).

Mediation has been shown to be beneficial in mechanism design with imperfect commitment by the principal.[15] Assuming that the principal can commit to a first-stage decision (as a function of the agent's report) but not to a second-stage decision, Bester and Strausz (2007) showed that the principal benefits from using a communication device that, with some probability, manipulates the report of the agent.[16] Mediation also helps to provide incentives for effort provision in partnerships (Rahman and Obara, 2010).[17] By secretly appointing an agent to act as a budget-breaker by paying a large sum to the others if the outcome is good, even if the appointment only occurs with a small probability, effort can be incentivized while preserving budget-balance.[18] In strategic information transmission (Crawford and Sobel, 1982), it is also beneficial to introduce a mediator that manipulates the communication between the informed party and the decision-maker (Goltsman et al., 2009). In fact, even *noise* can be beneficial (Blume et al., 2007).[19] Finally, the benefits of mediation have also been highlighted in the literature on repeated games.[20] Kandori (1991) and Mailath et al. (2002) pointed out that an imperfect monitoring technology can contribute to sustaining an equilibrium by working as a trustworthy randomization device on the equilibrium path, but an unreliable one off the equilibrium path (in a way that punishes a deviator).[21]

2.2 Applications: mechanism design with an outside game

The idea that rejecting to participate in a mechanism conveys information that influences the outcome of the outside game led Cramton and Palfrey (1995) to propose a two-stage mechanism in which the decision to participate is prior to the actual play of the mechanism. In the first stage, firms either accept or reject the mechanism. If the mechanism is unanimously accepted,

[15] In our setting, limited commitment by the principal can be bypassed because the principal can be treated like any other agent that receives a recommendation and chooses an action that cannot be contracted upon.

[16] A conclusion in the same spirit was obtained by Mitusch and Strausz (2005). Bester and Strausz (2001) had extended the revelation principle to the case of imperfect commitment, under the assumption that the principal perfectly observes the report of the single agent.

[17] Strausz (2012) showed that the kind of contracts considered by Rahman and Obara (2010) are permitted in the general framework of Myerson (1982).

[18] In related contributions: Rahman (2012) explained how an owner can incentivize a supervisor to exert effort by, with a small probability, secretly asking a worker to shirk and making a large payment to the supervisor if he detects the deviation; while Rahman (2014) showed that a mediator can induce firms to respect a collusive agreement under imperfect monitoring by, with a small probability, secretly asking a firm to produce at capacity, and making a large payment to the other firm if the resulting market price is relatively high.

[19] In the work of Blume et al. (2007), *noise* refers to a limited form of manipulation: with some probability, the output message is drawn from a distribution that is independent of the input message. In the present model, a probability of spurious rejection that is independent of the input message profile would result in the outside game being played under *passive beliefs*: players would not learn anything from disagreement.

[20] See Forges (1988), Lehrer (1992), Renault and Tomala (2004), Tomala (2009), Mertens et al. (2015), and Sugaya and Wolitsky (2016), among others.

[21] See Kandori (2002) and the references therein.

firms move to the second stage, in which they report their types and an outcome is enforced as a function of the type profile. If any firm rejects the mechanism, the set of rejectors is publicly announced, firms update their beliefs and play the outside game.[22] They proposed the concept of *ratifiability*: an outcome is ratifiable if it is a sequential equilibrium of this two-stage game and beliefs satisfy neologism-proofness (Grossman and Perry, 1986; Farrell, 1993).

As an application, Cramton and Palfrey (1995) showed that efficient collusion is not ratifiable when the outside game is Cournot competition with private information about costs.[23] By rejecting to participate in the cartel, a firm credibly signals that it is efficient, and this increases its profit in the outside game (due to strategic substitutability). It is this anticipated increase in profit that makes it optimal for the firm to reject to collude.[24]

Assuming a greater ability of the principal to structure the interaction, we find the opposite result. The principal can design beliefs following disagreement if she is able to merge the ratification and communication stages and employ a mediator who can produce spurious rejections that mimic genuine rejections. This reverses the result of Cramton and Palfrey (1995), and all the non-ratifiability results in the literature that hinge on the inability of the principal to control what is learned from disagreement. For example, non-ratifiability of efficient collusive agreements in second-price auctions with participation costs (Tan and Yilankaya, 2007).

More generally, trembling mechanisms enlarge the set of outcomes that a mechanism designer is able to induce when the outside option of the agents is a game.[25] This is relevant to the literature on collusion-proof mechanism design in the presence of informational frictions among the members of the coalition, where the collusive side-contract is itself a mechanism whose outside option is the status quo mechanism. Applications of collusion-proof mechanism design include the design of organizations (Tirole, 1986, 1992), supplier networks (Laffont and Martimort, 1997; Faure-Grimaud et al., 2003; Mookherjee and Tsumagari, 2004; Che and

[22]The separation between a ratifying stage and a communication stage limits the payoff that the principal can attain. In our setting, it would imply a probability of spurious rejection that is independent of the type profile, and, hence, no learning from disagreement (*passive beliefs*).

[23]Contrarily to what had been concluded by Cramton and Palfrey (1990) and Kihlstrom and Vives (1992) under the assumption that firms do not learn from disagreement.

[24]Examining non-mediated two-stage mechanisms in which the only thing firms observe before playing the outside game is the set of rejectors, Celik and Peters (2011) showed that some outcomes are only implementable through mechanisms that are rejected on-path. Considering an uninformed firm and an informed firm whose cost may be low or high, they explained that: in an equilibrium with full participation, if the uninformed firm deviates and rejects to participate, the outside game is played with no belief updating; while, in an equilibrium in which the informed firm rejects to participate if it has a high cost, if the uninformed firm deviates and rejects to participate, the outside game is played with full information. In their example, disclosure hurts the uninformed firm in expectation, meaning that the equilibrium rejection relaxes its participation constraint. In mechanisms without a ratification stage (one-stage mechanisms), the benefits of equilibrium rejections can be achieved by making the type announcements public even if some agent rejects the mechanism. This would avoid the potential loss from inducing equilibrium rejections.

[25]Jullien (2000) studied mechanism design with type-dependent outside options. In our setting, the outside option is not only type-dependent but also belief-dependent.

Kim, 2006), mechanisms for public good provision (Laffont and Martimort, 2000), or optimal auctions (Dequiedt, 2007; Pavlov, 2008; Che and Kim, 2009), under the threat of collusion.[26]

The design of dispute resolution schemes (Cooter and Rubinfeld, 1989; Hörner et al., 2015; Balzer and Schneider, 2016) is another plausible application of trembling mechanisms. Dispute resolution usually involves private communication between a mediator and conflicting parties, and the failure to settle a dispute may well lead to some form of litigation under private information. In fact, in their study on the design of alternative dispute resolution schemes between litigants with private information about the cost of collective evidence, Balzer and Schneider (2016) proposed a trembling mechanism, and showed that it made the optimal settlement mechanism robust to the introduction of *ex post* participation constraints. By making, with a small probability, an unacceptable proposal that originates a kind of spurious rejection, the mediator is able to induce the acceptor to believe that the rejector must have a low cost of collecting evidence. In their model, this makes it optimal for the acceptor to collect a lot of evidence, thereby reducing the payoff of the rejector sufficiently to deter rejections.

Collusion and dispute resolution do not exhaust the scope for applications. Analogous issues arise in the design of any kind of agreement when disagreement leads to subsequent interaction under adverse selection. Applications may thus include the design of climate-change agreements (Martimort and Sand-Zantman, 2016), and other dynamic contracting settings (Philippon and Skreta, 2012; Tirole, 2012; Board and Pycia, 2014; Jullien et al., 2016).

3 The model

A principal (P), without private information, designs and proposes a *trembling mechanism* to a finite set of agents, $I \equiv \{1, ..., n\}$. Let $I_P \equiv I \cup \{P\}$. Each agent $i \in I$ is privately informed about his type, $\theta_i \in \Theta_i$. The set of possible type profiles is assumed to be finite and denoted by $\Theta \equiv \Pi_{i \in I} \Theta_i$. The actual type profile of the agents is drawn according to a commonly known probability distribution, $\mu^0 \in \Delta(\Theta)$, assumed to have full support.[27]

A trembling mechanism incorporates a *mediation device*, which is defined by: a set of private input message profiles from the agents, $M^I \equiv \Pi_{i \in I} M_i^I$; a set of private output message profiles, $M^O \equiv \Pi_{i \in I_P} M_i^O$; and transition probabilities from inputs to outputs, $\tau : M^I \to \Delta(M^O)$. We denote by $\tau(m^O | m^I)$ the probability of the output being $m^O \in M^O$ conditionally on the input message profile having been $m^I \in M^I$.

We consider mediation devices with the following structure. Each agent $i \in I$ can privately

[26] Particular forms of collusion that consist in bribing a rival have been considered, for example, by Schummer (2000), Eső and Schummer (2004), Chen and Tauman (2006), and Rachmilevitch (2013).

[27] We denote by $\Delta(Z)$ the set of probability distributions over Z.

report a type or reject the mechanism: $M_i^I \equiv \Theta_i \cup \{R_i\}$, where R_i is the rejection message. The set of agents who send the rejection message is denoted by $I_R^I \equiv \left\{i \in I : m_i^I = R_i\right\}$. If there is at least one rejector ($I_R^I \neq \emptyset$), the output of the mediation device to each player $i \in I_P$ is the true set of rejectors plus a private signal: $m_i^O = (I_R^O, s_i)$ such that $I_R^O = I_R^I$ and $s_i \in S_i$, where S_i is finite. If all agents accept the mechanism ($I_R^I = \emptyset$), the mediation device either truthfully transmits the input message profile, $m_i^O = m^I$, $\forall i \in I_P$, or mimics rejection by one of the agents, transmitting $m_i^O = (I_R^O, s_i)$ such that $I_R^O \in I$ and $s_i \in S_i$, $\forall i \in I_P$.[28] The set of private output messages to player $i \in I_P$ is, therefore, $M_i^O \equiv \Theta \cup (\mathcal{I} \times S_i)$, where $\mathcal{I}$ is the set of nonempty subsets of I.

A mediation device of this kind can be decomposed into a trembling device and a correlating device. A *trembling device*, $\varepsilon : \Theta \to \Delta(I \cup \{0\})$, is defined by the probabilities of generating a spurious rejection by each agent $i \in I$ when the input message profile is $\theta \in \Theta$, denoted $\varepsilon^i(\theta)$.[29] With probability $\varepsilon^0(\theta) = 1 - \sum_{i \in I} \varepsilon^i(\theta)$, there is no spurious rejection. A *correlating device*, $\psi \equiv \left(\psi^i, \psi^{(i)}\right)_{i \in I}$, is defined by the distributions over profiles of private signals that follow a spurious rejection by each agent $i \in I$ when the input message profile is $\theta \in \Theta$, denoted $\psi^i : \Theta \to \Delta(S)$, where $S \equiv \Pi_{i \in I_P} S_i$; and by those that follow a genuine rejection by each agent $i \in I$ when the input message profile from acceptors is $\theta_{-i} \in \Theta_{-i}$, denoted $\psi^{(i)} : \Theta_{-i} \to \Delta(S)$.[30]

A mechanism is said to be *strictly trembling* if and only if: any message that may be received by player $j \in I_P \setminus \{i\}$ of type $\theta_j \in \Theta_j$ when agent $i \in I$ genuinely rejects the mechanism (and others report truthfully), may also be received by player j of type θ_j when (all agents report truthfully and) there is a spurious rejection by agent i. Formally: $\sum_{\theta_{-ij} \in \Theta_{-ij}} \psi_j^{(i)}(s_j|(\theta_j, \theta_{-ij})) > 0 \Rightarrow \sum_{\theta_{-j} \in \Theta_{-j}} \varepsilon^i(\theta_j, \theta_{-j})\, \psi_j^i(s_j|(\theta_j, \theta_{-j})) > 0$, $\forall s_j \in S_j, \forall \theta_j \in \Theta_j, \forall j \in I_P \setminus \{i\}, \forall i \in I$.[31] A mechanism is said to be *non-trembling* if and only if $\varepsilon^0(\theta) = 1$, $\forall \theta \in \Theta$.

A *trembling mechanism*, (ε, ψ, x), is thus defined by a mediation device, (ε, ψ), and an allocation, $x : \Theta \to \Delta(X)$, which yields a lottery over a finite set of consequences, X, as a function of the type announcements of the agents (unless the output of the trembling device is a rejection message). The resulting payoff (expected utility) of each player $i \in I_P$, denoted $\pi_i(x(m^O), \theta) \in \mathbb{R}$, is linear in the first variable (probabilities).

If the output of the trembling device is a rejection message, principal and agents play a

[28] It is not useful for our purposes to generate spurious rejections by more than one agent, nor to generate a spurious rejection when some agent genuinely rejects the mechanism.

[29] With probability $\varepsilon^i(\theta)$, the message "*agent i has rejected*" is transmitted to all players.

[30] We will investigate several scenarios regarding the characteristics of the correlating device: the case in which there is no correlating device; the case in which players receive a public signal; the case in which signals are uncorrelated with the input message profile; and, finally, the general case in which the distribution over profiles of private signals depends on the input message profile.

[31] We denote by ψ_j^i and $\psi_j^{(i)}$ the marginal distributions over S_j derived from the joint distributions ψ^i and $\psi^{(i)}$, respectively.

single-stage outside game by simultaneously choosing an action, $a_i \in A_i$, for $i \in I_P$.[32] The resulting payoff of each player $i \in I_P$, denoted $\pi_i^R(y(\theta), \theta) \in \mathbb{R}$, is a linear function of the joint distribution over actions, $y(\theta) \in \Delta(A)$, where $A \equiv \Pi_{i \in I_P} A_i$.[33]

The timing of the interaction is the following:

1. Nature draws the type profile of the agents, $\theta \in \Theta$; each agent $i \in I$ observes θ_i.
2. The principal proposes a trembling mechanism, (ε, ψ, x), to the agents.
3. Each agent $i \in I$ chooses a private input message, $m_i^I \in \Theta_i \cup \{R_i\}$.
4. The output message profile, $m^O \in M^O$, is drawn from $\tau(m^I)$; player $i \in I_P$ observes m_i^O.
5. If $I_R^O = \emptyset$, the outcome $x(m^I)$ is enforced. If $I_R^O \neq \emptyset$, principal and agents play the outside game by simultaneously choosing an action, $a_i \in A_i$, for $i \in I_P$.

We focus on a game in which the principal proposes a given trembling mechanism, (ε, ψ, x), which is said to be *incentive compatible* and *individually rational* if and only if there exists a sequential equilibrium of the game in which agents always accept to participate and report their types truthfully.

Our objective is to characterize the set of allocations that are *virtually t-feasible* in the sense that there exists a sequence of incentive compatible and individually rational strictly trembling mechanisms, $(\varepsilon^m, \psi^m, x^m)_{m \in \mathbb{N}}$, converging to a non-trembling mechanism that proposes the desired allocation, $(0, \psi, x)$.[34] Virtual t-feasibility of x thus means that an allocation arbitrarily close to x can be brought into effect with a probability arbitrarily close to one, in an equilibrium such that no unilateral deviation can lead to an information set off the equilibrium path.[35]

4 The single-agent case

We start by analyzing the case in which the principal interacts with a single agent. This case is much simpler because we only need to worry about the belief of the principal about the agent, and because there are only two players and one-sided private information in the outside game.

[32] A player $i \in I_P$ may be inactive in the outside game (e.g., his action set, A_i, may be a singleton).

[33] Linearity of $\pi_i(\cdot, \theta)$ and $\pi_i^R(\cdot, \theta)$ simply means that players maximize expected utility.

[34] Observe that, since Θ, I, S and X are finite, the space of possible mechanisms is a compact subset of a finite-dimension Euclidean space.

[35] In the usual sense: an allocation x is *virtually* feasible if and only if there exists a feasible allocation that is arbitrarily close to x (Abreu and Matsushima, 1992).

4.1 Non-mediated mechanism

As a benchmark, consider a mechanism that is non-trembling (i.e., does not produce spurious rejections) and non-correlating (i.e., does not send any additional signals to principal and agent accompanying the announcement of a rejection). Such a mechanism, denoted $(0, 0, x)$, is completely defined by the allocation, $x : \Theta \to X$, that is proposed.

Expecting the distribution over action profiles in the outside game, as a function of his type, to be given by $y : \Theta \to \Delta(A)$, the agent finds it optimal to participate and report his type truthfully if and only if the following incentive compatibility and individual rationality constraints are satisfied:

$$\pi_1(x(\theta), \theta) \geq \pi_1(x(\theta'), \theta), \ \forall \theta' \neq \theta, \ \forall \theta \in \Theta \tag{IC}$$

$$\pi_1(x(\theta), \theta) \geq \pi_1^R(y(\theta), \theta), \ \forall \theta \in \Theta. \tag{IR}$$

For the allocation x to be the outcome of a sequential equilibrium, in addition to (IC) and (IR) being satisfied for some outside option y, there must exist consistent beliefs following disagreement such that there exists a Bayesian Nash Equilibrium (BNE) of the outside game that induces y.

Let $\mu \in \Delta(\Theta)$ denote the commonly known belief of the principal about the type of the agent when the outside game is played (we can restrict μ to be commonly known because our solution concept is sequential equilibrium), and let $BNE(\mu)$ denote the set of distributions over action profiles induced by BNE of the outside game when the disagreement belief is μ.[36]

In a sequential equilibrium in which the mechanism is never rejected, beliefs following disagreement are formed off the equilibrium path. Therefore, any μ that is common knowledge is consistent. Notice that the "*no signaling what you don't know*" condition is trivially satisfied (there is nothing that the agent who deviates does not know).[37]

An allocation $x : \Theta \to X$ is said to be *0-feasible* if and only if the non-mediated mechanism, $(0, 0, x)$, is incentive compatible and individually rational.

Remark 1. *An allocation x is 0-feasible if and only if x satisfies (IC) and (IR) for some $y \in \cup_{\mu \in \Delta(\Theta)} BNE(\mu)$.*

[36]Formally: $y \in BNE(\mu)$ if and only if $y(\theta, a_P, a_1) = \sigma_P(a_P)\,\sigma_1(\theta, a_1)$, $\forall (\theta, a_P, a_1) \in \Theta \times A_P \times A_1$, where the strategy of the agent, $\sigma_1 : \Theta \to \Delta(A_1)$, is such that $\sigma_1(\theta)$ is a best-response to σ_P if the agent has type θ, for all $\theta \in \Theta$; and the strategy of the principal, $\sigma_P \in \Delta(A_P)$ is the best-response to σ_1 if she believes that the agent is of type $\theta \in \Theta$ with probability $\mu(\theta)$.

[37]In this benchmark, the interaction can be described by a two-stage game with observable actions, thus a sequential equilibrium is equivalent to a perfect Bayesian equilibrium in which beliefs formed off the equilibrium path are common knowledge and satisfy "*no signaling what you don't know*" (Fudenberg and Tirole, 1991).

With non-mediated mechanisms, the set of credible punishments is the union of the sets of BNE of the outside game over all possible disagreement beliefs, $\cup_{\mu\in\Delta(\Theta)} BNE(\mu)$.

4.2 Non-trembling mechanism

Another benchmark is a non-trembling mechanism that incorporates a correlating device. The difference with respect to a non-mediated mechanism is that, after a rejection, principal and agent receive private signals on which they may condition their strategies in the outside game.[38]

This means that the outside game becomes an extended outside game, whose Bayesian Nash Equilibria are, by definition, *Uninformed Mediator Bayesian Correlated Equilibria* (BCE^{U}) of the non-extended outside game. Denote by $BCE^U(\mu)$ the set of distributions over action profiles induced by the set of such equilibria when the disagreement belief is μ.[39]

An allocation $x : \Theta \to X$ is *nt-feasible* if and only if there is an incentive compatible and individually rational non-trembling mechanism proposing the allocation, $(0, \psi, x)$.

Remark 2. *An allocation x is nt-feasible if and only if it satisfies (IC) and (IR) for some $y \in \cup_{\mu\in\Delta(\Theta)} BCE^U(\mu)$.*

Hence, with non-trembling mechanisms, the set of credible punishments is the union of the sets of BCE^{U} of the outside game over all possible disagreement beliefs, $\cup_{\mu\in\Delta(\Theta)} BCE^U(\mu)$.

It is trivial that $BNE(\mu) \subseteq BCE^U(\mu)$, but the inclusion may not be strict. A correlating device relaxes the participation constraints whenever there exists $y \in BCE^U(\mu)$, for some $\mu \in \Delta(\Theta)$, that yields a lower payoff to an agent of some type when compared with any $y' \in BNE^U(\mu')$, for any $\mu' \in \Delta(\Theta)$. An example is given in Appendix A.1.

A caveat regarding the credibility of the punishments that sustain 0-feasibility and *nt*-feasibility is that, as the solution concept is sequential equilibrium, any interpretation by the principal about the information content of the agent's off-path choice to reject the proposal is allowed, since consistency of beliefs only implies that μ is common knowledge. A stronger refinement, such as the intuitive criterion (Cho and Kreps, 1987), some version of divinity (Cho and Sobel, 1990), or neologism-proofness (Grossman and Perry, 1986; Farrell, 1993), could further restrict beliefs formed off-path (and, as a result, the set of feasible allocations) by ruling

[38] Private signals may be recommendations of actions to make as a function of own type.

[39] Forges (1993) describes several coherent definitions of correlated equilibrium in games with incomplete information. This one is designated as *strategic form correlated equilibrium* and is defined as follows. Considering the outside game in its strategic form, a pure strategy by player $i \in I_P$ is a mapping $\tilde{a}_i : \Theta_i \to A_i$, and a pure strategy profile is a vector $\tilde{a} \in \tilde{A} \equiv \Pi_{i\in I_P} \tilde{A}_i$. A distribution over profiles of private recommendations, $y \in \Delta(\tilde{A})$, is a strategic form correlated equilibrium, $y \in BCE^U(\mu)$, if and only if it is optimal for players to obey the recommendations, i.e., $y(\tilde{a}_{-i}|\tilde{a}_i)\pi_i^R(\tilde{a}_i, \tilde{a}_{-i}) \geq y(\tilde{a}_{-i}|\tilde{a}_i)\pi_i^R(\tilde{a}'_i, \tilde{a}_{-i})$, $\forall \tilde{a}'_i \in \tilde{A}_i, \forall i \in I_P$.

out unreasonable conjectures of the principal about the possible types of the rejector and their relative probabilities.

4.3 Trembling devices and strict incentive compatibility

Now consider a trembling device that, when the agent reports type $\theta \in \Theta$, produces a spurious rejection with probability $\varepsilon(\theta) = \frac{\mu(\theta)}{\mu^0(\theta)}\epsilon$, where $0 < \epsilon < \min_{\theta\in\Theta} \mu^0(\theta)$ and $\mu \in \Delta(\Theta)$.[40] With such a device, we do not need to worry about beliefs formed off-path after a rejection, because, in an equilibrium in which the agent always accepts the mechanism and reports truthfully, rejection occurs with positive probability. The belief of the principal following disagreement is determined by Bayesian updating, which yields the probability distribution $\mu \in \Delta(\Theta)$.

Let $v_1(\theta', \theta)$ be the agent's expected payoff after a spurious rejection, where θ' is his report and θ is his type. The agent finds it optimal to participate and report truthfully if and only if:

$$\pi_1(x(\theta), \theta) + \tfrac{\varepsilon(\theta)}{1-\varepsilon(\theta)}\, v_1(\theta, \theta) \geq \tfrac{1-\varepsilon(\theta')}{1-\varepsilon(\theta)}\, \pi_1(x(\theta'), \theta) + \tfrac{\varepsilon(\theta')}{1-\varepsilon(\theta)} v_1(\theta', \theta),\ \forall \theta' \neq \theta,\ \forall \theta \in \Theta \qquad \text{(IC')}$$

$$\pi_1(x(\theta), \theta) \geq \pi_1^R(y(\theta), \theta),\ \forall \theta \in \Theta. \qquad \text{(IR)}$$

By generating spurious rejections that are correlated with the agent's report, trembling devices distort the incentives for truth-telling. Luckily, the distortion can be made arbitrarily small, since disagreement beliefs, μ, do not depend on the *ex ante* probability of spurious rejection, $\epsilon > 0$. Given any allocation, x, that strictly satisfies all the incentive compatibility (IC) conditions in a non-trembling mechanism, there is a trembling probability, $\epsilon > 0$, that is small enough for all the incentive compatibility (IC') conditions to remain strictly satisfied in the trembling mechanism.

The following assumption ensures that any allocation that satisfies (IR) and (IC) can be approximated by a sequence of allocations that satisfy (IR) and strictly satisfy (IC).[41]

Assumption 1. *There exists an allocation, x^f, that strictly satisfies (IC) and satisfies (IR) for any distribution y.*

The approximation can be made through a sequence of (IR) and strictly (IC) allocations that are weighted averages between the allocation that we are approximating (whose weight increases along the sequence) and x^f (whose weight vanishes in the limit).

A trembling device allows the principal to design the disagreement belief, μ, which becomes the result of Bayesian updating in equilibrium (instead of being a postulated off-the-

[40] The upper bound on ϵ guarantees that the conditional probability of spurious rejection is lower than 100%.

[41] I am grateful to Takuro Yamashita for suggesting this assumption and the subsequent argument, which he used in Yamashita (2014).

equilibrium-path belief that could violate appropriate refinements). The costs of this design are the strictly positive probability of reversion to the outside game and the possible necessity of strict incentive compatibility.[42] As a result, an allocation that is feasible with a non-trembling mechanism becomes only *virtually* feasible if the principal uses a strictly trembling mechanism.

4.4 Trembling mechanism with no correlating device

Let us start by considering the simplest kind of trembling mechanism (one with no correlating device): the mediator produces a spurious rejection with strictly positive probability, $\epsilon > 0$, but does not transmit any signals besides the announcement of rejection.

An allocation $x : \Theta \to X$ is said to be virtually *tnc-feasible* if and only if there exists a sequence of incentive compatible and individually rational strictly trembling mechanisms with no correlating device, $(\varepsilon^m, 0, x^m)_{m\in\mathbb{N}}$, that converges to a non-mediated mechanism that proposes the allocation, $(0, 0, x)$.

Proposition 1. *An allocation is virtually tnc-feasible if and only if it is* 0*-feasible.*

Despite the equivalence in Proposition 1, strictly trembling mechanisms have an advantage: beliefs are formed on-path, according to Bayes' rule. Thus, refinements based on restrictions of beliefs formed off-path are trivially satisfied because there are no information sets off-path.

4.5 Trembling mechanism with public correlating device

Allowing the mediator to send a public signal after announcing a rejection expands the set of credible threats to the convex hull of the set of Bayesian Nash Equilibria of the outside game.

The principal becomes able to induce any mixture of BNE distributions over action profiles, $y \in Conv(\cup_{\mu\in\Delta(\Theta)} BNE(\mu))$, because the disagreement belief, $\mu^k \in \Delta(\Theta)$, and the continuation equilibrium, $y^k \in BNE(\mu^k)$, can be made dependent on the public signal, $k \in K$.[43]

An allocation $x : \Theta \to X$ is said to be virtually *tpc-feasible* if and only if there exists a sequence of incentive compatible and individually rational strictly trembling mechanisms with public correlation, $(\varepsilon^m, \psi^{pc,m}, x^m)_{m\in\mathbb{N}}$, that converges to a non-trembling mechanism that proposes the allocation, $(0, \psi^{pc}, x)$.

[42] Strictness of all incentive compatibility conditions is sufficient, but not necessary, to accommodate the distortions caused by the trembling device.

[43] From Caratheodory's theorem, any given $y \in Conv(\cup_{\mu\in\Delta(\Theta)} BNE(\mu))$ can be written as a convex combination of a finite number, $|A|$, of distributions.

Proposition 2. *An allocation x is virtually tpc-feasible if and only if it satisfies (IC) and (IR) for some $y \in Conv(\cup_{\mu\in\Delta(\Theta)} BNE(\mu))$.*

A public correlating device allows the principal to choose the disagreement belief and the resulting continuation equilibrium as a function of the agent's report when the rejection is spurious. However, when the rejection is genuine, since the agent does not report his type, the principal cannot condition the punishment on the type reported by the agent.

Allowing the agent to announce his type when he rejects the mechanism would not enable the principal to harshen the punishment, because the agent would not be willing to provide information that would only be used to punish him. On the other hand, as shown by Lehrer and Sorin (1997), if principal and agent could send input messages that are rich enough to define an encryption code, the common signal could become in practice a profile of private signals because each player would not be able to decode the parcel of the common signal intended to be decoded by the rival.[44]

4.6 Trembling mechanism with extraneous correlating device

If the trembling device, when announcing a rejection, also sends a profile of private signals that is uncorrelated with the announcement made by the agent (i.e., is extraneous), the outside game becomes an extended game in which principal and agent observe correlated private signals whose distribution is independent of the agent's type, and may condition their strategies on these signals. As already mentioned, a BNE of this extended outside game is, by definition, a BCE^{U} of the non-extended outside game.

An allocation $x : \Theta \to X$ is said to be virtually *tec-feasible* if and only if there exists a sequence of incentive compatible and individually rational strictly trembling mechanisms with extraneous correlation, $(\varepsilon^m, \psi^{ec,m}, x^m)_{m\in\mathbb{N}}$, that converges to a non-trembling mechanism that proposes the allocation, $(0, \psi^{ec}, x)$.

Proposition 3. *An allocation is virtually tec-feasible if and only if it is nt-feasible.*

The set of credible punishments is the same as in the case of non-trembling mechanisms: the union of the sets of BCE^{U} of the outside game across the set of possible (common knowledge) disagreement beliefs, $\cup_{\mu\in\Delta(\Theta)} BCE^U(\mu)$.

This means that extraneous private signals can be used to punish the agent more harshly, whenever the outside game has some $y \in BCE^U(\mu)$, for some $\mu \in \Delta(\Theta)$, that yields a lower payoff to an agent of some type when compared to any $y' \in BNE(\mu')$, for any $\mu' \in \Delta(\Theta)$.[45]

[44] The scenario in which the mediator sends a profile of private signals is studied in Section 4.7.

[45] See Appendix A.1 for an example.

4.7 Trembling mechanism with general correlating device

We now address the general case where the mediator is able to send a profile of private signals to principal and agent according to a probability distribution that depends on the input message.

In an equilibrium with truthful participation: after a spurious rejection, since the agent reports truthfully (on-path), the mediator is informed and can thus send private recommendations to principal and agent that depend on the agent's type; after a genuine rejection (off-path), the mediator is uninformed and thus cannot condition recommendations on the agent's type. Unable to distinguish genuine from spurious rejections, the principal presumes that a rejection is spurious and obeys any recommendation that she would obey after a spurious rejection.[46]

Formally, after announcing a (spurious or genuine) rejection, the mediator sends a profile of private signals distributed according to $\psi : M^I \to \Delta(T \times A)$, where $T \equiv \Pi_{i \in I_P} T_i$ is a set of private signal profiles, and $A \equiv \Pi_{i \in I_P} A_i$ is the set of action profiles in the outside game. It is equivalent to think of private signals as being sent in two stages: first, through $\psi^T : M^I \to \Delta(T)$; second, through $\psi^A : M^I \times T \to \Delta(A)$. If the input message is $\theta \in \Theta$, the trembling device generates a spurious rejection with probability $\varepsilon(\theta) \equiv \frac{\mu(\theta)}{\mu^0(\theta)}\epsilon$, with $\epsilon > 0$. A spurious rejection is thus generated with *ex ante* probability ϵ, and the common knowledge disagreement belief is $\mu \in \Delta(\Theta)$. As will be shown, the common disagreement belief is not relevant as long as it has full support.[47]

The continuation game after a spurious rejection is an extended outside game in which principal and agent receive private signals that may be correlated with the type of the agent. This game was studied by Bergemann and Morris (2016), with their common prior corresponding to our common disagreement belief, μ. By definition, a BNE of this extended outside game is an *Informed Mediator Bayesian Correlated Equilibrium* (BCE^{I}) of the non-extended game.[48]

[46] After a genuine rejection, the mediator should send to the principal the recommendation that is the most harmful to the agent among those that the principal obeys. Whether the principal has incentives to obey it conditionally on a genuine rejection is not relevant. The principal obeys if and only if she has incentives to obey conditionally on a spurious rejection (which is infinitely more likely than a genuine rejection).

[47] The mediator could generate a mixture over disagreement beliefs, but this would not be useful. As long as it has full support, the common disagreement belief, μ, is irrelevant because the support of the set of recommendations that are obeyed in a BCE^{I} does not depend on μ (see Lemma 1 in Appendix A.4).

[48] The work of Bergemann and Morris (2016) is instrumental for our purpose. They considered an environment in which an informed mediator commits *ex ante* to an information structure, $\psi^T : \Theta \to \Delta(T)$, according to which a profile of private signals, $t \in T$, with $t = (t_P, t_1, ..., t_n)$ and $T \equiv \Pi_{i \in I_P} T_i$, is drawn from a distribution that depends on the state of nature, $\theta \in \Theta$. The state of nature is observed by the mediator but not by the players, who share a common prior, $\mu \in \Delta(\Theta)$, with full support. The mediator also commits *ex ante* to a decision rule, $\psi^A : T \times \Theta \to \Delta(A)$, according to which it sends a profile of private recommendations of actions to each player, $a = (a_P, a_1, ..., a_n) \in A \equiv \Pi_{i \in I_P} A_i$. The distribution over profiles of private recommendations may depend on the state of nature, $\theta \in \Theta$, and on the realization of the information signal, $t \in T$. Players obey the recommendations if it is in their interest. A decision rule that is always obeyed induces an outcome $\psi : \Theta \to \Delta(A)$ such that $\psi(a|\theta) = \sum_{t \in T} \psi^A(a|t, \theta)\, \psi^T(t, \theta)$, designated as a *Bayes Correlated Equilibrium*.

In our setting, the game that follows a spurious rejection is equivalent to the game considered by Bergemann and Morris (2016) under the restriction that ψ^T must, at least, transmit to each agent the information about his own type (which, in our setting, agents already possess). An *Informed Mediator Bayesian Correlated*

Bergemann and Morris (2016) showed that any relevant private signal besides a recommendation of an action to be chosen shrinks the set of BCE^{I}, i.e., shrinks the set of recommendation rules that are obeyed. Providing additional information to the players generates additional incentive compatibility conditions, which, taken together, are typically stronger and are never weaker than the pooled incentive compatibility condition that arises from a given recommendation when no further information is provided. Therefore, we can restrict private output messages to consist of a private recommendation of an action to be chosen, $a_i \in A_i$.

Off-path, after a genuine rejection, the messages received by principal and agent should be uncorrelated because it is advantageous to conceal the behavior of the principal from the agent. Any relevant signal sent to the agent after a genuine rejection can only increase his payoff by allowing him to condition his best-response (notice that the signal sent to the agent does not influence the decision of the principal, who attributes zero probability to the rejection having been genuine). Therefore, the trembling device should completely conceal from the agent the recommendation made to the principal (by sending an uninformative signal to the agent).[49]

In sum: on-path (after a spurious rejection), the trembling device privately sends recommendations according to some $y \in BCE^I(\mu)$; off-path (after a genuine rejection), the trembling device privately sends to the principal a recommendation, $a_P \in A_P$, that is also sent on-path, drawn according to $\sigma_P \in \Delta(supp(y_P|\mu))$.[50] After a genuine rejection, the principal wrongly presumes that she is on-path with a commonly known disagreement belief, μ, and obeys any recommendation that is also made on-path; while the agent knows that he is off-path and best-responds to the distribution over recommendations sent to the principal in case of a genuine rejection. This distribution, $\sigma_P \in \Delta(supp(BCE_P^I))$, is necessarily independent of the agent's type, and should be constructed in the way that minimizes the agent's best-response payoff.[51] It is possible that the harshest punishment depends on the agent's type. In that case, the principal must find the optimal trade-off between punishing different types, taking into account the shadow value of the respective participation constraints.

An allocation $x : \Theta \to X$ is said to be virtually *t-feasible* if and only if there exists a sequence of incentive compatible and individually rational strictly trembling mechanisms,

Equilibrium is, therefore, a *Bayes Correlated Equilibrium* in which each agent is at least informed about his own type. Equivalently, it is a *Bayesian Solution* (Forges, 1993, 2006).

[49] If the principal knew that the agent was not receiving the correlating signal that the agent receives on-path, she might prefer to disobey the recommendation.

[50] Given a distribution over action profiles, $y(\theta) \in \Delta(A_P \times A_1)$, we denote the marginal distribution over A_P by $y_P(\theta) \in \Delta(A_P)$. We define $supp(y_P|\mu) \equiv \cup_{\theta \in supp(\mu)} supp(y_P(\theta))$, and, writing $y_P \in BCE_P^I(\mu)$ if and only if $y \in BCE^I(\mu)$, we also define $supp(BCE_P^I(\mu)) \equiv \cup_{y \in BCE^I(\mu)} supp(y_P|\mu)$ and $supp(BCE_P^I) \equiv \cup_{\mu \in \Delta(\Theta)} supp(BCE_P^I(\mu))$.

[51] Restricting recommended actions to belong to the support of a single BCE^{I} is without loss of generality. If μ has full support, there exists $y \in BCE^I(\mu)$ whose support contains all the supports of all BCE^{I} under all possible disagreement beliefs (see Lemma 1 in Appendix A.4).

$(\varepsilon^m, \psi^m, x^m)_{m\in\mathbb{N}}$, converging to a non-trembling mechanism that proposes it, $(0, \psi, x)$.

Proposition 4. *An allocation x is virtually t-feasible if and only if it satisfies (IC) and (IR) for some y induced by (σ_P, σ_1) such that $\sigma_P \in \Delta(supp(BCE_P^I))$ and $\sigma_1(\theta)$ is a best-response to σ_P, $\forall\theta \in \Theta$.*

In this most general case, the set of credible punishments is the set of mixed strategies, $\sigma_P \in \Delta(A_P)$, whose support is contained in the union across disagreement beliefs, $\mu \in \Delta(\Theta)$, of the union of the supports of marginal distributions $y_P \in BCE_P^I(\mu)$. Equivalently, whose support is contained in the support of the marginal distribution of a BCE^{I} with maximal support.

Comparing the set of credible threats with general correlating devices with the one with extraneous correlating devices, we can distinguish three advantages of sending private signals according to a distribution that depends on the input message. First, the potential support of the principal's mixed action becomes larger (because any BCE^{U} is also a BCE^{I}).[52] Second, any distribution over that support is allowed (not only the equilibrium distribution). For example, if there is an action that punishes the agent more than any other, the principal can play it with 100% probability. Third, if the principal plays a mixed action, the correlating signal to the agent can be shut off (to prevent the agent from conditioning his best-response).

The second advantage is illustrated in the example of an all-pay auction with bid caps (Section 6.2.1). The first and the third advantages do not appear in our single-agent examples, where private information is only one-sided. They materialize in the Cournot triopoly example, where there are multiple agents, and thus private information is multi-sided (Section 6.1.3).

4.8 Summary

The benefit of using trembling mechanisms is the expansion of the set of credible punishments. Below, we compare the sets of credible punishments for different mechanism formats:

- Non-trembling, non-correlating: $y \in \cup_{\mu\in\Delta(\Theta)} BNE(\mu)$
- Non-trembling, correlating signals: $y \in \cup_{\mu\in\Delta(\Theta)} BCE^U(\mu)$
- Trembling, non-correlating: $y \in \cup_{\mu\in\Delta(\Theta)} BNE(\mu)$
- Trembling, public correlating signals: $y \in Conv(\cup_{\mu\in\Delta(\Theta)} BNE(\mu))$
- Trembling, extraneous correlating signals: $y \in \cup_{\mu\in\Delta(\Theta)} BCE^U(\mu)$
- Trembling, correlating signals: $y_P \in \Delta(\cup_{\mu\in\Delta(\Theta)} supp(BCE_P^I(\mu)))$.

[52] A $BCE^U(\mu)$ is a $BCE^I(\mu)$ in which recommendations are independent of the type of the agent.

It may seem that, in the absence of general correlating signals, trembling mechanisms do not improve on non-trembling mechanisms. There is actually an important improvement. Since a strictly trembling mechanism generates disagreement on-path, feasibility is robust to the introduction of refinements based on restrictions of beliefs formed off-path.

Regarding the usefulness of correlating signals, we verify that public signals only convexify the set of credible threats, while extraneous private signals are only advantageous in relatively specific outside games (by enlarging the space of punishments from the set of BNE to the possibly larger set of BCE^{U}). The usefulness of general private signals is more evident: credibility of a punishment only requires that any action recommended to the principal is also recommended in some BCE^{I}.[53]

5 The multiple-agent case

With multiple agents, further complexity arises from multi-sided private information. A genuine rejector faces uncertainty about the characteristics of the other agents, and his beliefs are not amenable to manipulation by the trembling device.[54] As a result, after a genuine rejection, the outside game is played under beliefs that exhibit a particular violation of the common prior assumption: acceptors (presuming that the rejection was spurious) update their beliefs according to Bayes' rule, and believe that all players do so; while the genuine rejector retains his prior belief, and knows that all other players have updated their beliefs.[55]

We will continue to consider candidate sequential equilibria in which agents always participate and report truthfully, and focus on whether any agent of any type can gain by unilaterally rejecting the mechanism. To extend the incentive compatibility (IC') and individual rationality (IR) conditions to the multiple-agent case, we must consider expected payoffs according to the *interim* expectation of agent $i \in I$ of type $\theta_i \in \Theta_i$ about the type profile, $\theta \in \Theta$, denoted $\mathbb{E}_{\theta|\theta_i}[\pi(\theta)] \equiv \sum_{\theta \in \Theta} \mu^0(\theta|\theta_i)\pi(\theta)$. We also introduce further notation: let $y^i(\theta)$ be the distribution over action profiles that agent $i \in I$ expects if he genuinely rejects the mechanism, as a function of the type profile, $\theta \in \Theta$; and let $v_i^j(\theta_i', \theta)$ be the payoff that agent $i \in I$ expects if there is a spurious rejection by agent $j \in I$, as a function of his report, $\theta_i' \in \Theta_i$, and the type profile, $\theta \in \Theta$.

If all other agents participate and report truthfully, it is optimal for agent $i \in I$ to participate

[53] It is important to keep in mind that the punishment strategy cannot depend on the agent's type.

[54] In the single-agent case, the single rival of the rejector does not have any private information.

[55] Here, and henceforth, the principal is referred to as an acceptor, although she proposes the mechanism and does not explicitly accepts it.

and report truthfully if and only if:

$$\mathbb{E}_{\theta|\theta_i}\left[\pi_i(x(\theta),\theta)+\sum_{j\in I}\frac{\varepsilon^j(\theta)}{\varepsilon^0(\theta)}\,v_i^j(\theta_i,\theta)\right] \geq \mathbb{E}_{\theta|\theta_i}\left[\frac{\varepsilon^0(\theta_i',\theta_{-i})}{\varepsilon^0(\theta)}\,\pi_i(x(\theta_i',\theta_{-i}),\theta)\right.$$
$$\left.+\sum_{j\in I}\frac{\varepsilon^j(\theta_i',\theta_{-i})}{\varepsilon^0(\theta)}v_i^j(\theta_i',\theta)\right],\ \forall\theta_i'\neq\theta_i,\ \forall\theta_i\in\Theta_i \quad \text{(IC')}$$

$$\mathbb{E}_{\theta|\theta_i}\left[\pi_i(x(\theta),\theta)\right] \geq \mathbb{E}_{\theta|\theta_i}\left[\pi_i^R(y^i(\theta),\theta)\right],\ \forall\theta_i\in\Theta_i. \quad \text{(IR)}$$

We say that (IC) is satisfied if and only if (IC') is satisfied with $\varepsilon^0(\theta)=1$, $\forall\theta\in\Theta$.

As in the single-agent case, we start by considering non-mediated and non-trembling mechanisms. Then, we focus on trembling mechanisms without correlating devices and with general correlating devices. The cases in which correlating devices are public or extraneous are only briefly discussed because they are complex and do not help to clarify the general case.[56]

We will work under the natural extension of Assumption 1 to the multiple-agent case.

Assumption 1'. *There exists an allocation, x^f, that, $\forall i\in I$, strictly satisfies (IC) and satisfies (IR) for any distribution y^i.*

5.1 Non-mediated mechanism

In the absence of mediation, as we consider sequential equilibria with participation and truthful reporting, a rejection by agent $i\in I$ can only directly influence beliefs about himself. The set of possible disagreement beliefs is composed by those that satisfy *"no signaling what you don't know"*: $\mathcal{B}_0^i\equiv\{\mu^i\in\Delta^*(\Theta):\ \mu^i(\theta|Z)=\mu^i(\theta_i|Z)\,\mu^0(\theta|\theta_i\cap Z),\ \forall\theta\in\Theta\}$, where $\Delta^*(\Theta)$ is the set of Bayesian conditional probability systems on Θ.[57]

After rejection by agent $i\in I$, the common belief becomes $\mu^i\in\mathcal{B}_0^i$: it is common knowledge that the probability that player $j\in I_P$ attributes to the type profile being $\theta\in\Theta$ is $\mu^i(\theta|\theta_j)$.[58]

[56] In the case of public correlation, it is necessary to trade off tailoring the punishment to the type profile of acceptors with the fact that such tailoring may convey information that allows the rejector to better respond to the punishment. In the case of extraneous correlation, it is necessary to deal with the mentioned violation of the common prior assumption: after a genuine rejection, acceptors update their beliefs and believe that all players do so, while the rejector maintains his prior belief knowing that other players have updated their beliefs.

[57] With multiple agents and non-trembling mechanisms, it is convenient to define a common belief as a conditional probability system (Myerson, 1986), because the disagreement belief may attribute zero marginal probability to some type of rejector and it may be necessary to consider the *interim* belief of a genuine rejector of this type. A conditional probability system is a $\mu\in\Delta^*(\Theta)$ such that, for every Z that is a nonempty subset of Θ, the conditional probability function $\mu(\cdot|Z)\in\Delta(\Theta)$ is such that $\mu(Z|Z)=1$ and $\mu(Z''|Z)=\mu(Z''|Z')\,\mu(Z'|Z)$, for all $Z''\subseteq Z'\subseteq Z$. We will continue to denote $\mu(Z')\equiv\mu(Z'|\Theta)$, $\forall Z'\subseteq\Theta$.

[58] Notice that $\mu^i\in\mathcal{B}_0^i$ implies that the posterior of player i is unchanged: $\mu^i(\theta|\theta_i)=\mu^i(\theta_i|\theta_i)\mu^0(\theta|\theta_i\cap\theta_i)=\mu^0(\theta|\theta_i)$, $\forall\theta\in\Theta$. The use of conditional probability systems allows $\mu^i(\theta|\theta_i)$ to be well defined even if $\mu^i(\theta_i)=0$.

Proposition 5. *An allocation x is* 0*-feasible if and only if, for each $i \in I$, it satisfies (IC) and (IR) for some $y^i \in \cup_{\mu \in \mathcal{B}_0^i} BNE(\mu)$.*

5.2 Non-trembling mechanism

Even without producing spurious rejections, a mediator can enlarge the set of credible threats. Being informed about the type profile of acceptors, $\theta_{-i} \in \Theta_{-i}$ (but not about the type of the rejector), the mediator can make recommendations that, besides correlating players' actions (as in the single-agent case), may convey information about the type profile of acceptors.

Designate a BNE of the extended outside game in which a mediator makes recommendations that depend on θ_{-i} but not on θ_i as a *Partially Informed Mediator Bayesian Correlated Equilibrium*, and denote by $BCE^{-i}(\mu)$ the set of such equilibria when the disagreement belief is $\mu \in \Delta^*(\Theta)$.

Proposition 6. *An allocation x is nt-feasible if and only if, for each $i \in I$, it satisfies (IC) and (IR) for some $y^i \in \cup_{\mu \in \mathcal{B}_0^i} BCE^{-i}(\mu)$.*

5.3 Trembling mechanism with no correlating device

When a strictly trembling mechanism is used, although we focus on equilibria with participation and truth-telling, beliefs of acceptors in the face of disagreement are determined on-path. Acceptors always presume that the rejection is spurious, and update their beliefs accordingly.

If the mediator does not send additional signals to the players besides the announcement of rejection, the beliefs of an acceptor when playing the outside game result from two pieces of information: the rejection, which is public; and the information about own type, which is private. Acceptors can be seen as having a common disagreement belief, which results from Bayesian updating of the common prior, μ^0, after rejection is publicly observed.[59]

The principal can induce any disagreement belief that is compatible with each acceptor's private information. For an example of incompatibility, suppose that μ^i attributed zero probability to agent $j \in I \setminus \{i\}$ being of type $\hat{\theta}_j$, which means that a spurious rejection by agent i is never produced if agent j reports type $\hat{\theta}_j$. Then, if agent j is of type $\hat{\theta}_j$, he is able to infer that a rejection by agent i has been genuine, and forms beliefs off-path. To rule out this possibility, the disagreement belief must attribute strictly positive marginal

[59] Agents receive their private information before publicly observing a rejection. But since posterior beliefs do not depend on the order in which public and private information is processed, they can be thought of as resulting from a common disagreement belief that each agent combines with his private information.

probabilities to all types of all acceptors (not necessarily to all type profiles of acceptors): $\mu^i \in \mathcal{B}^i_\varepsilon \equiv \{\mu \in \Delta(\Theta) : \mu_j(\theta_j) > 0, \forall \theta_j \in \Theta_j, \forall j \neq i\}$.

The principal designs the disagreement belief through the trembling device. For the common belief after a spurious rejection by agent $i \in I$ to be $\mu^i \in \Delta(\Theta)$, the probability of spurious rejection by agent i should be $\varepsilon^i(\theta) = \frac{\mu^i(\theta)}{\mu^0(\theta)}\epsilon^i$, with $0 < \epsilon^i < \frac{1}{|I|}\min_{\theta\in\Theta}\mu^0(\theta)$, $\forall i \in I$. Following a spurious rejection, players share a common disagreement belief, $\mu^i \in \mathcal{B}^i_\varepsilon$, which means that the posterior of player $j \in I_P$ of type $\theta_j \in \Theta_j$ is $\mu^i(\cdot|\theta_j) \in \Delta(\Theta)$. This is common knowledge. Following a genuine rejection by agent $i \in I$, the beliefs of acceptors are the same as after a spurious rejection, and this is also common knowledge. In contrast, the genuine rejector maintains his prior belief, $\mu^0(\cdot|\theta_i) \in \Delta(\Theta)$, and this is only known by the rejector himself.

Let $R^i_{BNE(\mu)}$ be the set of mappings $y^{(i)} : \Theta \to \Delta(A)$ that can be induced by a mixed strategy profile $(\sigma^i_{-i}, \sigma^{(i)}_i)$, such that $\sigma^i_{-i} \in BNE_{-i}(\mu)$ and $\sigma^{(i)}_i(\theta_i)$ is a best-response to σ^i_{-i} under the prior belief $\mu^0(\cdot|\theta_i)$, $\forall \theta_i \in \Theta_i$.[60]

Proposition 7. *If an allocation x, for each $i \in I$, satisfies (IC) and (IR) for some $y^{(i)} \in \cup_{\mu\in\mathcal{B}^i_\varepsilon} R^i_{BNE(\mu)}$, then x is virtually tnc-feasible.*

The exact converse is not true because lower hemi-continuity of $BNE(\mu)$ payoffs with respect to μ may fail at the boundary (Einy et al., 2012). As $\mathcal{B}^i_\varepsilon$ is not closed, a sequence of punishments enacted along a sequence of strictly trembling mechanisms may converge to a $BNE_{-i}(\mu)$ with $\mu \notin \mathcal{B}^i_\varepsilon$, which may be a significantly harsher punishment than any $BNE_{-i}(\mu)$ with $\mu \in \mathcal{B}^i_\varepsilon$.

Proposition 8. *If an allocation x is virtually tnc-feasible, then, for each $i \in I$, it satisfies (IC) and (IR) for some $y^{(i)} \in \cup_{\mu\in\Delta(\Theta)} R^i_{BNE(\mu)}$.*

The set of credible punishments is the union of the sets of BNE strategy profiles (restricted to acceptors), $BNE_{-i}(\mu)$, at least across beliefs that satisfy the interiority condition, $\mu \in \mathcal{B}^i_\varepsilon$.

Perhaps the main takeaway is that since the probability of a spurious rejection by one agent can depend on the types announced by all agents, disagreement beliefs do not have to satisfy "*no signaling what you don't know*".[61] With multiple agents, therefore, trembling mechanisms enlarge the set of credible disagreement beliefs (relatively to non-trembling mechanisms).

[60] We say that $z_{-i} \in BNE_{-i}(\mu)$ if and only if there exists $z \in BNE(\mu)$ and $z_i : \Theta_i \to \Delta(A_i)$ such that $z(\theta) = z_{-i}(\theta_{-i})\, z_i(\theta_i)$, $\forall \theta \in \Theta$.

[61] If the probability of spurious rejection by an agent could only depend on his report, the mediator would only be able to influence beliefs about the rejector. Disagreement beliefs would thus satisfy "*no signaling what you don't know*".

5.4 Trembling mechanism with public correlating device

With multiple agents, public signals allow the mediator to condition the disagreement belief and the associated punishment on the information provided by acceptors. However, such conditioning transmits information to the rejector about the type profile of his rivals. There is thus a trade-off between: on the one hand, tailoring the punishment to the characteristics of acceptors; and, on the other hand, concealing from the rejector the punishment and the characteristics of acceptors.

Even if we restrict the public signals to be independent of type profile of acceptors, a public correlating device convexifies the set of credible threats (as in the single-agent case).[62] Without that restriction, an even larger set of punishments can be made credible.[63]

5.5 Trembling mechanism with extraneous correlating device

If the mediator can send an extraneous profile of private signals, the set of credible punishments becomes the set of Uninformed Mediator Bayesian Correlated Equilibria under beliefs that violate the common prior assumption in the way already described. After a genuine rejection: while the common belief of acceptors can be chosen by design, and acceptors believe that it is common to all players, the rejector retains his prior belief and knows the belief of acceptors.[64]

5.6 Trembling mechanism with general correlating device

The set of credible punishments can become larger (and further relax the participation constraints) if the mediator is able to transmit a profile of private signals whose distribution depends on the input message profile. As in the single-agent case, these private signals can be restricted w.l.o.g. to consist of recommendations of actions to choose in the outside game. A correlating device can thus be defined by the distributions over profiles of private signals that follow a spurious and a genuine rejection by agent $i \in I$, denoted $\psi^i : \Theta \to \Delta(A)$ and $\psi^{(i)} : \Theta_{-i} \to \Delta(A)$, respectively.

Suppose that, when the input is $\theta \in \Theta$, the mediator: produces a spurious rejection by agent $i \in I$ with probability $\varepsilon^i(\theta) = \frac{\mu^i(\theta)}{\mu^0(\theta)}\epsilon^i$, where $0 < \epsilon^i < \frac{1}{|I|}\min_{\theta\in\Theta}\mu^0(\theta)$, and μ^i has full support,

[62] Although independent of the type profile, such a public signal is not extraneous because its distribution depends on whether the rejection has been spurious or genuine.

[63] A complete characterization would require the analysis of an extended outside game in which an informed mediator publicly recommends actions, taking into account the fact that a genuine rejector retains his prior belief and makes inferences about the type profile of acceptors from the observation of these public recommendations. Again, it is worth mentioning the result by Lehrer and Sorin (1997), according to which public recommendations become *de facto* private if players can send input messages that are rich enough.

[64] The common disagreement belief of acceptors must satisfy the same interiority condition as in the scenario in which the mediator does not send any additional signals: $\mu^i \in \mathcal{B}^i_\varepsilon$.

and sends a profile of private recommendations, $a \in A$, according to a distribution $\psi^i(\theta)$ such that $\psi^i \in BCE^I(\mu^i)$.[65] Since, after rejection by agent $i \in I$ is observed, the common belief is μ^i (from Bayes' rule), by definition, players always find it optimal to obey the recommendations if and only if, given the common belief μ^i, the mapping $\psi^i : \Theta \to \Delta(A)$ is an Informed Mediator Bayesian Correlated Equilibrium, $BCE^I(\mu^i)$.

After a genuine rejection by agent $i \in I$ (off-path), the set of recommendations that acceptor $j \in I_P \setminus \{i\}$ of type $\theta_j \in \Theta_j$ obeys is the set of recommendations that acceptor j obeys after a spurious rejection (on-path) under the same circumstances: the union across $\theta_{-j} \in \Theta_{-j}$ of the supports of marginal distributions $\psi_j^i(\theta_j, \theta_{-j})$. Off-path, acceptor $j \in I_P \setminus \{i\}$ of type $\theta_j \in \Theta_j$ can thus be induced to obey a recommendation if and only if it is a recommendation that he receives with positive probability in ψ^i. From Lemma 1, there exists $z^i \in BCE^I(\mu^i)$ whose support contains the supports of all BCE$^{\text{I}}$. Let us, therefore: set $\psi^i = z^i$, let $\mathcal{A}_j(\theta_j) \equiv \cup_{\theta_{-j} \in \Theta_{-j}} supp\left[\psi_j^i(\theta_j, \theta_{-j})\right]$, and let $\mathcal{A}_{-i}(\theta_{-i}) \equiv \Pi_{j \in I_P \setminus \{i\}} \mathcal{A}_j(\theta_j)$, $\forall \theta_{-i} \in \Theta_{-i}$.[66] Acceptors can be induced to obey any punishment $y_{-i}^{(i)} : \Theta_{-i} \to \Delta(A_{-i})$ such that $y_{-i}^{(i)}(\theta_{-i}) \in \Delta(\mathcal{A}_{-i}(\theta_{-i}))$.[67] The recommendation made to each acceptor $j \in I_P \setminus \{i\}$ will typically not accord with ψ^i, but each acceptor will believe that it does.

Finally, it is important to understand what should the mediator recommend to a genuine rejector. It is straightforward that, since the private signal sent to a genuine rejector does not affect the incentives of acceptors (because genuine rejections occur with zero probability), an informative signal may benefit – but never harm – the rejector by allowing him to condition his best-response. Hence, a genuine rejector should be sent a non-informative signal.

Proposition 9. *An allocation x is virtually t-feasible if and only if, for each $i \in I$, it satisfies (IC) and (IR) for some $y^{(i)}$ induced by $(y_{-i}^{(i)}, \sigma_i^{(i)})$ such that $y_{-i}^{(i)}(\theta_{-i}) \in \Delta(\mathcal{A}_{-i}(\theta_{-i}))$, $\forall \theta_{-i} \in \Theta_{-i}$, and $\sigma_i^{(i)}(\theta_i)$ is a best-response to $y_{-i}^{(i)}$, $\forall \theta_i \in \Theta_i$.*[68]

[65] From Lemma 1 (in Appendix A.4), the belief μ^i is irrelevant (as long as it has full support) because, given any μ^i with full support, there exists $z^i \in BCE^I(\mu^i)$ whose support contains all the actions played in at least one BCE$^{\text{I}}$. Only the support of the BCE$^{\text{I}}$ played after a spurious rejection is relevant because the recommendations that acceptors obey after a genuine rejection are exactly those in the support of that BCE$^{\text{I}}$. Hence, after a spurious rejection, the mediator should recommend a BCE$^{\text{I}}$ with maximal support, such as z^i. It also follows from Lemma 1 that it is not useful to induce different acceptors to have different beliefs off-path (given that they have a common disagreement belief on-path).

[66] From Lemma 1: $\mathcal{A}_j(\theta_j) = \cup_{\mu^i \in \Delta(\Theta)} \cup_{\phi^i \in BCE^I(\mu^i)} \cup_{(\theta_j, \theta_{-j}) \in supp(\mu^i)} supp\left[\phi_j^i(\theta_j, \theta_{-j})\right]$.

[67] The principal should choose the punishment that maximizes her payoff by relaxing the participation constraints as much as possible. Observe that the punishment that is the most harmful to the rejector may depend on the type of the rejector. However, the choice of punishment cannot be contingent on the rejector's type (which is only known by the rejector himself).

[68] The punishment of a genuine rejector $i \in I$ can be induced if and only if: for each acceptor $j \in I_P \setminus \{i\}$, recommendations made to j with positive probability belong to $\mathcal{A}_j(\theta_j)$. Contrarily to the scenario with no correlating device, lower hemi-continuity of the support of BCE$^{\text{I}}$ with respect to μ holds at the boundary.

6 Examples

6.1 Collusion in oligopoly

6.1.1 Cournot duopoly

Consider a duopoly in which firms simultaneously choose quantities to supply under one-sided private information. Firm A (agent) has private information about its unit cost, $\theta_A \in \left\{0, \frac{1}{3}\right\}$, while the unit cost of firm P (principal) is zero ($\theta_P = 0$). Demand is such that the profit function of firm $i \in \{A, P\}$ is given by $\pi_i = \left(1 - q_i - \frac{1}{2}q_j - \theta_i\right) q_i$, where $\{i, j\} = \{A, P\}$.[69]

Suppose that firm P is able to propose an enforceable "take-it-or-leave-it" collusive agreement involving side-payments.[70] Let $\mu_A \in [0, 1]$ denote the commonly known probability that firm P attributes to $\theta_A = \frac{1}{3}$ in case of disagreement. For each $\mu_A \in [0, 1]$, there is a menu of contracts that maximizes joint-profit and leaves each type of firm A indifferent between accepting or rejecting (truth-telling constraints are not binding).[71]

Due to strategic substitutability, the profit of firm P is strictly increasing in μ_A. A high-cost firm produces less, leaving greater residual demand. Expecting to face a high-cost rival, firm P produces more, which reduces the profit of firm A.

In the sequential equilibrium that firm P prefers, the disagreement belief is $\mu_A = 1$. However, a sequential equilibrium with $\mu_A > \mu_A^0$, where μ_A^0 is the prior belief, is not neologism-proof. The full set of types is a credible veto set: if firm P believes that a rejector may be of either type (and thus keeps its prior belief), both types of firm A have strict incentives to reject if $\mu_A > \mu_A^0$. This means that a rejector would likely be able to convince the principal to change his belief from μ_A to the prior belief whenever $\mu_A > \mu_A^0$.

A mediator with the ability to mimic rejection by firm A allows firm P to approximate its preferred sequential equilibrium, in which $\mu_A = 1$, in a way that survives any restriction of beliefs formed off-path. Firm P should use a mediator to propose the menu of contracts that maximizes joint-profit and leaves firm A indifferent between accepting or rejecting with $\mu_A = 1$.[72] If firm A chooses the contract designed for $\theta_A = 0$, it is enforced; if firm A chooses the contract designed for $\theta_A = \frac{1}{3}$, it is enforced with probability $1 - \epsilon$, where $\epsilon > 0$. Upon choice of the contract designed for $\theta_A = \frac{1}{3}$, the mediator generates a spurious disagreement

[69] This is a version of the model of Singh and Vives (1984) with one-sided private information. We consider a very fine grid of possible output choices, including all the values that are obtained below.

[70] The enforceability assumption is more tenable in legal cartels, such as export cartels (for example, those operating in the U.S. under the Webb-Pomerene Export Trade Act of 1918 or the Export Trading Company Act of 1982) or agriculture cartels (for example, those operating in the U.S. under a Federal Marketing Order).

[71] This is shown in Appendix C.1.

[72] This menu is, precisely: $(q_A, q_P, t) \in \left\{\left(\frac{1}{3}, \frac{1}{3}, -\frac{5}{324}\right), \left(\frac{1}{9}, \frac{4}{9}, \frac{1}{81}\right)\right\}$. See Appendix C.1.

with probability ϵ, and Bayesian updating implies that disagreement beliefs are $\mu_A = 1$.

Message: With non-trembling mechanisms, requiring neologism-proofness reduces the payoff attainable by firm P significantly because rejection can no longer signal inefficiency ($\mu_A \leq \mu_A^0$). With trembling mechanisms, neologism-proofness can be required at a negligible cost.

6.1.2 Cournot triopoly (without correlating device)

Consider a similar model, but with three firms. The unit cost of firm P (principal) is zero ($\theta_P = 0$), while firms A and B (agents) have private information about their unit costs, $\theta_i \in \left\{0, \frac{1}{3}\right\}$, $i \in \{A, B\}$, which are independently and identically distributed. The profit function of firm $i \in \{A, B, P\}$ is given by $\pi_i = \left(1 - q_i - \frac{1}{2}q_j - \frac{1}{2}q_k - \theta_i\right) q_i$, where $\{i, j, k\} = \{A, B, P\}$.

Let the common disagreement belief be that $\theta_i = \frac{1}{3}$ with probability μ_i, for $i \in \{A, B\}$, and that θ_A and θ_B are independent. For each $(\mu_A, \mu_B) \in [0, 1]^2$, there is a menu of contracts that maximizes joint-profit and leaves each type of each firm indifferent between accepting or rejecting (truth-telling constraints are not binding). Again, due to strategic substitutability, the profit of firm P is strictly increasing in μ_A and in μ_B.[73]

Even if we do not require neologism-proofness, the "*no signaling what you don't know*" property of sequential equilibrium restricts beliefs about firm B to remain unchanged ($\mu_B = \mu^0$) if firm A rejects the agreement. The worst possible beliefs for firm A are thus $(\mu_A, \mu_B) = (1, \mu^0)$, i.e., 100% probability that firm A has high cost and unchanged belief about firm B.

A trembling mechanism (without a correlating device) allows the principal to avoid the "*no signaling what you don't know*" restriction and induce, approximately, her preferred disagreement beliefs: $(\mu_A, \mu_B) = (1, 1)$.[74] These are the beliefs that maximize the sum of the outputs of the principal and the acceptor.

The mediator can induce these beliefs by generating spurious rejections with a probability that depends on the profile of reports: if both firms report high cost, the mediator generates a spurious rejection by agent $i \in \{A, B\}$ with probability $\frac{\epsilon}{2(1+\epsilon)}$, with $\epsilon > 0$; if a single firm reports high cost, the mediator generates a spurious rejection by this firm with probability $\frac{\epsilon^2}{2(1+\epsilon)}$. The total probability of spurious rejection is ϵ, and the common disagreement belief after a rejection by firm A is $(\mu_A, \mu_B) = (1, \frac{1}{1+\epsilon})$, which converges to $(1, 1)$ as $\epsilon \to 0$.

Message: With multiple informed firms, trembling mechanisms allow the principal to generate disagreement beliefs that not only survive restrictions of beliefs formed off-path, but can also be significantly more favorable to the principal than any belief that is consistent in the absence of a mediator by allowing a rejection to release optimally designed information about acceptors.

[73] All this is shown in Appendix C.2.

[74] Beliefs could depend on the identity of the rejector, but this is not useful in this example.

6.1.3 Cournot triopoly (with correlating device)

Now suppose that, after a (spurious or genuine) rejection, the mediator can send private recommendations to the firms concerning how much to produce. The ability of the mediator to work as an informed correlating device allows the principal to relax the participation constraints even further.[75] Observe that the mediator is perfectly informed when the rejection is spurious (on-path), because agents have truthfully reported their costs. For simplicity, assume that costs are equiprobable: $\mu^0 = \frac{1}{2}$.

In an equilibrium in which recommendations are obeyed, denote by $\underline{q}_i$ and $\overline{q}_i$ the minimum and maximum output of firms $i \in \{A, B, P\}$. The maximum output of firm $i \in \{A, B, P\}$ cannot be higher than the best-response to the minimum outputs of its rivals when $\theta_i = 0$. That is: $\overline{q}_i \leq \frac{1}{2} - \frac{1}{4}(\underline{q}_j + \underline{q}_k)$, where $\{i, j, k\} = \{A, B, P\}$. Similarly, the minimum output of firm i cannot be lower than the best-response to the maximum outputs of its rivals when firm i has the highest possible cost: $\underline{q}_i \geq \frac{1}{3} - \frac{1}{4}(\overline{q}_j + \overline{q}_P)$, where $\{i, j\} = \{A, B\}$; and $\underline{q}_P \geq \frac{1}{2} - \frac{1}{4}(\overline{q}_A + \overline{q}_B)$.

As a result of these restrictions, the following system of equations yields upper and lower bounds on the minimum and maximum outputs that firms produce in equilibrium:

$$\begin{cases} \overline{\overline{q}}_A = \frac{1}{2} - \frac{1}{4}(\underline{\underline{q}}_B + \underline{\underline{q}}_P) \\ \overline{\overline{q}}_B = \frac{1}{2} - \frac{1}{4}(\underline{\underline{q}}_A + \underline{\underline{q}}_P) \\ \overline{\overline{q}}_P = \frac{1}{2} - \frac{1}{4}(\underline{\underline{q}}_A + \underline{\underline{q}}_B) \\ \underline{\underline{q}}_A = \frac{1}{3} - \frac{1}{4}(\overline{\overline{q}}_B + \overline{\overline{q}}_P) \\ \underline{\underline{q}}_B = \frac{1}{3} - \frac{1}{4}(\overline{\overline{q}}_A + \overline{\overline{q}}_P) \\ \underline{\underline{q}}_P = \frac{1}{2} - \frac{1}{4}(\overline{\overline{q}}_A + \overline{\overline{q}}_B) \end{cases} \Leftrightarrow \begin{cases} \overline{\overline{q}}_A = \overline{\overline{q}}_B = \frac{53}{135} \\ \overline{\overline{q}}_P = \frac{59}{135} \\ \underline{\underline{q}}_A = \underline{\underline{q}}_B = \frac{17}{135} \\ \underline{\underline{q}}_P = \frac{41}{135}. \end{cases}$$

The punishment that can be inflicted on firm A if it genuinely rejects the agreement cannot be harsher than having, with 100% probability, firm P producing $q_P = \frac{59}{135}$ and firm B producing $q_B = \frac{53}{135} - \frac{\theta_B}{2}$. These bounds seem impossible to attain because they are best-responses under beliefs that are extremely asymmetric. However, surprisingly, the mediator can induce obedience of recommendations that approximate these bounds. As a result, if firm A genuinely rejects the agreement, the output from its rivals will be approximately equal to $\overline{\overline{q}}_B - \frac{\theta_B}{2} + \overline{\overline{q}}_P$.

To understand how the existence of a mediator allows these bounds to be approximated, keep in mind that the mediator is informed (we are considering a spurious rejection after both firms have truthfully reported their unit costs). This allows the distribution over recommendations to be contingent on the type profile. Observe also that the recommendations that firms B and P obey after a spurious rejection by firm A (on-path) are also obeyed after a genuine rejection

[75] Such an extended game, in which an informed mediator commits to a distribution over profiles of private recommendations that is conditional on the type profile has been studied by Bergemann and Morris (2016).

by firm A (off-path), because the two kinds of rejection are not distinguishable.

We will construct a stochastic recommendation that is always obeyed after a spurious rejection by firm A, such that the maximum outputs recommended to firms B and P, denoted $\overline{q}_B$ and $\overline{q}_P$, are arbitrarily close to the upper and lower bounds, $\overline{\overline{q}}_B$ and $\overline{\overline{q}}_P$. The mediator will recommend $\overline{q}_B$ and $\overline{q}_P$ with 100% probability after a spurious rejection. The stochastic recommendation to be made after a spurious rejection by firm A is constructed as follows.

1. With probability $1-\epsilon_0$, recommend the full-information Nash equilibrium outputs: $(q_A, q_B, q_P) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ if $(\theta_A, \theta_B) = (0,0)$; $(q_A, q_B, q_P) = (\frac{13}{27}, \frac{13}{27}, \frac{5}{54})$ if $(\theta_A, \theta_B) = (0, \frac{1}{3})$; $(q_A, q_B, q_P) = (\frac{5}{54}, \frac{13}{27}, \frac{13}{27})$ if $(\theta_A, \theta_B) = (\frac{1}{3}, 0)$; $(q_A, q_B, q_P) = (\frac{5}{27}, \frac{5}{27}, \frac{11}{27})$ if $(\theta_A, \theta_B) = (\frac{1}{3}, \frac{1}{3})$. With $\epsilon_0 = 0$, these recommendations would be obeyed, and the (incentive compatibility) obedience constraints would be strictly satisfied.[76] Choose $\epsilon_0 > 0$ that is small enough for the obedience constraints associated with these recommendations to be remain strictly satisfied for any possible recommendations made with probability ϵ_0. This guarantees that these output recommendations are obeyed independently of the remainder of the construction. Let $q^0 \equiv (\underline{q}_A^0, \overline{q}_A^0, \underline{q}_B^0, \overline{q}_B^0, \underline{q}_P^0, \overline{q}_P^0) = (\frac{5}{27}, \frac{13}{27}, \frac{5}{27}, \frac{13}{27}, \frac{1}{3}, \frac{11}{27})$.

2. Construct the sequence $\{q^n\}_{n\in\mathbb{N}}$ as follows. Let: $\underline{q}_i^n$ be the best-response by firm i, when it has the highest possible cost, to $(\overline{q}_j^{n-1}, \overline{q}_k^{n-1})$, where $\{i,j,k\} = \{A,B,P\}$. Explicitly: $\underline{q}_A^n = \frac{1}{3} - \frac{1}{4}(\overline{q}_B^{n-1} + \overline{q}_P^{n-1})$, $\underline{q}_B^n = \frac{1}{3} - \frac{1}{4}(\overline{q}_A^{n-1} + \overline{q}_P^{n-1})$ and $\underline{q}_P^n = \frac{1}{2} - \frac{1}{4}(\overline{q}_A^{n-1} + \overline{q}_B^{n-1})$. Similarly, let $\overline{q}_i^n$ be the best-response by firm i, when it has the lowest possible cost, to $(\underline{q}_j^{n-1}, \underline{q}_k^{n-1})$, where $\{i,j,k\} = \{A,B,P\}$. Explicitly: $\overline{q}_A^n = \frac{1}{2} - \frac{1}{4}(\underline{q}_B^{n-1} + \underline{q}_P^{n-1})$, $\overline{q}_B^n = \frac{1}{2} - \frac{1}{4}(\underline{q}_A^{n-1} + \underline{q}_P^{n-1})$ and $\overline{q}_P^n = \frac{1}{2} - \frac{1}{4}(\underline{q}_A^{n-1} + \underline{q}_B^{n-1})$. This transformation is a contraction with modulus $\frac{1}{2}$. Therefore, it has a single fixed point (the bounds that we wish to approximate), which is the limit of the sequence, and convergence is very fast.[77]

3. With probability given by $(1-\epsilon_n)\Pi_{m=0}^{n-1}\epsilon_m$, recommend: $(\underline{q}_A^n, \overline{q}_B^{n-1}, \overline{q}_P^{n-1})$ or $(\underline{q}_A^{n-1}, \overline{q}_B^n, \underline{q}_P^{n-1})$, equiprobably, if $(\theta_A, \theta_B) = (\frac{1}{3}, 0)$; $(\overline{q}_A^n, \underline{q}_B^{n-1}, \underline{q}_P^{n-1})$ or $(\overline{q}_A^{n-1}, \underline{q}_B^n, \overline{q}_P^{n-1})$, equiprobably, if $(\theta_A, \theta_B) = (0, \frac{1}{3})$; $(\overline{q}_A^{n-1}, \overline{q}_B^{n-1}, \underline{q}_P^n)$, if $(\theta_A, \theta_B) = (0,0)$; $(\underline{q}_A^{n-1}, \underline{q}_B^{n-1}, \overline{q}_P^n)$, if $(\theta_A, \theta_B) = (\frac{1}{3}, \frac{1}{3})$. Choose $\epsilon_n > 0$ sufficiently small for all the obedience constraints to remain strictly satisfied independently of the remainder of the construction (i.e., independently of what is recommended with the remaining probability, $\Pi_{m=0}^n \epsilon_m$). Recommendations $\underline{q}_i^{n-1}$ and $\overline{q}_i^{n-1}$, not being best-responses in this case, are obeyed because the obedience constraints are pooled with the much more likely cases (previous iterations) in which they are best-responses. Recommendations $\underline{q}_i^n$ and $\overline{q}_i^n$ are obeyed because they are best-responses, and the obedience constraints are strictly satisfied.

With a sufficient number of iterations, the support of recommendations made after a spurious rejection (on-path) approximates the bounds as closely as desired. Recommendations made

[76] We are considering a very fine, but finite, grid of possible output choices.

[77] This is shown in Appendix .

to firms B and P after a genuine rejection by firm A are deterministic, being given by the maximum outputs recommended on-path: $(\overline{q}_B^N - \frac{\theta_B}{2}, \overline{q}_P^N)$, where N is the number of iterations.[78]

Message: Trembling mechanisms with a correlating device can originate significantly harsher punishments than the harshest punishment attainable without correlating signals, by making each acceptor j believe that: all rivals $k \neq j$ have high costs; all rivals $k \neq j$ believe that all rivals $l \neq k$ have low costs; all rivals $k \neq j$ believe that all rivals $l \neq k$ believe that all rivals $m \neq l$ have high costs; and so on.

6.2 Collusion in auctions

6.2.1 All-pay auction with bid cap

Two lobbyists, $i \in \{A, P\}$, compete for a government contract by simultaneously choosing how much to spend, $b_i \in [0, m]$, where m is the cap on spending.[79] The highest spender gets the contract, worth $v_P > 0$ to lobbyist P (principal) and $v_A > v_P$ to lobbyist A (agent). In case of a tie, each wins with 50% probability. Supposing that the cap on bids has an intermediate value, $m \in \left(\frac{v_P}{2}, v_P\right)$, non-cooperative bids are as follows: lobbyist A randomizes over $(0, 2m - v_P] \cup \{m\}$, with a mass point at m; while lobbyist P randomizes over $[0, 2m - v_P] \cup \{m\}$, with mass points at 0 and m. Their expected payoffs are $\pi_A = v_A - v_P$ and $\pi_P = 0$, respectively.[80]

Suppose that the principal proposes a collusive agreement to the agent, offering to withdraw from the contest in exchange for a payment of $v^P - \delta$, where $\delta > 0$. If the agent rejects, non-cooperative bidding ensues. The agent strictly gains from the agreement, as his payoff increases to $v_A - v_P + \delta$.

If the principal has access to a mediator that can mimic the rejection of a proposal, she can obtain a higher payment. The mediator should generate a spurious rejection with probability $\epsilon > 0$, and, after a spurious rejection, recommend a bid to the principal according to the non-cooperative equilibrium distribution over $[0, 2m - v_P] \cup \{m\}$. If the agent genuinely rejects the agreement, the mediator should recommend a bid equal to m with 100% probability.

Without a trembling mechanism, the principal would not obey such a recommendation. After observing a rejection, the principal would know that the randomization is not proper and would thus perform the randomization herself. With a trembling mechanism, the principal, after observing a rejection, believes with 100% probability that it is spurious, and thus believes that the randomization is proper. It is the strictly positive probability of spurious rejection

[78] As in the previous scenarios, it is straightforward to verify that the truth-telling constraints are satisfied.

[79] Since our theory of trembling mechanisms assumed finite action sets, the conclusions of this example hold only asymptotically. It is also worthwhile remarking that, in this example, there is no private information.

[80] This model was proposed and analyzed by Che and Gale (1998b).

that makes the principal obey the faulty randomization, by concealing that it is faulty.

We conclude that, using the mediator as a randomization device, the principal can credibly threaten to bid m in case of a genuine rejection. Facing such a threat: if $m > \frac{v_A}{2}$, the agent withdraws and has zero payoff; if $m < \frac{v_A}{2}$, the agent bids at the cap and gets a payoff of $\frac{v_A}{2} - m$. The disagreement payoff of the agent is reduced to $max\left\{0, \frac{v_A}{2} - m\right\}$, thus the agent becomes willing to pay the principal $v_A - max\left\{0, \frac{v_A}{2} - m\right\} - \delta$, which is more than $v^P - \delta$.

Message: Even without private information, trembling mechanisms can be useful by inducing the principal to always make the highest bid in the support of a Nash equilibrium of the game.

6.2.2 Second-price auction with participation costs

In a second-price auction with participation costs, trembling mechanisms may be crucial for a cartel to be established. Without mediation, as shown by Tan and Yilankaya (2007), a cartel agreement is not ratifiable. Rejecting the agreement credibly signals a high valuation, which makes it unprofitable for the rivals to support the participation cost. As a result, a high-value bidder is better off rejecting the collusive agreement than ratifying it. However, the same agreement, proposed using a trembling mechanism, would be accepted by all agents.[81]

Message: Non-ratifiability of a collusive agreement is not an issue if trembling mechanisms are used. Since disagreement beliefs are designed to inflict an informational loss on rejectors, agents are at least willing to accept a proposal that gives them what would be their non-cooperative payoffs in the absence of a proposal.

6.3 Other possible applications

6.3.1 Dispute resolution

Rejection of a settlement can release information that is relevant for a future decision about whether to litigate and about what effort to make in the litigation process.[82] Therefore, it may be important to generate the disagreement beliefs that are the most adverse for a rejector.[83]

In a contemporaneous study on the design of alternative dispute resolution schemes, Balzer and Schneider (2016) achieve this using a kind of trembling mechanism. In their model, a mediator who wishes to maximize the *ex ante* probability of settlement proposes a mechanism that specifies, as a function of litigants' reported types, a division of a unit of surplus and a probability of reversion to an all-pay contest. They showed that imposing *ex post* individual

[81] Without participation costs, beliefs are irrelevant and thus trembling mechanisms would not be useful.

[82] See Cooter and Rubinfeld (1989).

[83] In Hörner et al. (2015), a mediator can recommend war, with some of the flavor of a spurious rejection.

rationality in addition to *interim* individual rationality has a negligible impact on the probability of settlement if the mediator is able to communicate privately with each litigant. If this is possible, the mediator should, with a small probability: privately make an unacceptable proposal to one of the litigants, generating a kind of spurious rejection (on-path); and truthfully disclose the type profile of the litigants before the contest. As a result, if a litigant rejects an acceptable settlement (off-path), the rival will presume that the rejection was caused by the mediator having proposed an unacceptable settlement and will believe the mediator when it announces that the rejector has a low cost of collecting evidence.

6.3.2 Sequential contracting

When two principals contract sequentially with the same agent, the first principal may extract additional surplus from the agent by manipulating the otherwise off-path beliefs of the second principal.[84] For example, consider an agent with private information about his valuation for a good that is supplied by two principals, in an environment where the first principal proposes a "take-it-or-leave-it" menu of contracts to the agent, and, if the agent rejects, the second principal has the opportunity to contract with the agent. Employing a mediator that, with a small probability, generates disagreement if the agent picks the contract designed for an agent with high valuation, the first principal reduces the outside payoff of the agent by inducing in the second principal the belief that an agent that rejects the proposal of the first principal must have a high valuation.

6.3.3 Dynamic contracting

In a two-period dynamic screening model with limited commitment, Deb and Said (2015) showed that a seller can gain from inducing rejections in the first period to incentivize herself to charge higher prices in the second period. They constrained rejections to be genuine (the proposal must be designed in a way that makes it optimal for some types to reject), while trembling mechanisms dispense with restrictions on the utility offered to the types that the principal wants to reject.[85]

[84] See Calzolari and Pavan (2006a,b, 2008), and Pavan and Calzolari (2009).

[85] In the model of Deb and Said (2015), some buyers arrive early while others arrive at the last-minute. Those who arrive early can wait for the last-minute to contract. When contracting early, the seller is not able to commit to the contract that she will offer at the last-minute. Therefore, the last-minute proposal is an endogenous outside option for the early proposal. Rejections of the early proposal change the composition of demand at the last-minute and, therefore, may mitigate the seller's loss due to limited commitment. Deb and Said (2015) conclude that the seller should distort the early contract in a way that induces rejection by an intermediate set of types.

6.3.4 Informed principal

Although outside our scope, the use of a mediator can also allow an informed principal to manipulate off-path beliefs about her own private information. If the informed principal is able to transmit her proposal through a communication device that, with a small probability, modifies the proposal, an agent that receives a proposal which is different from the one that he should have received according to the candidate equilibrium will update his beliefs about the principal in a Bayesian way (amenable of design through the communication device).[86]

For example, in models of multilateral vertical contracting where a supplier is constrained to make private bilateral offers to multiple downstream firms, belief updating off-path by a downstream firm occurs if it receives an offer that is different from the one prescribed by the candidate equilibrium. It is common to assume *passive beliefs*: continue to expect the proposals made to the other downstream firms to be in accordance with the candidate equilibrium (Hart and Tirole, 1990; Segal, 1999). However, an alternative known as *wary beliefs* has been put forward: expect the proposals made to the other downstream firms to be optimal for the supplier, given the proposal just received (McAfee and Schwartz, 1994; Rey and Vergé, 2004).

Suppose that a supplier can use a mediation device that transforms her proposal into another proposal with a small probability. Then, if a downstream firm receives an unexpected proposal, it may infer that is was due to a tremble by the mediation device and will not update its beliefs about the proposals that the supplier has sent to the other downstream firms. In this limited sense, trembling mechanisms provide a foundation for assuming passive beliefs.

7 Conclusion

In principal-agent environments where the outside payoff of an agent depends on what others infer from his rejection, the principal can relax participation constraints using trembling mechanisms. These are mediated stochastic mechanisms. Stochastic because the agreement may not be enforced even if all agents accept to participate (with a small probability, there is a spurious rejection). Mediated because participation decisions are not observed by others (who thus cannot distinguish a genuine rejection from a spurious one). Besides signal-jamming genuine rejections, the mediator collects information and makes non-binding recommendations. These are trustworthy on-path, and thus trusted off-path when they induce acceptors to punish a genuine rejector. Understanding this will hopefully be useful for the design of organizations under the threat of collusion, and for the design of mediation between privately informed parties.

[86] In an example presented by Rabin and Sobel (1996), the informed agent, who moves first, has a higher payoff in a pooling equilibrium which does not satisfy reasonable refinements than in a partially separating equilibrium that satisfies these refinements. The informed agent would thus gain from using a trembling mechanism to manipulate the otherwise off-path beliefs of the uninformed agent.

A Some remarks on correlated equilibrium

A.1 BCE$^{\mathrm{U}}$ vs BNE

An outside game in which a BCE$^{\mathrm{U}}$ can entail a harsher punishment than any BNE is described in Figure 1.[87] Despite being a complete information game, the example illustrates the point.

	C1	C2	C3	C4
R1	0 0	3 6	6 3	4 4
R2	6 3	0 0	3 6	4 4
R3	3 6	6 3	0 0	4 4
R4	4 4	4 4	4 4	5 5

Figure 1: A correlated equilibrium may be a harsher punishment than any Nash equilibrium.

In this degenerate outside game, a BCE$^{\mathrm{U}}$ is simply a correlated equilibrium, while a BNE is simply a Nash equilibrium. While the unique Nash equilibrium payoffs are $(5, 5)$, there is a correlated equilibrium in which expected payoffs are $(5, 4)$.[88]

A.2 BCE$^{\mathrm{I}}$ vs BCE$^{\mathrm{U}}$

It is easy to find examples in which a $BCE^I(\mu)$ strategy can punish the agent more than any $BCE^U(\mu)$ strategy. Figure 2 presents one where the agent (column player) does not play.

The principal plays $R3$ in the single BCE$^{\mathrm{U}}$ (which is the single BNE), yielding a payoff of 4 to the agent (column player). The BCE$^{\mathrm{I}}$ in which the principal plays $R1$ if the agent is of type 1 and $R2$ if the agent is of type 2 yields a payoff of 0 to the agent.

Notice that the principal cannot implement such a harsh punishment using trembling mechanisms, because, after a genuine rejection, the mediator cannot make a recommendation that is contingent on the type of the agent. Still, the mediator can recommend $R1$ with probability $\alpha \in [0, 1]$ and $R2$ with probability $1 - \alpha$, yielding an outside payoff of $3(1 - \alpha)$ to the agent of type 1 and an outside payoff of 3α to the agent of type 2.

[87] This example is attributed to Robert J. Aumann. Versions of it appeared in the works of Nau and McCardle (1990) and Evangelista and Raghavan (1996), from where this version was adapted.

[88] Observe that a profile of correlated equilibrium strategies that places probability $\frac{2}{9}$ on each outcome whose payoffs are $(6, 3)$, i.e., on $(R1, C2)$, $(R2, C3)$ and $(R3, C1)$, and probability $\frac{1}{9}$ on each outcome whose payoffs are $(3, 6)$, i.e., on $(R1, C3)$, $(R2, C1)$ and $(R3, C2)$, yields expected payoffs of $(5, 4)$.

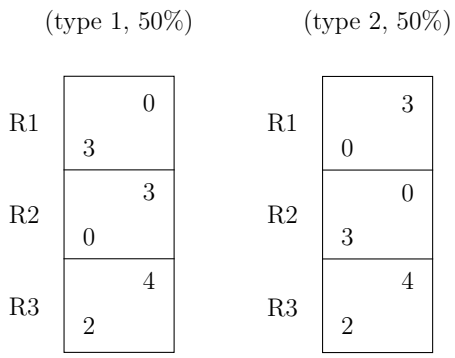

Figure 2: A BCE^I may be a harsher punishment than any BCE^U.

A.3 Choosing an action in the support

After a spurious rejection, the recommendations of the trembling device must constitute a $BCE^I(\mu)$ strategy profile. After a genuine rejection, the principal will obey any action in the support of a $BCE^I_P(\mu)$ strategy.

A strategy with support contained in that of an equilibrium strategy can punish the agent more than any equilibrium strategy. For example, in the matching pennies game in Figure 3: while the payoff of the column player in the unique BCE$^{\text{I}}$ is -4 (the unique BCE$^{\text{I}}$ is the unique Nash equilibrium), the pure strategy $R1$ reduces the payoff of the column player to -7.

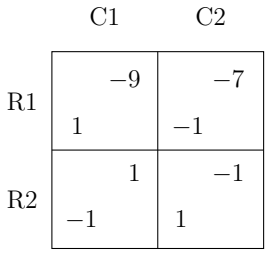

Figure 3: A strategy in $\Delta(supp(BCE^I))$ may punish more than any BCE^I strategy.

A.4 On the structure of BCE$^{\text{I}}$

There exists a single BCE$^{\text{I}}$ whose support is equal to the union of the supports of all BCE$^{\text{I}}$ for all common beliefs. On-path, the mediator could randomize over beliefs and over different BCE$^{\text{I}}$, but this would not enlarge the set of recommendations obeyed off-path (which is the support of the distribution over recommendations made on-path) relatively to that single BCE$^{\text{I}}$.

Lemma 1. *Let $\mu \in \Delta(\Theta)$ have full support. There exists $y \in BCE^I(\mu)$ such that, for all $i \in I_P$: $supp(y_i|\mu) = supp(BCE^I_i)$.*[89]

[89]Recall that $supp(BCE^I_i) \equiv \cup_{\mu\in\Delta(\Theta)} \cup_{y\in BCE^I(\mu)} supp(y_P|\mu)$.

Proof. Given a common belief with full support, $\mu \in \Delta(\Theta)$, let us construct $y \in BCE^I(\mu)$ whose support contains all actions $a_j \in A_j$, for all $j \in I_P$, that are played with strictly positive probability in some BCE$^{\text{I}}$. Collect all such a_j, for all $j \in I_P$, and denote this set by $\mathcal{A}$. For each action $a^k \in \mathcal{A}$, there exists a common belief, μ^k, and a BCE$^{\text{I}}$ under that belief, $y^k \in BCE^I(\mu^k)$, in which a^k is played with strictly positive probability. There exists a set of signals that transforms μ into a distribution over posteriors whose support contains $\left\{\mu^1, ..., \mu^K\right\}$, where K is the cardinality of $\mathcal{A}$. Concretely, consider a random signal with possible realizations $\left\{s^1, ..., s^K\right\}$ where the probability of s^k if the type profile is θ is given by $\frac{\mu^k(\theta)}{\mu(\theta)}\epsilon^k$, with $\epsilon^k > 0$, for $k \in K$. Imposing $\sum_k \epsilon^k < \min_\theta \mu(\theta)$ guarantees that the conditional probability of some signal being released is smaller than unity. The posterior probability that results from the observation of s^k is μ^k, for $k \in K$. Conditionally on signal s^k being released, the mediator recommends $y^k \in BCE^I(\mu^k)$. As a result, all actions $a^k \in \mathcal{A}$ are played with strictly positive probability. □

B Proofs

B.1 Single-agent case

Proof of Proposition 1. ($\Rightarrow$) If x is virtually *tnc*-feasible, there exists a sequence of strictly trembling mechanisms, $(\varepsilon^m, 0, x^m)_{m\in\mathbb{N}}$, converging to $(0, 0, x)$, each having a sequential equilibrium in which the agent participates and reports truthfully. Consider a sequence of such equilibria, which are characterized by: an *ex ante* probability of spurious rejection, ϵ^m; a disagreement belief, μ^m; and a type-dependent distribution over action profiles, $y^m \in BNE(\mu^m)$, induced by a strategy profile, (σ_P^m, σ_1^m), where $\sigma_1^m : \Theta \to \Delta(A_1)$ is such that $\sigma_1^m(\theta)$ is a best-response to σ_P^m if the agent has type θ, $\forall\theta \in \Theta$, while $\sigma_P^m \in \Delta(A_P)$ is a best-response to σ_1^m if the principal believes that the agent is of type $\theta \in \Theta$ with probability $\mu(\theta)$. Consider a subsequence, $(\epsilon^m, \mu^m, y^m, \sigma_P^m, \sigma_1^m)_{m\in M}$, that converges to $(0, \mu, y, \sigma_P, \sigma_1)$. The limit strategy profile, (σ_P, σ_1), induces the limit distribution over action profiles, y. Since $\pi_P^R(y(\theta), \theta)$ and $\pi_1^R(y(\theta), \theta)$ are continuous with respect to y, the limit strategies are mutual best-responses, thus, $y \in BNE(\mu)$. Since (IC') and (IR) are satisfied along the sequence, x satisfies (IC) and (IR) when the outside option is y.

($\Leftarrow$) From Remark 1, if an allocation x is 0-feasible, there exists a belief, $\mu \in \Delta(\Theta)$, and a type-dependent distribution over action profiles, $y \in BNE(\mu)$, such that x satisfies (IC) and (IR). Given Assumption 1 and linearity of $\pi_1(x(\theta), \theta)$ in the first variable, we can construct a sequence of allocations, $(x^m)_{m\in\mathbb{N}}$, converging to x, that strictly satisfy (IC), and satisfy (IR) given the same outside option, y. For example, letting $x^m \equiv \frac{m}{m+1}x + \frac{1}{m+1}x^f$, $\forall m \in \mathbb{N}$. Each allocation in the sequence, x^m, satisfies (IC') and (IR) for sufficiently small $\epsilon^m > 0$ (since π_1^R is bounded, v_1 is also bounded, thus ϵ^m can be chosen irrespectively of the behavior in the outside game after a spurious rejection). Therefore, we can construct a sequence of pairs $(x^m, \epsilon^m)_{m\in\mathbb{N}}$, converging to $(x, 0)$, such that (IC') and (IR) are satisfied along the sequence. The sequence of strictly trembling mechanisms, $(\varepsilon^m, 0, x^m)_{m\in\mathbb{N}}$,

becomes completely defined by letting $\varepsilon^m(\theta) = \frac{\mu(\theta)}{\mu^0(\theta)}\epsilon^m$, which implies that the disagreement belief is μ along the sequence. As a result, following a genuine rejection, the independent strategies that induce $y \in BNE(\mu)$ are sequentially rational. Hence, each mechanism in the sequence has a sequential equilibrium in which the agent participates and reports truthfully. As $\lim_{m\to\infty} \epsilon^m = 0$, the sequence of mechanisms converges to the non-mediated mechanism, $(0, x)$. □

Proof of Proposition 2. ($\Rightarrow$) Let x be virtually *tpc*-feasible. There exists a sequence of strictly trembling mechanisms with public correlation, $(\varepsilon^m, \psi^{pc,m}, x^m)_{m\in\mathbb{N}}$, converging to $(0, \psi^{pc}, x)$, each having a sequential equilibrium in which the agent participates and reports truthfully. In mechanism $(\varepsilon^m, \psi^{pc,m}, x^m)_{m\in\mathbb{N}}$, if the agent reports type $\theta \in \Theta$, a spurious rejection is produced and a public signal, $k \in K$, is sent with probability $\varepsilon_k^m(\theta) = \frac{\mu_k^m(\theta)}{\mu^0(\theta)}\epsilon_k^m$. Observation of $k \in K$ induces the disagreement belief $\mu_k^m \in \Delta(\Theta)$, and the distribution over action profiles, $y_k^m \in BNE(\mu_k^m)$. If there is a genuine rejection (off-path), signal $k \in K$, such that $\epsilon_k^m > 0$ (so that the principal cannot detect that the rejection has been genuine), is sent with probability p_k^m. This implies that the same belief, μ_k^m, and the same behavior, $y_k^m \in BNE(\mu_k^m)$, is induced. The principal behaves in the same way because she does not distinguish a genuine rejection from a spurious rejection. The agent could behave differently, but is not interested in deviating because he is best-responding to the behavior of the principal. Let $y^m \in Conv(\cup_{\mu\in\Delta(\Theta)} BNE(\mu))$ be the mixture over $\{y_k^m\}_{k\in K}$ induced by probabilities p_k^m. Each (x^m, ϵ^m) satisfies (IC') and (IR) with y^m as the outside option. Taking limits, as $(x^m, \epsilon^m, y^m) \to (x, 0, y)$, we find that x satisfies (IC) and (IR), with y as the outside option. Continuity of $\pi_P^R(\cdot, \theta)$ and $\pi_1^R(\cdot, \theta)$ implies that $y \in Conv(\cup_{\mu\in\Delta(\Theta)} BNE(\mu))$.

($\Leftarrow$) Let y be a mixture over a set of distributions over action profiles, $\{y_k\}_{k\in K}$, where $y_k \in BNE(\mu_k)$, with probability weights p_k (from Caratheodory's theorem, K can be assumed to be finite). As explained in the proof of Proposition 1, construct a sequence of allocations, $(x^m)_{m\in\mathbb{N}}$, that satisfy (IR) and strictly satisfy (IC), whose limit is x. Let the associated sequence of strictly trembling mechanisms, $(\varepsilon^m, \psi^{pc,m}, x^m)_{m\in\mathbb{N}}$, be such that, following report $\theta \in \Theta$, a spurious rejection is generated and signal $k \in K$ is released with probability $\varepsilon_k(\theta) = \frac{\mu_k(\theta)}{\mu^0(\theta)} p_k \epsilon^m$, where $\epsilon^m > 0$ is sufficiently small for (IC) to remain strictly satisfied. After a genuine rejection (off-path), signal $k \in K$ is released with probability p_k. The resulting belief is μ_k, and the resulting behavior is $y_k \in BNE(\mu_k)$. Each of the mechanisms in the sequence has, therefore, a sequential equilibrium in which: the agent participates and reports truthfully; a spurious rejection is generated with probability ϵ^m; and the outside option is y. As $\epsilon^m \to 0$, the sequence of strictly trembling mechanisms converges to a non-trembling mechanism, $(0, \psi^{pc}, x)$. □

Proof of Proposition 3. ($\Rightarrow$) If x is virtually *tec*-feasible, there exists a sequence of strictly trembling mechanisms, $(\varepsilon^m, \psi^{ec,m}, x^m)_{m\in\mathbb{N}}$, converging to $(0, \psi^{ec}, x)$, each having a sequential equilibrium in which the agent participates and reports truthfully. Such an equilibrium is characterized by: a probability of spurious rejection, ϵ^m; a disagreement belief, μ^m; and a joint distribution over actions for each type of the agent, $y^m \in BCE^U(\mu^m)$. Consider a subsequence, (ϵ^m, μ^m, y^m), that converges to

$(0, \mu, y)$. Since $\pi_P^R(y(\theta), \theta)$ and $\pi_1^R(y(\theta), \theta)$ are continuous with respect to y, optimality of obeying the extraneous private recommendations along the sequence implies optimality of obeying the extraneous private recommendations in the limit: $y \in BCE^U(\mu)$. Since (IC') and (IR) are satisfied along the sequence, they are also satisfied in the limit. Hence, x satisfies (IC) and (IR) with y as outside option.

($\Leftarrow$) From Remark 2, if an allocation x is nt-feasible, there exists a belief, $\mu \in \Delta(\Theta)$, and a distribution over action profiles in the outside game, $y \in BCE^U(\mu)$, such that (IC) and (IR) are satisfied. The fact that $y \in BCE^U(\mu)$ means that there exists a distribution over profiles of private signals, $\psi^{ec} \in \Delta(A_P \times A_1^{|\Theta|})$, recommending an action for the principal to execute and a plan of actions for the agent to execute as a function of his type, which is such that $y(\theta, a_P, a_1) = \psi^{ec}(a_P, a_1(\theta))$, $\forall(\theta, a_P, a_1) \in \Theta \times A_P \times A_1$, and such that neither the principal nor the agent gain from disobeying. Notice that ψ^{ec} must be the same after a spurious and a genuine rejection (because it is extraneous). Assumption 1 and linearity of $\pi_1(\cdot, \theta)$ allow us to construct a sequence of allocations, $(x^m)_{m \in \mathbb{N}}$, converging to x, that strictly satisfy (IC), and satisfy (IR) when the outside option is y. Let $\varepsilon^m(\theta) = \frac{\mu(\theta)}{\mu^0(\theta)}\epsilon^m$ be the probability of a spurious rejection following report $\theta \in \Theta$, which implies that the disagreement belief is μ. Since each x^m satisfies (IC') and (IR) as long as $\epsilon^m > 0$ is sufficiently small, we can construct a sequence $(x^m, \epsilon^m)_{m \in \mathbb{N}}$, converging to $(x, 0)$, such that (IC') and (IR) are satisfied along the sequence. All mechanisms in the sequence send extraneous profiles of private signals according to ψ^{ec}. Thus, the outside option is y along the sequence. We conclude that each mechanism in the sequence has a sequential equilibrium in which the agent participates and reports truthfully. As $\epsilon^m \to 0$, the sequence converges to a non-trembling mechanism, $(0, \psi^{ec}, x)$. □

Proof of Proposition 4. ($\Rightarrow$) Let $(\varepsilon^m, \psi^m, x^m)_{m \in \mathbb{N}}$ be a sequence of strictly trembling mechanisms converging to a non-trembling mechanism, $(0, \psi, x)$, such that each mechanism in the sequence has a sequential equilibrium in which the agent participates and reports truthfully. Consider mechanism $m \in \mathbb{N}$ in the sequence. After a spurious rejection, common disagreement beliefs are μ^m and principal and agent play some $z^m \in BCE^I(\mu^m)$. After a genuine rejection (off-path): the principal is recommended an action in $supp(z_P^m|\mu^m)$ (she obeys, as after a spurious rejection); the agent best-responds to the distribution of recommendations made to the principal. Let $\sigma_P^m \in \Delta(supp(z_P^m|\mu^m))$ be the mixed action that the principal plays off-path, following a genuine rejection; and let $\sigma_1^m : \Theta \to \Delta(A_1)$ be a best-response to σ_P^m. Denote by $y^m : \Theta \to \Delta(A)$ the resulting joint distribution over action profiles. Considering a subsequence that converges, denote the limit of σ_P^m by $\sigma_P \in \Delta(supp(BCE_P^I))$, and let $\sigma_1(\theta)$ be a best-response to σ_P when the agent is of type θ. Since there is participation and truth-telling along the sequence, x satisfies (IC) and (IR) when rejection leads to (σ_P, σ_1).

($\Leftarrow$) Let σ_P consist in playing $a^k \in supp(BCE_P^I)$ with probability p_k. Consider a common disagreement belief with full support, $\mu \in \Delta(\Theta)$, and a BCE$^{\text{I}}$ distribution over action profiles, $y \in BCE^I(\mu)$, whose support contains the support of σ_P. Existence of y follows from Lemma 1. Construct the trembling device of each mechanism $m \in \mathbb{N}$ so that if the input message is θ, a spurious rejection is generated with probability $\varepsilon(\theta) \equiv \frac{\mu(\theta)}{\mu^0(\theta)}\epsilon^m$, where $\epsilon^m > 0$. Along the sequence, the common disagreement belief is μ. Instruct the mediator to recommend $y \in BCE^I(\mu)$ after a spurious rejection (in

all mechanisms in the sequence). Following a genuine rejection, the mediator recommends a^k with probability p_k to the principal, and sends no meaningful signal to the agent. Construct $(x^m, \epsilon^m)_{m\in\mathbb{N}}$ as in the previous proofs, so that (IC') and (IR) are satisfied with y induced by $(\sigma_P, \sigma_1(\theta))$, where σ_1 is a best-response to σ_P. Consider a candidate sequential equilibrium in which the agent participates and reports truthfully. It is optimal for principal and agent to obey the recommendations made after a spurious rejection. Therefore, the principal also obeys the recommendation following a genuine rejection (believing that the rejection has been spurious, which is infinitely more likely). After a genuine rejection, the agent anticipates that the principal will play σ_P, and plays a best-response, $\sigma_1(\theta)$. This means that the distribution over action profiles is induced by $(y_P, y_1(\theta))$, $\forall\theta \in \Theta$. Since (x^m, ϵ^m) satisfies (IC') and (IR), no deviation from the candidate sequential equilibrium is beneficial. Letting, $\epsilon^m \to 0$, the mechanism converges to a non-trembling mechanism, $(0, \psi, x)$. □

B.2 Multiple-agent case

Proof of Proposition 5. ($\Rightarrow$) If x is 0-feasible, there is a sequential equilibrium of the non-mediated mechanism, $(0, 0, x)$, in which agents participate and report their types truthfully. In this sequential equilibrium, the common belief following a rejection by agent $i \in I$ (off-path), denoted μ^i, belongs to $\mathcal{B}_0^i$; and the resulting distribution over action profiles, y^i, belongs to $BNE(\mu^i)$. Since agents participate and report their types truthfully, (IC) and (IR) are satisfied for each $i \in I$, given each y^i.

($\Leftarrow$) Since, for each $i \in I$, $\mu^i \in \mathcal{B}_0^i$ and $y^i \in BNE(\mu^i)$, we can construct a sequential equilibrium of the non-mediated mechanism, $(0, 0, x)$, in which agents participate and report their types truthfully. For each $i \in I$, let μ^i be the common belief following a rejection by agent $i \in I$ (off-path), and let y^i be the resulting distribution over action profiles. □

Proof of Proposition 6. ($\Rightarrow$) Since x is *nt*-feasible, there exists a sequential equilibrium of a non-trembling mechanism, $(0, \psi, x)$, in which agents participate and report their types truthfully. The common belief following a rejection by agent $i \in I$ (off-path), denoted μ^i, belongs to $\mathcal{B}_0^i$ (remember that under $\mu^i \in \mathcal{B}_0^i$, the belief of agent i of type $\theta_i \in \Theta_i$ is the same as under μ^0); and the continuation distribution over action profiles, y^i, constitutes a $BCE^{-i}(\mu^i)$, by definition. Since agents participate and report truthfully, (IC) and (IR) are satisfied for each $i \in I$, given y^i.

($\Leftarrow$) Let $\mu^i \in \mathcal{B}_0^i$ and $y^i \in BCE^{-i}(\mu^i)$ be such that x satisfies (IC) and (IR), $\forall i \in I$. We will construct a sequential equilibrium of a non-trembling mechanism, $(0, \psi, x)$, in which agents participate and report their types truthfully. Let a pure strategy by player $i \in I$ be a mapping $\tilde{a}_i : \Theta_i \to A_i$, and denote the finite set of such pure strategies by $\tilde{A}_i$. For each $i \in I$, let μ^i be the common belief following a genuine rejection by agent i, and let $y^i : \Theta_{-i} \to \Delta(A_{-i} \times \tilde{A}_i)$ be the subsequent distribution over profiles of private recommendations (agent i receives a recommendation for each of his possible types). The recommendations are obeyed because $y^i \in BCE^{-i}(\mu^i)$, by definition. Since, $\forall i \in I$, (IC)

and (IR) are satisfied with y^i being the outside option, there is a sequential equilibrium of $(0, \psi, x)$ in which agents participate and report their types truthfully. □

Proof of Proposition 7. Using Assumption 1' and linearity of $\pi_i(x(\theta), \theta)$ in the first variable, we can construct a sequence of allocations, $(x^m)_{m\in\mathbb{N}}$, converging to x, that strictly satisfy (IC) and satisfy (IR), for each $i \in I$, given the outside option, $y^{(i)}$. We know that $y^{(i)}$ is induced by $(\sigma^i_{-i}, \sigma^{(i)}_i)$ such that $\sigma^i_{-i} \in BNE_{-i}(\mu^i)$, with $\mu^i \in \mathcal{B}^i_\varepsilon$. Construct a sequence of trembling devices by setting $\varepsilon^{i,m}(\theta) = \frac{\mu^i(\theta)}{\mu^0(\theta)}\epsilon^m$, for each $i \in I$, with $\lim_{m\to\infty} \epsilon^m = 0$. The only difference across trembling devices along the sequence is the value of $\epsilon^m > 0$. For each $m \in \mathbb{N}$, $\epsilon^m > 0$ must be sufficiently small for (IC') and (IR) to be satisfied, $\forall i \in I$, given the outside option $y^{(i)}$. This completes the construction of a sequence of strictly trembling mechanisms, $(\varepsilon^m, 0, x^m)_{m\in\mathbb{N}}$, that converges to $(0, 0, x)$. Consider a candidate sequential equilibrium in which: agents participate truthfully; a spurious rejection by agent $i \in I$ leads to a mixed action profile $(\sigma^i_{-i}, \sigma^i_i) \in BNE(\mu^i)$; and a genuine rejection by agent i leads to the same mixed action profile by the acceptors, $\sigma^i_{-i} \in BNE_{-i}(\mu^i)$, and a best-response by the rejector, $\sigma^{(i)}_i$, which induces $y^{(i)}$. Since (IC') and (IR) are satisfied along the sequence, this is a sequential equilibrium. □

Proof of Proposition 8. Since x is virtually *tnc*-feasible, there exists a sequence of strictly trembling mechanisms, $(\varepsilon^m, 0, x^m)_{m\in\mathbb{N}}$, converging to $(0, 0, x)$, each having a sequential equilibrium in which agents participate and report truthfully. Each equilibrium in the sequence is characterized by: the probability of spurious rejection by each agent $i \in I$, denoted $\varepsilon^{i,m} : \Theta \to [0, 1]$; the resulting disagreement belief, $\mu^{i,m}$; the strategy of each acceptor $j \in I_P \setminus \{i\}$, following a rejection by agent $i \in I$, denoted $\sigma^{i,m}_j : \Theta_j \to \Delta(A_j)$; the strategy of a spurious rejector $i \in I$, denoted $\sigma^{i,m}_i : \Theta_i \to \Delta(A_i)$; and the strategy of a genuine rejector $i \in I$, denoted $\sigma^{(i),m}_i : \Theta_i \to \Delta(A_i)$. Since the mechanisms in the sequence are strictly trembling, all types of all players play the outside game on-path. Notice also that $\sigma^{i,m} \equiv (\sigma^{i,m}_j)_{j\in I_P}$ is a $BNE(\mu^{i,m})$ strategy profile: each $\sigma^{i,m}_j : \Theta_j \to \Delta(A_j)$ is such that $\sigma^{i,m}_j(\theta_j)$ is a best-response to the other players' strategies, $(\sigma^{i,m}_l)_{l\in I_P\setminus\{j\}}$, if agent j has type θ_j, $\forall\theta_j \in \Theta_j$. The best-response of agent j of type θ_j is calculated given the belief that the type profile of the other agents is $\theta_{-j} \in \Theta_{-j}$ with probability $\mu^{i,m}(\theta|\theta_j)$. On the other hand, the strategy profile that results from a genuine rejection, denoted $\sigma^{(i),m} \equiv (\sigma^{(i),m}_i, (\sigma^{i,m}_j)_{j\in I_P})$, is not typically a $BNE(\mu^{i,m})$. The genuine rejector best-responds to $(\sigma^{i,m}_j)_{j\in I_P}$ given his belief that the type profile of the other players is $\theta_{-i} \in \Theta_{-i}$ with probability $\mu^0(\theta|\theta_i)$. The genuine rejector does not update his belief from μ^0 to $\mu^{i,m}$ as a result of his own deviation (although acceptors believe, and the rejector knows that acceptors believe, that $\mu^{i,m}$ is the common disagreement belief).

Given the sequence $((\varepsilon^{i,m}, \mu^{i,m}, \sigma^{i,m}, \sigma^{(i),m}_i)_{i\in I})_{m\in\mathbb{N}}$, pick a subsequence that converges and denote the limit by $(0, \mu^i, \sigma^i, \sigma^{(i)}_i)_{i\in I}$. The *ex ante* payoff function in the limit, denoted $\mathbb{E}_{\mu^{i,m}}\left[\pi^R_j(z(\theta), \theta)\right]$, is bilinear in z and μ, and is uniformly bounded. This implies that, if $\mu^i_j(\theta_j) > 0$, the limit strategy $\sigma^i_j(\theta_j)$ is a best-response to σ^i_{-j}: if an alternative strategy of agent j of type θ_j, denoted $\sigma^{i'}_j(\theta_j)$, was strictly preferred in the limit, the same strategy would be strictly preferred for sufficiently large

m, which would be a contradiction. Although $\mu_j^{i,m}(\theta_j) > 0$, along the sequence, it is possible that $\mu_j^i(\theta_j) = 0$ in the limit. In that case, considering the limit conditional probability, the objective of agent j of type θ_j becomes well defined and the same kind of contradiction is reached. We conclude that $\sigma^i \in BNE(\mu^i)$. Acceptors $j \in I_P \setminus \{i\}$ play the same strategy profile, $\sigma_{-j}^i \in BNE_{-i}(\mu^i)$, after a genuine rejection. The limit strategy of a genuine rejector, $\sigma_i^{(i)}$, is also a best-response in the limit (for the same reason as before). Let $y^{(i),m}$ and $y^{(i)}$ denote the distributions over joint actions induced by the strategy profiles played after a genuine rejection along the sequence and in the limit, respectively. Since (IC') and (IR) are satisfied along the sequence (with the outside option being $y^{(i),m}$), (IC) and (IR) are also satisfied in the limit (with the outside option being $y^{(i)}$). □

Proof of Proposition 9. ($\Rightarrow$) Let $(\varepsilon^m, \psi^m, x^m)_{m\in\mathbf{N}}$ be a sequence of strictly trembling mechanisms converging to a non-trembling mechanism, $(0, \psi, x)$, such that each mechanism in the sequence has a sequential equilibrium in which the agent participates and reports truthfully. Consider the equilibrium of mechanism $(\varepsilon^m, \psi^m, x^m)$. After a spurious rejection by agent $i \in I$, given the type profile, $\theta \in \Theta$: a common belief, $\mu^{i,m} \in \mathcal{B}^\theta \equiv \{\mu \in \Delta(\Theta) : \mu_j(\theta_j) > 0, \forall j \in I\}$ results from Bayesian updating; and players receive, and obey, recommendations according to $y^{i,m} \in BCE^I(\mu^{i,m})$. After a genuine rejection by agent $i \in I$ (off-path): each acceptor $j \in I_P \setminus \{i\}$ is recommended, and obeys (presuming that the rejection has been spurious), an action, $a_j \in A_j$, in the support of $y_j^{i,m}$; the profile of recommendations, a_{-i}, is drawn from a distribution $y^{(i),m}(\theta_{-i})$; the rejector, i, chooses a best-response, $a_i \in A_i$, or follows a recommendation that allows him to obtain a higher payoff (the worst that can happen to player i is to receive a meaningless private signal, which prevents him from conditioning his best-response). Since there is a sequential equilibrium with truthful participation, x^m satisfies (IC') and (IR) for a joint distribution over action profiles, $y^{(i),m} : \Theta \to \Delta(A)$, induced by $y_{-i}^{(i),m} \in \Delta(\mathcal{A}_{-i}(\theta_{-i}))$, for each $\theta_{-i} \in \Theta_{-i}$ (which is the distribution over acceptors' action profiles following a genuine rejection by agent i) and by a best-response from the rejector, $\sigma_i^{(i),m} : \Theta \to \Delta(A_i)$.

Consider a subsequence $\left(y_{-i}^{(i),m}\right)_{m\in M}$ that converges and denote the limit by $y_{-i}^{(i)}$. It is clear that $y_{-i}^{(i)} \in \Delta(\mathcal{A}_{-i}(\theta_{-i}))$, because the support of a sequence of distributions "cannot increase in the limit". Denote the limit of $\sigma_i^{(i),m}$ by $\sigma_i^{(i)}$. Since the best-response mapping is upper hemi-continuous, $\sigma_i^{(i)}$ is a best-response to $y_{-i}^{(i)}$. Participation and truth-telling along the sequence implies participation and truth-telling in the limit: x satisfies (IC) and (IR) when rejection leads to $(y_{-i}^{(i)}, \sigma_i^{(i)})$.

($\Leftarrow$) Since $y_{-i}^{(i)}(\theta_{-i}) \in \Delta(\mathcal{A}_{-i}(\theta_{-i}))$, $\forall \theta_{-i} \in \Theta_{-i}$, from Lemma 1, $\exists y^i \in BCE^I(\mu^i)$, where μ^i is any common belief with full support, s.t., for each $j \in I_P \setminus \{i\}$: $a_j \in supp\left[y_j^{(i)}(\theta_{-i})\right]$ implies $a_j \in supp(y_j^i)$.

Construct a trembling device such that if the input message is $\theta \in \Theta$, a spurious rejection by agent $i \in I$ is generated with probability $\varepsilon^i(\theta) \equiv \frac{\mu^i(\theta)\epsilon^m}{\mu^0(\theta)|I|}$, where $0 < \epsilon^m < \min_{\theta\in\Theta} \mu^0(\theta)$. Following a spurious rejection by agent i, the disagreement belief is μ^i, and the mediator recommends $y^i(\theta)$. Observe that all actions in the support of $y_j^{(i)}(\theta_{-i})$ are chosen with positive probability by player j when he has type θ_j, after a spurious rejection by agent i. After a genuine rejection by agent i, let the mediator send recommendations to the acceptors according to $y_{-i}^{(i)}(\theta_{-i})$. These recommendations are obeyed, because acceptors $j \in I_P \setminus \{i\}$ believe that a spurious rejection has occurred and the

recommendations are made according to $y^i \in BCE^I(\mu^i)$.

Since x satisfies (IC) and (IR) for the given $y^{(i)}$, we can define (in the usual way) a sequence of pairs, $(x^m, \epsilon^m)_{m\in\mathbf{N}}$, such that (IC') and (IR) are satisfied along the sequence for the given $y^{(i)}$. Consider a candidate sequential equilibrium in which agents always participate and report truthfully. As we have seen, it is optimal for players to obey the recommendations made after a spurious rejection, and, therefore, acceptors also obey the recommendations made after a genuine rejection (believing that the rejection has been spurious, which is infinitely more likely). After a genuine rejection, the rejector anticipates that the acceptors will play $y^{(i)}_{-i} : \Theta_{-i} \to \Delta(A_{-i})$, and plays a best-response, $\sigma_i(\theta_i)$. This means that the distribution over action profiles is in fact induced by $(y^{(i)}_{-i}, \sigma^{(i)}_i)$ such that $y^{(i)}_{-i} \in \Delta(\mathcal{A}_{-i}(\theta_{-i}))$, $\forall\theta_{-i} \in \theta_{-i}$, and $\sigma^{(i)}_i(\theta_i)$ is a best-response to $y^{(i)}_{-i}$, $\forall\theta_i \in \Theta_i$. Since (x^m, ϵ^m) satisfies (IC') and (IR), no deviation from the candidate sequential equilibrium is beneficial. Letting, $\epsilon^m \to 0$, the mechanism converges to a non-trembling mechanism. □

C Examples

C.1 Cournot duopoly

Claim: For each $\mu_A \in [0, 1]$, there is a menu of contracts that maximizes joint-profit and leaves each type of firm A indifferent between accepting or rejecting (incentive compatibility constraints are not binding). The profit of firm P is strictly increasing in μ_A.

Proof: Joint-profit maximization under complete information yields: $(q_A, q_P) = (\frac{1}{3}, \frac{1}{3})$ if $\theta_A = 0$, and $(q_A, q_P) = (\frac{1}{9}, \frac{4}{9})$ if $\theta_A = \frac{1}{3}$. The resulting profits of firm A (before side-payments) are $\pi_A(0) = \frac{1}{6}$ and $\pi_A(\frac{1}{3}) = \frac{1}{27}$. In case of rejection, firm P produces $q^R_P = \frac{18+2\mu_A}{45}$, and the profits of firm A are $\pi^R_A(0) = \left(\frac{36-\mu_A}{90}\right)^2$ and $\pi^R_A(\frac{1}{3}) = \left(\frac{21-\mu_A}{90}\right)^2$. The side-payments to firm A that leave it indifferent between accepting or rejecting are $t(0) = \frac{-54-72\mu_A+\mu_A^2}{90^2}$ and $t(\frac{1}{3}) = \frac{141-42\mu_A+\mu_A^2}{90^2}$. Strict incentive compatibility can easily be checked: if firm A announces $\theta_A = \frac{1}{3}$ when $\theta_A = 0$, its payoff is $\pi^D_A(0) = \left(1 - \frac{1}{9} - \frac{2}{9}\right)\frac{1}{9} + \frac{141-42\mu_A+\mu_A^2}{90^2} = \frac{741-42\mu_A+\mu_A^2}{90^2} < \pi^R_A(0)$; if firm A announces $\theta_A = 0$ when $\theta_A = \frac{1}{3}$, its payoff is $\pi^D_A(\frac{1}{3}) = \left(1 - \frac{1}{3} - \frac{1}{6} - \frac{1}{3}\right)\frac{1}{3} + \frac{-54-72\mu_A+\mu_A^2}{90^2} = \frac{396-72\mu_A+\mu_A^2}{90^2} < \pi^R_A(\frac{1}{3})$. Since both side-payments are strictly decreasing in μ_A, the profit of firm P is strictly increasing in μ_A. □

C.2 Cournot triopoly (without correlating device)

Claim: Suppose that the common disagreement belief is that $\theta_i = \frac{1}{3}$ with probability μ_i, for $i \in \{A, B\}$, and that θ_A and θ_B are independent. For each $(\mu_A, \mu_B) \in [0, 1]^2$, there is a menu of contracts that maximizes joint-profit and leaves each type of each firm indifferent between

accepting or rejecting (incentive compatibility constraints are not binding). The profit of firm P is strictly increasing in μ_A and in μ_B.

Proof: Joint-profit maximization yields: $(q_A, q_B, q_P) = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ if $(\theta_A, \theta_B) = (0, 0)$; $(q_i, q_j, q_P) = (\frac{1}{3}, 0, \frac{1}{3})$ if $(\theta_i, \theta_j) = (0, \frac{1}{3})$; $(q_i, q_j, q_P) = (0, \frac{1}{3}, \frac{1}{3})$ if $(\theta_i, \theta_j) = (\frac{1}{3}, 0)$; and $(q_A, q_B, q_P) = (\frac{1}{12}, \frac{1}{12}, \frac{5}{12})$ if $(\theta_A, \theta_B) = (\frac{1}{3}, \frac{1}{3})$. The resulting profits (before side-payments) of firm A, $\pi_A(\theta_A, \theta_B)$, are $\pi_A(0, 0) = \frac{1}{8}$, $\pi_A(0, \frac{1}{3}) = \frac{1}{6}$, $\pi_A(\frac{1}{3}, 0) = 0$, and $\pi_A(\frac{1}{3}, \frac{1}{3}) = \frac{1}{36}$. Under the common disagreement belief: the output of firm P is $q_P = \frac{9+\mu_A+\mu_B}{27}$; and the outputs of firms A and B are $q_A(\theta_A) = \frac{18-\mu_A+2\mu_B}{54} - \frac{\theta_A}{2}$ and $q_B(\theta_B) = \frac{18-\mu_B+2\mu_A}{54} - \frac{\theta_B}{2}$. The aggregate output of firms P and B is: $q_P + q_B(\theta_B) = \frac{36+4\mu_A+\mu_B}{54} - \frac{\theta_B}{2}$. If firm A rejects, it maintains its prior about θ_B, denoted μ^0, and produces: $q_A = \frac{72+9\mu^0-4\mu_A-\mu_B}{216} - \frac{\theta_A}{2}$. This yields an expected profit given by: $\pi_A^R(\theta_A) = \left(\frac{72+9\mu^0-4\mu_A-\mu_B}{216} - \frac{\theta_A}{2}\right)^2$. Consider a mechanism that maximizes joint-profit and leaves each type of each firm indifferent between accepting or rejecting. If firm A announces $\theta_A = \frac{1}{3}$ when $\theta_A = 0$, its profit is: $\pi_A^D(0) = \pi_A^R(\frac{1}{3}) + \frac{1}{3}\mu^0\frac{1}{12} = \left(\frac{36+9\mu^0-4\mu_A-\mu_B}{216}\right)^2 + \frac{\mu^0}{36}$. Since $\pi_A^R(0) - \pi_A^D(0) = \frac{54-9\mu^0-4\mu_A-\mu_B}{648} > 0$, lying is strictly sub-optimal. If firm A announces $\theta_A = 0$ when $\theta_A = \frac{1}{3}$, its profit is: $\pi_A^D(\frac{1}{3}) = \pi_A^R(0) - \frac{1}{3}\left[(1-\mu^0)\frac{1}{4} + \mu^0\frac{1}{3}\right] = \left(\frac{72+9\mu^0-4\mu_A-\mu_B}{216}\right)^2 - \frac{3+\mu^0}{36}$. Again, lying is strictly sub-optimal: $\pi_A^R(\frac{1}{3}) - \pi_A^D(\frac{1}{3}) = \frac{9\mu^0+4\mu_A+\mu_B}{648}$. The payoffs of firms A and B are strictly decreasing in μ_A and in μ_B. Hence, the profit of firm P is strictly increasing in μ_A and in μ_B. □

For simplicity, suppose that $\mu^0 = \frac{1}{2}$. With non-trembling mechanisms, the "*no signaling what you don't know*" restriction implies that $\mu_B = \frac{1}{2}$ if firm A rejects. Therefore, the worst beliefs for the rejector are $(\mu_A, \mu_B) = \left(1, \frac{1}{2}\right)$, which yield: an expected output from rivals $q_P + \frac{1}{2}[q_B(0) + q_B(\frac{1}{3})] = \frac{2}{3}$, own output $q_A(\theta_A) = \frac{1}{3} - \frac{\theta_A}{2}$, and own profit $\pi_A^R(\theta_A) = \left(\frac{1}{3} - \frac{\theta_A}{2}\right)^2$. With trembling mechanisms, the worst beliefs, $(\mu_A, \mu_B) = (1, 1)$, can be induced. This entails: $q_P + \frac{1}{2}[q_B(0) + q_B(\frac{1}{3})] = \frac{73}{108}$, $q_A(\theta_A) = \frac{143}{432} - \frac{\theta_A}{2}$, and $\pi_A^R(\theta_A) = \left(\frac{143}{432} - \frac{\theta_A}{2}\right)^2$.

C.3 Cournot triopoly (with correlating device)

Let $d(x, y) \equiv |\underline{x}_A - \underline{y}_A| + |\overline{x}_A - \overline{y}_A| + |\underline{x}_B - \underline{y}_B| + |\overline{x}_B - \overline{y}_B| + |\underline{x}_P - \underline{y}_P| + |\overline{x}_P - \overline{y}_P|$ and let $Tx \equiv \left(\frac{1}{3} - \frac{\overline{x}_B+\overline{x}_P}{4}, \frac{1}{2} - \frac{\underline{x}_B+\underline{x}_P}{4}, \frac{1}{3} - \frac{\overline{x}_A+\overline{x}_P}{4}, \frac{1}{2} - \frac{\underline{x}_A+\underline{x}_P}{4}, \frac{1}{2} - \frac{\overline{x}_A+\overline{x}_B}{4}, \frac{1}{2} - \frac{\underline{x}_A+\underline{x}_B}{4}\right)$.

Claim: The transformation $T : [0, 1]^6 \to [0, 1]^6$ is a contraction with modulus $\frac{1}{2}$.

Proof: Observe that:

$$\begin{aligned} d(Tx, Ty) &= \frac{1}{4}|\overline{x}_B - \overline{y}_B + \overline{x}_P - \overline{y}_P| + \frac{1}{4}|\underline{x}_B - \underline{y}_B + \underline{x}_P - \underline{y}_P| + \frac{1}{4}|\overline{x}_A - \overline{y}_A + \overline{x}_P - \overline{y}_P| \\ &\quad + \frac{1}{4}|\underline{x}_A - \underline{y}_A + \underline{x}_P - \underline{y}_P| + \frac{1}{4}|\overline{x}_A - \overline{y}_A + \overline{x}_B - \overline{y}_B| + \frac{1}{4}|\underline{x}_A - \underline{y}_A + \underline{x}_B - \underline{y}_B| \\ &\leq \frac{1}{2}\left(|\overline{x}_B - \overline{y}_B| + |\overline{x}_P - \overline{y}_P| + |\underline{x}_B - \underline{y}_B| + |\underline{x}_P - \underline{y}_P| + |\overline{x}_A - \overline{y}_A| + |\underline{x}_A - \underline{y}_A|\right) \\ &= \frac{1}{2}d(x, y). \end{aligned}$$

□

References

Abreu, D. and Matsushima, H. (1992). Virtual implementation in iteratively undominated strategies: complete information. *Econometrica*, 60(5):993–1008.

Balzer, B. and Schneider, J. (2016). Managing a conflict: alternative dispute resolution in contests. *mimeo, February 2016*.

Banks, J. S. and Sobel, J. (1987). Equilibrium selection in signaling games. *Econometrica*, 55(3):647–661.

Bergemann, D. and Morris, S. (2013). Robust predictions in games with incomplete information. *Econometrica*, 81(4):1251–1308.

Bergemann, D. and Morris, S. (2016). Bayes correlated equilibrium and the comparison of information structures in games. *Theoretical Econ*, 11(2):487–522.

Bester, H. and Strausz, R. (2001). Contracting with imperfect commitment and the revelation principle: the single agent case. *Econometrica*, 69(4):1077–1098.

Bester, H. and Strausz, R. (2007). Contracting with imperfect commitment and noisy communication. *J Econ Theory*, 136(1):236–259.

Blume, A., Board, O. J., and Kawamura, K. (2007). Noisy talk. *Theoretical Econ*, 2(4):395–440.

Blume, L., Brandenburger, A., and Dekel, E. (1991). Lexicographic probabilities and choice under uncertainty. *Econometrica*, 59(1):61–79.

Board, S. and Pycia, M. (2014). Outside options and the failure of the coase conjecture. *American Econ Rev*, 104(2):656–671.

Calzolari, G. and Pavan, A. (2006a). Monopoly with resale. *RAND J Econ*, 37(2):362–375.

Calzolari, G. and Pavan, A. (2006b). On the optimality of privacy in sequential contracting. *J Econ Theory*, 130(1):168–204.

Calzolari, G. and Pavan, A. (2008). On the use of menus in sequential common agency. *Games Econ Behavior*, 64(1):329–334.

Celik, G. and Peters, M. (2011). Equilibrium rejection of a mechanism. *Games Econ Behavior*, 73(2):375–387.

Che, Y.-K. and Gale, I. (1998a). Standard auctions with financially constrained bidders. *Rev Econ Studies*, 65(1):1–21.

Che, Y.-K. and Gale, I. L. (1998b). Caps on political lobbying. *American Econ Rev*, 88(3):643–651.

Che, Y.-K. and Kim, J. (2006). Robustly collusion-proof implementation. *Econometrica*, 74(4):1063–1107.

Che, Y.-K. and Kim, J. (2009). Optimal collusion-proof auctions. *J Econ Theory*, 144(2):565–

603.

Chen, C.-L. and Tauman, Y. (2006). Collusion in one-shot second-price auctions. *Econ Theory*, 28(1):145–172.

Cho, I.-K. and Kreps, D. M. (1987). Signaling games and stable equilibria. *Quarterly J Econ*, 102(2):179–221.

Cho, I.-K. and Sobel, J. (1990). Strategic stability and uniqueness in signaling games. *J Econ Theory*, 50(2):381–413.

Cooter, R. D. and Rubinfeld, D. L. (1989). Economic analysis of legal disputes and their resolution. *J Econ Literature*, 27(3):1067–1097.

Cramton, P. C. and Palfrey, T. R. (1990). Cartel enforcement with uncertainty about costs. *Int Econ Rev*, 31(1):17–47.

Cramton, P. C. and Palfrey, T. R. (1995). Ratifiable mechanisms: learning from disagreement. *Games Econ Behavior*, 10(2):255–283.

Crawford, V. P. and Sobel, J. (1982). Strategic information transmission. *Econometrica*, 50(6):1431–1451.

Deb, R. and Said, M. (2015). Dynamic screening with limited commitment. *J Econ Theory*, 159:891–928.

Dequiedt, V. (2007). Efficient collusion in optimal auctions. *J Econ Theory*, 136(1):302–323.

Einy, E., Haimanko, O., and Tumendemberel, B. (2012). Continuity of the value and optimal strategies when common priors change. *Int J Game Theory*, 41(4):829–849.

Eső, P. and Schummer, J. (2004). Bribing and signaling in second price auctions. *Games Econ Behavior*, 47(2):299–324.

Evangelista, F. S. and Raghavan, T. (1996). A note on correlated equilibrium. *Int J Game Theory*, 25(1):35–41.

Farrell, J. (1993). Meaning and credibility in cheap-talk games. *Games Econ Behavior*, 5(4):514–531.

Faure-Grimaud, A., Laffont, J.-J., and Martimort, D. (2003). Collusion, delegation and supervision with soft information. *Rev Econ Studies*, 70(2):253–279.

Forges, F. (1986). An approach to communication equilibria. *Econometrica*, 54(6):1375–1385.

Forges, F. (1988). Communication equilibria in repeated games with incomplete information. *Mathematics of Operations Research*, 13(2):191–231.

Forges, F. (1993). Five legitimate definitions of correlated equilibrium in games with incomplete information. *Theory and Decision*, 35(3):277–310.

Forges, F. (2006). Correlated equilibrium in games with incomplete information revisited. *Theory and Decision*, 61(4):329–344.

Fudenberg, D. and Tirole, J. (1986). A "signal-jamming" theory of predation. *RAND J Econ*, 17(3):366–376.

Fudenberg, D. and Tirole, J. (1991). Perfect Bayesian equilibrium and sequential equilibrium. *J Econ Theory*, 53(2):236–260.

Gavious, A., Moldovanu, B., and Sela, A. (2002). Bid costs and endogenous bid caps. *RAND J Econ*, 33(4):709–722.

Goltsman, M., Hörner, J., Pavlov, G., and Squintani, F. (2009). Mediation, arbitration and negotiation. *J Econ Theory*, 144(4):1397–1420.

Govindan, S. and Wilson, R. (2005). Refinements of Nash equilibrium. In Durlauf, S. and Blume, L., editors, *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, London, 2nd edition.

Grossman, S. J. and Perry, M. (1986). Perfect sequential equilibrium. *J Econ Theory*, 39(1):97–119.

Hart, O. and Tirole, J. (1990). Vertical integration and market foreclosure. *Brookings papers on economic activity: Microeconomics*, 1990:205–286.

Hörner, J., Morelli, M., and Squintani, F. (2015). Mediation and peace. *Rev Econ Studies*, 82(4):1483–1501.

Jullien, B. (2000). Participation constraints in adverse selection models. *J Econ Theory*, 93(1):1–47.

Jullien, B., Pouyet, J., and Sand-Zantman, W. (2016). An offer you can't refuse: Early contracting with endogenous threat. *forthcoming in Rand J Econ*.

Kandori, M. (1991). Cooperation in finitely repeated games with imperfect private information. *mimeo*.

Kandori, M. (2002). Introduction to repeated games with private monitoring. *J Econ Theory*, 102(1):1–15.

Kihlstrom, R. and Vives, X. (1992). Collusion by asymmetrically informed firms. *J Econ & Management Strategy*, 1(2):371–396.

Kohlberg, E. and Mertens, J.-F. (1986). On the strategic stability of equilibria. *Econometrica*, 54(5):1003–1037.

Kreps, D. M. and Wilson, R. (1982). Sequential equilibria. *Econometrica*, 50(4):863–894.

Laffont, J.-J. and Martimort, D. (1997). Collusion under asymmetric information. *Econometrica*, 65(4):875–911.

Laffont, J.-J. and Martimort, D. (2000). Mechanism design with collusion and correlation. *Econometrica*, 68(2):309–342.

Lehrer, E. (1992). Correlated equilibria in two-player repeated games with nonobservable

actions. *Mathematics of Operations Research*, 17(1):175–199.

Lehrer, E. and Sorin, S. (1997). One-shot public mediated talk. *Games Econ Behavior*, 20(2):131–148.

Mailath, G. J., Matthews, S. A., and Sekiguchi, T. (2002). Private strategies in finitely repeated games with imperfect public monitoring. *Contributions in Theoretical Econ*, 2(1).

Mailath, G. J., Okuno-Fujiwara, M., and Postlewaite, A. (1993). Belief-based refinements in signalling games. *J Econ Theory*, 60(2):241–276.

Martimort, D. and Sand-Zantman, W. (2016). A mechanism design approach to climate-change agreements. *J European Econ Association*, 14(3):669–718.

McAfee, R. P. and Schwartz, M. (1994). Opportunism in multilateral vertical contracting: Nondiscrimination, exclusivity, and uniformity. *American Econ Rev*, 84(1):210–230.

Mertens, J.-F. (1989). Stable equilibria – a reformulation: Part I. Definition and basic properties. *Mathematics of Operations Research*, 14(4):575–625.

Mertens, J.-F. (1991). Stable equilibria – a reformulation: Part II. Discussion of the definition, and further results. *Mathematics of Operations Research*, 16(4):694–753.

Mertens, J.-F., Sorin, S., and Zamir, S. (2015). *Repeated games*. Cambridge University Press.

Mitusch, K. and Strausz, R. (2005). Mediation in situations of conflict and limited commitment. *J Law, Econ, and Organization*, 21(2):467–500.

Mookherjee, D. and Tsumagari, M. (2004). The organization of supplier networks: effects of delegation and intermediation. *Econometrica*, 72(4):1179–1219.

Myerson, R. B. (1982). Optimal coordination mechanisms in generalized principal–agent problems. *J Mathematical Econ*, 10(1):67–81.

Myerson, R. B. (1986). Multistage games with communication. *Econometrica*, 54(2):323–358.

Nau, R. F. and McCardle, K. F. (1990). Coherent behavior in noncooperative games. *J Econ Theory*, 50(2):424–444.

Olszewski, W. and Siegel, R. (2016). Bid caps in contests. *mimeo, April 2016*.

Pavan, A. and Calzolari, G. (2009). Sequential contracting with multiple principals. *J Econ Theory*, 144(2):503–531.

Pavlov, G. (2008). Auction design in the presence of collusion. *Theoretical Econ*, 3(3):383–429.

Philippon, T. and Skreta, V. (2012). Optimal interventions in markets with adverse selection. *American Econ Rev*, 102(1):1–28.

Rabin, M. and Sobel, J. (1996). Deviations, dynamics, and equilibrium refinements. *J Econ Theory*, 68(1):1–25.

Rachmilevitch, S. (2013). Endogenous bid rotation in repeated auctions. *J Econ Theory*,

148(4):1714–1725.

Rahman, D. (2012). But who will monitor the monitor? *American Econ Rev*, 102(6):2767–2797.

Rahman, D. (2014). The power of communication. *American Econ Rev*, 104(11):3737–3751.

Rahman, D. and Obara, I. (2010). Mediated partnerships. *Econometrica*, 78(1):285–308.

Renault, J. and Tomala, T. (2004). Communication equilibrium payoffs in repeated games with imperfect monitoring. *Games Econ Behavior*, 49(2):313–344.

Rey, P. and Vergé, T. (2004). Bilateral control with vertical contracts. *RAND J Econ*, 35(4):728–746.

Schummer, J. (2000). Manipulation through bribes. *J Econ Theory*, 91(2):180–198.

Segal, I. (1999). Contracting with externalities. *Quarterly J Econ*, 114(2):337–388.

Selten, R. (1975). Reexamination of the perfectness concept for equilibrium points in extensive games. *Int J Game Theory*, 4(1):25–55.

Singh, N. and Vives, X. (1984). Price and quantity competition in a differentiated duopoly. *RAND J Economics*, 15(4):546–554.

Strausz, R. (2012). Mediated contracts and mechanism design. *J Econ Theory*, 147(3):1280–1290.

Sugaya, T. and Wolitsky, A. (2016). Bounding equilibrium payoffs in repeated games with private monitoring. *forthcoming in Theoretical Econ*.

Szech, N. (2015). Tie-breaks and bid-caps in all-pay auctions. *Games Econ Behavior*, 92:138–149.

Tan, G. and Yilankaya, O. (2007). Ratifiability of efficient collusive mechanisms in second-price auctions with participation costs. *Games Econ Behavior*, 59(2):383–396.

Tirole, J. (1986). Hierarchies and bureaucracies: On the role of collusion in organizations. *J Law, Econ, and Organization*, 2(2):181–214.

Tirole, J. (1992). Collusion and the theory of organizations. In Laffont, J.-J., editor, *Advances in Economic Theory: Sixth World Congress*, volume II. Cambridge University Press.

Tirole, J. (2012). Overcoming adverse selection: How public intervention can restore market functioning. *American Econ Rev*, 102(1):29–59.

Tomala, T. (2009). Perfect communication equilibria in repeated games with imperfect monitoring. *Games Econ Behavior*, 67(2):682–694.

van Damme, E. (1989). Stable equilibria and forward induction. *J Econ Theory*, 48(2):476–496.

Yamashita, T. (2014). Strategic and structural uncertainties in robust implementation. *mimeo, April 2014*.