Subversive Conversations*

Nemanja Antic† Archishman Chakraborty‡ Rick Harbaugh§

This version: July, 2024

Abstract

Two players with common interests exchange information to make a decision. But they fear scrutiny. Their unencrypted communications will be observed by another agent with different interests who can object to their decision. We show how the players can implement their ideal decision rule using a back and forth conversation. Such a subversive conversation reveals enough information for the players to determine their best decision but not enough information for the observer to determine whether the decision was against his interest. Our results show how conversations can maintain deniability even in the face of leaks, hacks, and other public exposures.

JEL Classification: C72, D71, D72, D82.

Keywords: conversations, deniability, subversion, cheap talk, persuasion.

*We thank the editor and four anonymous referees, as well as conference participants at the Junior Theory Workshop at U. Bonn, Decentralization Conference at U. Michigan, Stonybrook International Game Theory Conference, North-South Chicago Theory Conference, the Midwest Theory Conference, and the NBER Organizational Economics Meetings; as well as seminar participants at the Delhi School of Economics, Monash University, Northwestern University, Queen Mary College, Toulouse School of Economics, University of Bath, UCLA, Norwegian Business School, University of Arizona and Arizona State University. For helpful comments, we also thank David Austen-Smith, Sandeep Baliga, Gabriel Carroll, Eddie Dekel, Wouter Dessein, Wioletta Dziuda, Georgy Egorov, Jeff Ely, Tim Feddersen, Daniel Garrett, Parikshit Ghosh, Faruk Gul, Jason Hartline, Philip Kalikman, Andreas Kleiner, Aaron Kolb, Elliot Lipnowski, Meg Meyer, Gregory Pavlov, Marilyn Pease, Nicola Persico, Doron Ravid, Ludovic Renou, Patrick Rey, Ariel Rubinstein, Alvaro Sandroni, Joel Sobel, Lars Stole, Jean Tirole, Bilge Yilmaz and Bill Zame. Edited by Emir Kamenica.

†Kellogg School of Management, Northwestern University; nemanja.antic@kellogg.northwestern.edu.

‡Syms School of Business, Yeshiva University; archishman@yu.edu.

§Kelley School of Business, Indiana University; riharbau@indiana.edu.

1 Introduction

People with similar interests need to share information to make a decision. But their discussions may be observed by other people with different interests. Minutes of government meetings are often on the public record. Deliberations of corporate boards can be accessible to other stakeholders. Communications between parties to a merger may be subpoenaed by anti-trust regulators. Employee messages may be subject to discovery in lawsuits. Even if communication is private, the chance of exposure always remains. Emails can be hacked, codes can be broken, firewalls can be breached and whistleblowers can go public.[1]

When communication is public or exposure is a concern, does decision-making suffer? We study this problem of communication under scrutiny. Two players with private information and common interests must exchange their information to decide whether to accept a proposal or not. An uninformed observer with partially opposed interests sees the players' messages and decisions. The observer could be a regulator, a supervisor, or the wider public. In some states the players and observer agree on the best decision, while in other states they do not. The players want to avoid controversy, protests, or penalties that may be incurred if the observer believes the decision was against the observer's interests.

We focus on the possibility of subversion. The players subvert when they share enough information to take the same ideal decision they would take in the absence of scrutiny, while concealing enough information to maintain *plausible deniability* that they acted against the observer's interests. Because of this deniability constraint, the players cannot immediately reveal their information but must use a back-and-forth conversation. As the conversation progresses, they share increasingly detailed information but only once a suitable context has been created by previous statements. A subversive conversation is an indirect mechanism that allows the players to get their first best outcomes even when the conversation is, or might become, public.

Consider a committee of two managers evaluating whether to accept or reject a new mining project which has environmental costs and economic benefits. The public (the observer) cares more about the environment than the firm does. Manager X only knows the project's economic benefit $x \in \{0, 1/2, 1\}$ while manager Y only knows the environmental cost $y \in \{0, 1/2, 1\}$. The two managers both prefer acceptance if the project is good ($x > y$) or mediocre ($x = y$), and rejection if it is bad ($x < y$). The uninformed public finds bad and mediocre projects equally undesirable and is willing to accept the project if and only if it has at least an even chance of being good. This gives rise to a deniability constraint faced by the managers: each time they

[1]Prominent exposures include the leak of Climategate emails by a server breach, subpoena of private documents and messages in the VW Dieselgate and Purdue Pharma settlements, whistleblowing by a government employee that led to a presidential impeachment, and data extraction from cellphones that led to arrests of Hong Kong activists. Silberman and Bruno (2017) recount many cases where subpoenaed emails and memos were key pieces of evidence in antitrust litigation. Even for attorney-client communication they advise participants "to assume somehow every word will get published". Attorney-client privilege can be legally trumped by the "crime-fraud exception" as shown by a case involving the handling of classified documents by a former president.

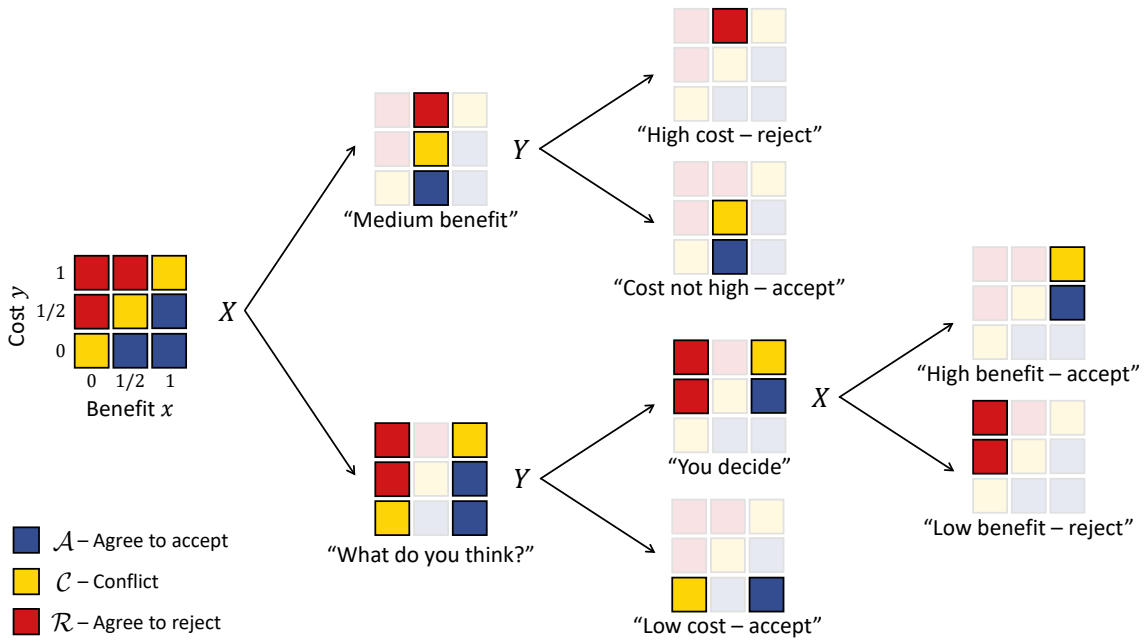

Figure 1: Conversation tree

accept the proposal, the public must believe the project is at least as likely to be good as not. Priors are uniform on x and y. Communication is costless.

Figure 1 depicts a conversation which dynamically pools and separates types, allowing the managers to determine if the project is truly bad while concealing from the public whether it is good or mediocre. If the benefit is medium, manager X says so in the first round of this conversation, as seen in the upper branch of the tree. Then if the cost is high manager Y rejects the project since it is clearly bad. Otherwise Y accepts the project. The public learns the project is as likely to be good as not, while only Y knows whether it is truly good or only mediocre.

If the benefit is low or high then in the first round X passes the conversation over to Y as seen in the lower branch of the tree. If the cost is low then the project is either good or mediocre, so Y accepts it. Only X knows the actual quality, but strategically reveals nothing. If instead the cost is medium or high then Y passes the conversation back to X. Having learned that the cost is not low, in the third round X now rejects the project if the benefit is low since the project is clearly bad. If instead the benefit is high then X accepts the project. The project is as likely to be good as mediocre, but only Y knows which and says nothing further.

By the end of the conversation the managers have pooled all mediocre and good states where they want to accept the project, while identifying all bad states where they want to reject it. Since they take their ideal decision in each state and achieve their first-best outcome, there is no incentive to deviate from this strategy. As long as the managers follow the subversive communication protocol, they will meet the deniability constraint and any leaks or other disclosure of their exchanges will not reveal that they knowingly acted against the public interest.

We examine binary decision problems like this one, but with richer information structures and more general types of conflict between the players and the observer. Using constructive methods like in Figure 1, as well as a belief-based approach based on Aumann and Hart (1986), we show that in a wide range of situations the committee can take its favored decisions via a conversation that is sufficiently informative for them but conceals enough from the observer. When subversion is possible, it is as if the committee is free to choose as it pleases.

Subversion is always possible when the choice is between two ex ante identical alternatives and the impartial observer wants the committee to pick the better of the two, e.g., the choice between two job market candidates. Each committee member is informed about the quality one candidate and the committee is biased towards a particular candidate. Although their information is dispersed, they can use a conversation to achieve the same ideal outcome in the presence of the observer as they would under completely private communication. Neither player reveals all that she knows immediately but instead waits for the right moment. Initially each player conceals unfavorable news about the likely rank of the favored candidate, while also waiting to reveal good news in order to create a favorable context in case her partner has bad news.

This existence result for the case of two ex ante identical alternatives extends to situations where the committee's bias is in favor of the alternative that is ex ante more likely to be better, to cases with correlated information, and to uncertainty about the magnitude or sign of the bias. More generally, we study changes to preferences and priors under which a given conversation remains subversive. These changes identify "invariance" properties of a subversive conversation. They implicitly define a set of games for which the same conversation is subversive. This robustness feature lowers the design burden on the committee—it can use the same conversation across different games and also when the observer is uncertain about the committee's exact preferences.

Our results imply accountability and regulatory compliance may be difficult to ensure, even for transparent organizations. In the leading example above, if the managers use a subversive conversation in reporting to a leader who shares their preferences, any ex post scrutiny will not find the latter had reason to stop the project.[2] This is better than maintaining deniability by foregoing decision-relevant information from subordinates, which can lead to inefficient outcomes. Guiding questions that manage the conversation, or other suitable protocols, can yield plausible deniability for leadership and still allow them to obtain the information they need for a decision.[3]

[2] As the White House Counsel said to President George W. Bush regarding enhanced interrogation techniques, "Mr. President, I think for your own protection you don't need to know the details of what's going on here." See Garicano and Rayo (2016).

[3] If organizations face similar decision problems regularly, they have an incentive to design communication protocols that will work for every realization of the state. To quote Shannon (1948), "The system must be designed to operate for each possible selection [of a message], not just the one which will actually be chosen since

Hiring and personnel committees are often accused of bias toward candidates from similar backgrounds. To combat this problem, many institutions have implemented policies such as documentation of hiring and promotion explanations and auditing of committee communications. In addition, anti-discrimination laws exist and committee communications can be subpoenaed. But our results imply these measures are unlikely to eliminate biased decisions. Transparency and greater stakeholder representation on boards are also concerns in the ongoing debate on corporate governance. As long as management controls information flows and the deliberative process, stakeholder representation by itself may not be enough to prevent management interests from being fully served.

Applied more broadly, the observer in our model could be the general public when their interests diverge from technocratic experts on a policy issue such as free trade, Covid, or climate change. Even if open government, sunshine laws and similar regulations force deliberations out from behind closed doors, careful technocrats can still promote their agenda, although the form of their communication may become more roundabout.[4] By controlling the process of information exchange, experts can manufacture consent and undermine the will of the majority. Viewed more positively, the ability of experts to engage in fact-based decision making is less affected by public interference than might be expected. Similarly, activists organizing under state surveillance can be successful if they are careful about how they speak.

We assume that the observer hears the unencrypted messages the players send to each other. In equilibrium, he fully understands the intended meaning of messages in the same way as the players. In practice, communication may be insecure because encryption might be broken; or the conversation must be in public and take place in ordinary language. Encryption is not necessary for our existence results since, unlike the cryptography literature, we assume some commonality of interest between all relevant parties. So our results are relevant to understanding secure versus insecure communication in many commercial, diplomatic, and national security contexts where interests are only partially opposed.

Subversion requires the "conflict" set $\mathcal{C}$ be no larger in measure than the "agree to accept" set $\mathcal{A}$ as in Figure 1. This *non-negative slack* condition, $\Pr[\mathcal{A}] \geq \Pr[\mathcal{C}]$, captures a commonality of interest with the observer. It is obtained by aggregating the deniability constraint the players face each time they take a decision. Under it, the observer would like to accept the proposal if all he knows is that the committee wants to accept it, although he may not always agree if he learns the exact reason why the committee wants to do so. With dispersed information the committee has to figure out its own reasons publicly, so ex ante commonality of interest with the observer is not sufficient for subversion. As they share information to determine whether they want to

this is unknown at the time of design." It does not matter for our results if the firm can commit to its plans. We model them as ex ante plans of action that must be interim incentive compatible (Green and Stokey, 2007).

[4]The convoluted patterns of bureaucratic communication have been satirized in fiction and film, e.g., in the BBC television sitcom *Yes Prime Minister*. As the principal character Sir Humphrey Appleby put it, in an episode titled Official Secrets, "The purpose of minutes is not to record events, it is to protect people."

accept or reject the project, the players also need to hide information to maintain non-negative slack and preserve the commonality of interest with the observer in every continuation game generated by the conversation.[5]

We show that in some games this need to maintain non-negative slack will make it impossible for the committee to take its ideal decisions with positive probability even if it deliberates for ever. We call such games conversational dead-ends. A necessary condition for subversion is to avoid a dead-end as a continuation game. A self-similar dead-end has the additional property that the players cannot even communicate any information to each other if they are to maintain non-negative slack. For such a game, any attempt at subversion results only in continuation games that are identical to the original game. In the paper we show that for finite type problems, there are three kinds of self-similar dead-ends, each a game with a binary type space for each player. Any other dead-end can be partitioned via a conversation into one or more of the three self-similar ones and so they are the key end games that prevent subversion in larger games with more types.

To identify necessary *and* sufficient conditions for existence of a subversive conversation, we restrict attention to finite environments with independent priors and utilize the characterization of bimartingales in Aumann and Hart (1986). Such a belief-based approach tracks how alternating statements by X and Y generate a bimartingale that describes updated beliefs on the x or y dimension respectively, and it has been used by Forges (1990a) and Aumann and Hart (2003) to show that multi-round communication can expand the equilibrium set in cheap talk games. Applying this approach to our problem, we show that if a conversation is subversive its associated bimartingale must converge to a set of terminal beliefs consistent with subversion. The convergence occurs if and only if priors lie in (a suitable version of) the biconvex hull of the set of subversive terminal beliefs. For a self-similar dead-end, subversion is impossible because the bimartingale will get stuck at the prior, as the committee must be uninformative if it is to maintain non-negative slack.

The rest of the paper is organized as follows. Section 2 sets up the model. In Section 3 we construct robust subversive conversations when there are two ex ante identical choices. We also establish invariance properties of subversive conversations that can be used to generalize these results. Section 4 investigates the limits of subversion. Conversational dead-ends are described in Section 4.1 and necessary and sufficient conditions for existence in Section 4.2. Section 4.3 discusses the interrelations between our results. Section 5 contains the literature review and Section 6 contains the concluding remarks. The Appendices contain proofs of results not contained in the main text as well as additional results.

[5]In contrast to our model of communication that is (or may become) public, if the players could guarantee completely secure communication then ex ante commonality of interest would be necessary *and* sufficient for subversion since they could take their ideal decision without revealing any additional information.

2 Players, preferences and information

A committee is composed of two players, X and Y (both "she"). Player X privately observes $x \in \mathcal{S}_X \subseteq \mathbb{R}$, while player Y privately observes $y \in \mathcal{S}_Y \subseteq \mathbb{R}$. Let $s = (x, y) \in \mathcal{S} = \mathcal{S}_X \times \mathcal{S}_Y$ denote the state of the world. We assume player types are independent. Let P and Q denote the cumulative distribution of x and y respectively, with $G = P \times Q$.

The two players have common interests and face a binary decision to either accept or reject a proposal. Their common payoff from rejecting the proposal is normalized to zero, while the payoff from accepting it equals $u(s) \in \mathbb{R}$. Let $\mathcal{R}$ be the (measurable) set of states where the committee prefers to reject the proposal, i.e., $u(s) < 0$ for $s \in \mathcal{R}$, with $u(s) > 0$ otherwise.[6]

The players communicate in discrete time, $t = 1, 2, ...$ If t is odd, player X may take a decision (accept or reject), or she may not (the "null" decision). She also sends a cheap talk message to the other player. Player Y does the same when t is even. We assume the set of possible messages M is rich enough to allow each player to reveal any measurable subset of her types. Let $m_t \in M$ denote a message sent in round t, and $d_t \in D = \{A, R, N\}$ denote, respectively, an accept, reject or null decision in round t. The game terminates as soon as a player takes a (non-null) decision and the payoffs of the players are then determined.[7]

Let $m^t \in M^t$ denote a history of messages, $d^t \in D^t$ a history of decisions, and $h^t = (m^t, d^t) \in H^t = M^t \times D^t$ a history, each of length t. Let $\omega_t \in \Omega = [0, 1]$ denote the draw of a uniformly distributed random variable, that describes the privately observed randomization by the player who moves in round t, and it is independent of randomizations in other rounds. A *protocol* $\xi \equiv \{\xi^t\}$ for the committee is defined by the maps $\xi^t \equiv (\sigma^t, \alpha^t) : \mathcal{S} \times H^{t-1} \times \Omega \to M \times D$, where ξ^t is measurable with respect to the information of the player who makes a move in round $t \in \mathbb{N}$. The protocol ξ has two components, a *conversation* $\sigma \equiv \{\sigma^t\}$ and an *action plan* $\alpha \equiv \{\alpha^t\}$. In each round t, the conversation specifies a (possibly random) message as a function of the state and history, $\sigma^t(s, h^{t-1}, \omega_t) \in M$, while the action plan specifies a (possibly random) decision, $\alpha^t(s, h^{t-1}, \omega_t) \in D$. Thus, $\xi = (\sigma, \alpha)$ is a behavior strategy profile, with $\xi_i = (\sigma_i, \alpha_i)$ the strategy of player $i \in \{X, Y\}$.

A protocol ξ together with the prior G gives rise to a probability distribution over histories. We will say ξ is *a.s.-finite* if it takes all its (non-null) decisions in finite time with probability one, from the perspective of each type of each player, i.e., Q-a.s. given any $x \in \mathcal{S}_X$ and P-a.s. given any $y \in \mathcal{S}_Y$. A protocol is *finite* if it is possible to specify in advance a round by which it takes all its decisions. We focus on *subversions*. A subversive protocol must be a.s.-finite and it

[6] We assume the players strictly prefer one action or the other in each state in order to ensure their ideal decision rule is unique. This decision is a collective action taken by the committee, although our results extend to some cases of individual decisions taken by each player, such as problems of pure coordination.

[7] We set payoffs to zero if neither player ever takes a decision. Instead of allowing either player to take the decision unilaterally, we could equally assume a particular player has decision rights, or allow a decision to be taken after both players vote in favor or ratify it, without altering anything substantive.

must implement the committee's first-best optimal decision rule:

$$\alpha(s, h^{t-1}, \omega_t) = R \Rightarrow s \in \mathcal{R},$$
$$\alpha(s, h^{t-1}, \omega_t) = A \Rightarrow s \in \mathcal{R}^c \equiv \mathcal{S} - \mathcal{R}.$$

Notice it does not matter for our results if the players can commit to a subversive protocol or not since neither player has an incentive to deviate from it.[8]

As described so far, it is easy to create a subversive protocol. Player X can reveal the value of x to Y who then knows the state $s = (x, y)$ and can take the committee-optimal decision. But we suppose that the players face a constraint. Their conversation and decision will be observed ex post (i.e., after a decision is taken but before payoffs are realized) by another agent who has a conflict of interest with the committee. We call this agent the observer ("he"). Because of the conflict of interest, the observer may object to or overrule the committee's decisions. The players must communicate in a manner that ensures the observer never objects. We call this constraint on the committee the deniability constraint and describe it now in more detail.

The observer's payoff when the proposal is rejected is normalized to zero. His payoff when it is accepted is $v(s) = 1$ if s belongs to some (measurable) set $\mathcal{A} \subseteq \mathcal{S}$, with $v(s) = -1$ otherwise. So the observer prefers to accept the proposal if $s \in \mathcal{A}$ and reject it otherwise. We assume $\mathcal{A} \subseteq \mathcal{R}^c$ so that whenever the observer wants to accept the proposal so does the committee.[9] Let $\mathcal{C} = \{\mathcal{A} \cup \mathcal{R}\}^c$ denote the conflict zone, the set of states where the committee prefers to accept the proposal but the observer does not. In contrast, the sets $\mathcal{A}$ and $\mathcal{R}$ denote the acceptance and rejection zones, where all parties agree on the decision. Since the extensive form of the communication game will be held fixed throughout, the acceptance, rejection and conflict sets together with the priors define a game $\Gamma = \{\mathcal{A}, \mathcal{C}, \mathcal{R}; P, Q\}$.

Fix a game Γ and suppose a protocol ξ is subversive. The observer will infer $s \in \mathcal{R}$ following a decision to reject. In this case all parties agree that rejection is best, so the observer never objects to such a decision. But if the committee takes a decision $d_t = A$ after some history of communication then the observer learns $s \in \mathcal{A} \cup \mathcal{C} = \mathcal{R}^c$. So he may object to the decision if he assigns a large enough likelihood to the event $s \in \mathcal{C}$ where he prefers to reject the proposal.

Let $H_A^t(\xi) = \{h^t \in H^t \mid d_t = A,\ d_{t'} = N,\ t' < t\}$ be the set of t-round histories that can be generated by subversive protocol ξ and that terminate in a decision to accept. The observer will not object to the decision if and only if $\Pr[\mathcal{A} \mid h^t] \geq 1/2$, $h^t \in H_A^t(\xi)$. Since in a subversion a decision to accept leads the observer to infer that $s \in \mathcal{A} \cup \mathcal{C}$, we can rewrite this inequality as

$$\Pr[\mathcal{A} \mid h^t, \mathcal{A} \cup \mathcal{C}] \geq \Pr[\mathcal{C} \mid h^t, \mathcal{A} \cup \mathcal{C}], \quad h^t \in H_A^t(\xi). \tag{DC}$$

When the committee accepts the proposal after a history h^t, the deniability constraint (DC)

[8] Our default assumption is of no commitment and our equilibrium notion is Bayesian Nash equilibrium.

[9] We relax this assumption in Section B.3.

says the observer thinks it is (weakly) more likely that the true state belongs to $\mathcal{A}$ as opposed to $\mathcal{C}$ and so he would also prefer acceptance.[10]

Definition A *protocol* $\xi = (\sigma, \alpha)$ is subversive if it is a.s.-finite and implements the committee's optimal decision rule while satisfying the deniability constraint (DC). A *conversation* σ is subversive if (σ, α) is a subversive protocol for some action plan α.

We close this section with a few remarks. Any history h^t generated by a subversive protocol ξ defines a continuation game. In this continuation game, the residual part of $\mathcal{A} \cup \mathcal{C}$, after deleting any states ruled out by h^t, may be subsets of $\mathbb{R}^1$, or even finite, and it may not be possible to use Bayes' Rule to evaluate (DC). In such cases, we use (generalized) probability densities to compute posteriors. In addition, a continuum of types can create measure theoretic paradoxes that contradict the law of iterated expectations. To rule these out, we impose the following *admissibility restriction* on a subversive protocol ξ: the deniability constraint (DC) must hold not only for each element of $H_A^t(\xi)$ but also when we integrate over all histories that belong to any measurable subset of $H_A^t(\xi)$.

If a subversive protocol exists, then by the law of iterated expectations, $\Pr[\mathcal{A} \mid h^t] \geq \Pr[\mathcal{C} \mid h^t]$ after every h^t. This implies we must have $\Pr[\mathcal{A}] \geq \Pr[\mathcal{C}]$, an ex ante *non-negative slack* condition that is necessary for the committee to be able to subvert. It also implies the observer has no incentive to object after histories where the committee has not yet made a decision. So if the committee can subvert under ex post scrutiny, it can also subvert when its deliberations are observed contemporaneously. Indeed, we can allow the observer (and not the committee) to have formal authority over decisions, with the committee simply recommending a decision.

We have in mind situations where the observer's role is passive and the communication between the players can only be scrutinized after the fact by the observer. He cannot influence the procedural rules of committee deliberations (e.g., restrict the length of communication, constrain the message space, or interject). The players are free to design the procedural rules, or these rules are given by tradition. They also have the freedom to choose the equilibrium protocol, subject only to the deniability constraint.[11] This constraint could itself be a primitive of the model. It could arise out of cultural, social or psychological norms, and represent the players' own desire to avoid scandal, outrage or being seen to act against the public interest.[12]

[10] The deniability constraint is similar to the "balance of probabilities" burden of proof faced by courts in U.S. civil cases. When (DC) is met, the balance of probabilities favors "acquittal", i.e., allowing the committee to accept the proposal. So the committee will also be acquitted under the more demanding "reasonable doubt" burden of proof used for criminal trials.

[11] In the game where the observer actively takes decisions, this amounts to selecting the equilibrium (if one exists) in which the observer accepts the proposal whenever indifferent, and the players implement their ideal decision rule. Other equilibria exist but we focus on subversive equilibria.

[12] To understand what can be implemented when the observer (or a social planner) has full control over communication design, one can use direct, truth-telling mechanisms. But the Revelation Principle cannot be used if mechanisms are constrained to limit the amount of information revealed publicly. We provide one reason for such constraints (namely, deniability); and focus on indirect mechanisms that all implement the committee's

3 Constructing subversions

We begin by considering two benchmark models of natural interest similar to the introductory example of Figure 1. In the first benchmark, the committee is biased towards one of the two choices while the observer is impartial and prefers whichever choice is better. In the second benchmark, the observer is biased against one of the two choices while the committee is impartial and meritocratic. As will become clear, the two models are not mirror-images of each other. For each model we construct a subversive conversation that allows the committee to take its ideal decision while meeting deniability. We also show that these conversations need not change under changes to committee preferences or uncertainty about them by providing general invariance properties of a subversive conversation.

3.1 Biased committee, impartial observer

Let $\mathcal{A} = \mathcal{L} \equiv \{(x, y) \in \mathcal{S} \mid y \leq x\}$ be the set of states where both the observer and the committee prefer to accept the project, $\mathcal{R} \subseteq \mathcal{L}^c$ the set of states where they both prefer to reject it, and $\mathcal{C} \subseteq \mathcal{L}^c$ the zone of conflict where the committee prefers to accept but the observer prefers to reject, with $\mathcal{R} \cup \mathcal{C} = \mathcal{L}^c$. Since the observer prefers whichever choice is better, but the committee is biased towards one of the two choices, we call this the biased committee model.

We suppose x and y are iid with common cdf $P = Q$ that is continuous and increasing (i.e., invertible). Using a one-to-one transformation of x and y to their quantiles, we can restrict attention to uniform priors on $\mathcal{S} = [0, 1]^2$ and interpret quantiles as the actual types. Notice that $\mathcal{A} = \mathcal{L}$ in this transformed space and we still have an instance of the biased committee model. We show later how our results extend to priors that are not continuous, or identical, or independent. For now, given uniform priors, a biased committee game Γ is fully specified by preferences that pin down the sets $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$, with $\mathcal{A} = \mathcal{L}$. We have the following result.

Proposition 1 *There exists a conversation that is subversive for every biased committee game.*

Figure 2 illustrates Proposition 1. Priors are uniform without loss of generality. Panel (a) depicts the sets $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$. The remaining panels describe a subversive conversation. In round 1 of this conversation, X sends one of two messages. She says "*in*" when $x \in [1/4, 3/4]$ and "*out*" when $x \notin [1/4, 3/4]$. Player Y does the same in round 2—she says "*in*" when $y \in [1/4, 3/4]$ and "*out*" otherwise. The four elements of the partition of the state space created by the first two messages are depicted in panels (b) through (e).

When the two players send different messages in the first two rounds, the player who said *in* reveals her exact type, following which the other player takes the committee's optimal decision. As shown in panels (b) and (c), the deniability constraint (DC) is met each time the committee

optimal decision rule, identifying the ones that maintain deniability.

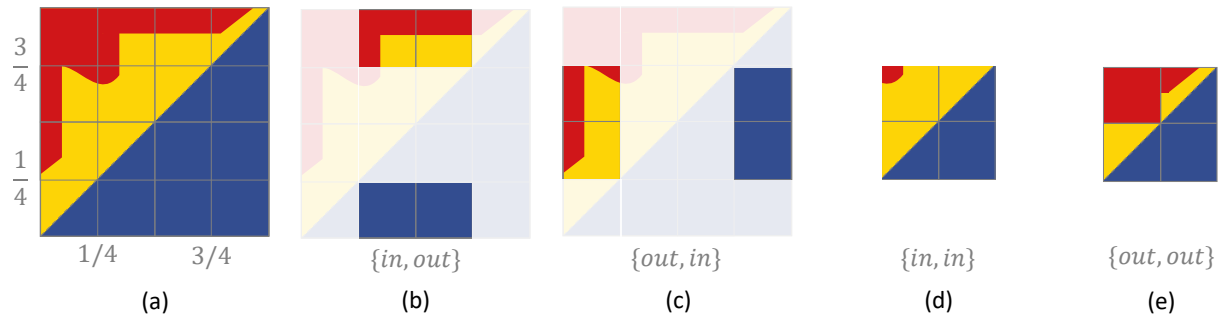

Figure 2: Recursive conversation for the biased committee model

accepts the proposal—the residual measure of $\mathcal{A}$ is at least as large as the residual measure of $\mathcal{C}$, given the observed history of messages, including the revealed type.

Notice the importance of the history of prior messages that creates a suitable context before a player can safely take a decision and ensure deniability. For instance, if X revealed $x = 1/4$ at the beginning of the conversation, (DC) would not be met if Y subsequently accepted the proposal. But it is safe for X to reveal such a type after Y reveals $y \notin [1/4, 3/4]$. The same is true for Y when she reveals $y = 3/4$. To ensure deniability, no player reveals very unfavorable news in the first two rounds of the conversation. Each player also conceals some favorable news initially in order to compensate for any unfavorable news that may be revealed by her partner later in the conversation. This gradual process of creating and refining suitable contexts creates slack in the deniability constraint where none existed initially.

Panels (d) and (e) depict the two remaining continuation games, when the first two messages are identical. The residual state space in each game is itself an instance of the biased committee model, if we paste the components together and rescale. We proceed recursively in each of these two biased committee continuation games by supposing that the conversation restarts in the manner described above. This recursion results in committee's optimal decisions taken a.s. in finite time, conditional on each type of each player. We have found a subversive conversation.

Notice that this argument does not depend on the particular properties of the game depicted in Figure 2(a). Because the sets $\mathcal{C}$ and $\mathcal{R}$ that pin down the committee's preferences vary from game to game, the committee needs to adapt its decisions to each particular game. But the conversation that precedes the decision can be designed in advance. It does not vary from game to game. Thus, Proposition 1 asserts the existence of a *robust* subversive conversation which applies to every instance of the biased committee model, i.e., to every specification of the sets $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$ with $\mathcal{A} = \mathcal{L}$, and every specification of the iid priors $P = Q$. Since the committee may encounter similar games frequently, this robustness feature simplifies communication design within organizations. In addition, if the committee's preferences are not common knowledge, a robust conversation allows subversion regardless of observer beliefs about the committee's actual

preferences.[13]

The conversation of Figure 2 is robust because the player who takes the decision necessarily knows both x and y. The other player reveals her exact type in the conversation that precedes the decision. Call a conversation *fine* if it always results in the player taking the decision being fully informed. A subversive protocol (σ, α) is a *fine subversion* if σ is a fine subversive conversation.

Subset property. A fine subversive conversation for a game $\Gamma = \{\mathcal{A}, \mathcal{C}, \mathcal{R}; P, Q\}$ is also a fine subversive conversation for any other game $\Gamma' = \{\mathcal{A}', \mathcal{C}', \mathcal{R}'; P, Q\}$ with $\mathcal{A}' \supseteq \mathcal{A}$, $\mathcal{R}' \supseteq \mathcal{R}$ and $\mathcal{C}' \subseteq \mathcal{C}$.

The subset property follows immediately from the deniability constraint (DC) and the fact that subset relations are preserved after intersections with the residual set of states $\mathcal{S}\left(h^t\right)$ reached after any history h^t generated by the fixed conversation. Since the player who takes the decisions is fully informed in a fine subversion, she can always tailor her decision to suit her preferences. So (DC) will be met in Γ' when she takes a decision, as long as it was met in Γ.

The subset property allows us to construct a fine subversion for a "worst case" and use it to extend existence to other scenarios, yielding a robust subversion. We employ this worst case approach again in the next section to construct a robust subversive conversation for the biased observer model and subsequently generalize the subset property in Section 3.3. Note for now that the subset property also allows us to relax our assumptions on the common prior and allow for atoms. When the common cdf $P = Q$ has atoms, we can still use the quantile function and work with quantiles as the underlying types, except that now a range of quantiles map into the same type. But this is immaterial since all these quantiles have the same payoff as the original type they map into. The only material difference that arises from allowing atomic types is that now we can have $\mathcal{A} \supseteq \mathcal{L}$ after the transformation to quantiles. By the subset property, the conversation described in Figure 2 also works for this generalized biased committee model that includes the finite type case.[14]

3.2 Biased observer, impartial committee

We now consider a model where the committee is impartial and prefers the better one of the two choices but the observer is biased against one choice. To this end, suppose $\mathcal{S} = [0,1]^2$ and let $\mathcal{R} = \mathcal{U} \equiv \{(x,y) \in \mathcal{S} \mid y \geq x\}$. Further, let $\mathcal{A} \supseteq \mathcal{L}_b \equiv \{(x,y) \in \mathcal{S} \mid y \leq x - b\}$ and $\mathcal{C} = \{\mathcal{A} \cup \mathcal{R}\}^c$. The parameter $b > 0$ represents a bound on the bias of the observer. Assume the common cdf $P = Q$ is uniform and that $b \equiv 1 - 1/\sqrt{2}$, which ensures non-negative slack (a necessary condition for subversion) in the worst case $\mathcal{A} = \mathcal{L}_b$ that is depicted in panel (a) of Figure 3.[15]

[13] Robust subversive conversations may not be unique. See Section B.1 for one that is finite.

[14] See Section B.2 for the results under non-iid priors.

[15] Like in the model of Section 3.1, uniform priors could be interpreted as the outcome of a quantile transformation of an invertible common cdf $P = Q$. The non-negative slack condition we employ translates to the restriction that the probability that $s \in \mathcal{C}$ is bounded above by b, conditional on each x and on each y.

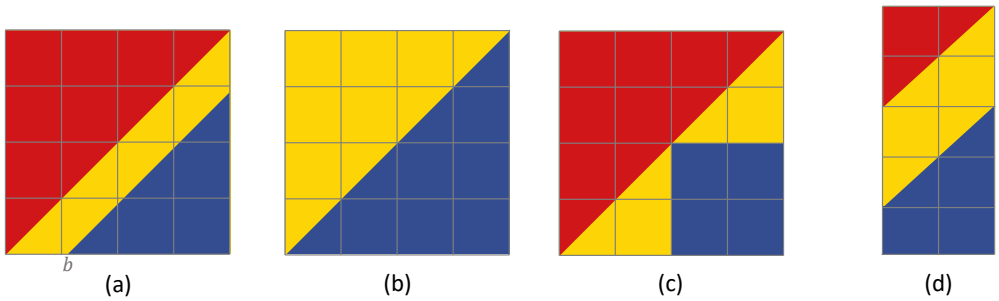

Figure 3: Biased observer model (a) and continuation games (b)-(d)

Call the set of games with these properties the biased observer model. Notice that in the absence of any information from the committee, the biased observer would like to reject the proposal, a distinction from the biased committee model where he would be willing to accept it.[16] In spite of this ex ante skepticism on the part of the observer, we have the following result.

Proposition 2 *There exists a conversation that is subversive for every biased observer game.*

In the proof of Proposition 2, we construct a fine subversion for the worst case $\mathcal{A} = \mathcal{L}_b$ which must then be a robust subversion for the biased observer model, via the subset property. As shown in the Appendix, the conversation we construct yields a set of continuation games where decisions can be taken immediately, together with three other continuation games that are depicted in panels (b) through (d) of Figure 3. Panel (b) of Figure 3 is a biased committee game Γ_b, that was solved recursively via the conversation depicted in Figure 2. The game Γ_c in panel (c) captures another natural economic situation—the committee wants to accept the proposal if the benefit x exceeds the cost y, while the observer wants to accept it only if the benefit and the cost are each better than average. The game Γ_d depicted in panel (d) is an asymmetric game where the committee is more in favor of accepting the proposal than the observer but all parties are willing to make tradeoffs. As shown in the proof of Proposition 2, the continuation games of panels (c) and (d) can also be solved via recursive constructions similar to the one described in Figure 2.

3.3 Invariance properties of subversive conversations

Propositions 1 and 2 show that the same conversation can be subversive in a range of different games, i.e., the conversation is invariant to the particular game. Invariance is important because it extends existence results established for one game to others. The transformation to quantiles is an invariance property because the conversation can always be in terms of quantiles. This

[16]The biased observer model would be an equivalent mirror-image version of the biased committee model (via switching the names of the accept and reject decisions) if, for $s \in \mathcal{C}$, the committee preferred to reject the proposal while the observer preferred to accept it. But we assume the opposite. The structure of subversive conversations is different across the two models.

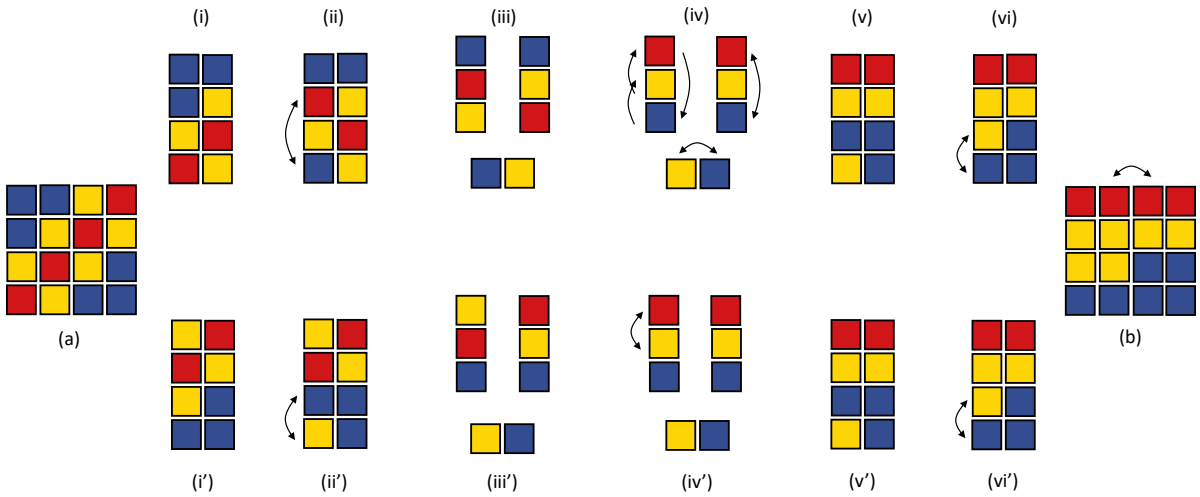

Figure 4: Relabeling property

is an example of a more general *relabeling property* that we define below. Robustness is a second reason that invariance properties are important. For instance, the subset property of fine subversions established the existence of robust subversions in the models of Sections 3.1 and 3.2. This is an example of a more general invariance property that we call *decision-measurability*.

Panels (a) and (b) of Figure 4 depict two games with priors assumed to be uniform. Panel (b) is similar to the models we have considered so far. The two signals x and y can be interpreted as a benefit and a cost of accepting the proposal. A high benefit can compensate for a high cost and make the proposal acceptable. The game in panel (a) depicts a different case where the magnitudes of x and y matter less, and whether or not the signals "confirm" each other matters more. All parties prefer to reject when the two signals match and prefer to accept when the signals are sufficiently dissimilar.[17]

Figure 4 shows how a subversive conversation for the game in panel (a), depicted by the vertical and horizontal partitioning of the state space in the subsequent panels, is also a subversive conversation for the game in panel (b), after some permutations of rows and columns in each continuation game created by the conversation. These permutations are inessential relabelings of the residual type space at every continuation game generated by the fixed conversation.

Starting from the game in panel (a), panels (i) and (i') show the two continuation games created by X in round 1 when she partitions her type space into the two left columns and the two right ones. In panels (ii) and (ii') we perform a row permutation in each continuation game. Panel (iii) and (iii') show the next two moves where Y first partitions her type space into the bottom row and its complement, and in the latter case X follows by partitioning her residual type space into the left and right columns. Panels (iv) and (iv') depict some permutations of the

[17]For instance, a defendant could be on trial, with matching signals indicating that his alibis check out and he should be acquitted. The committee prefers to convict if there is any mismatch, while the observer prefers to do so only if the mismatch is sufficiently large.

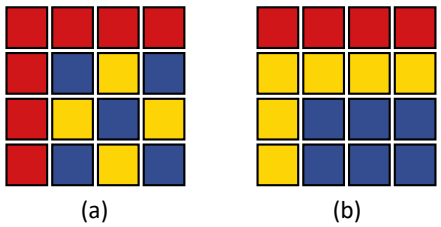

Figure 5: Decision-measurability property

residual type space at this stage for each continuation game. At this point one player can reveal her exact type and the other can take the committee's ideal decision. In the remaining panels we put together these continuation games, performing some more permutations at each stage, to obtain the game in panel (b). This game shares the same subversive conversation, subject to relabelings of the type space at each continuation game generated by the fixed conversation, as shown in the figure. We now formally define the relabeling property.

Fix a subversive conversation σ for a game $\Gamma = \{\mathcal{A}, \mathcal{C}, \mathcal{R}; P, Q\}$. Let $\mathcal{S}_i\left(h^t\right)$ denote the residual type space of player $i = X, Y$, generated by σ after history h^t, with $\mathcal{S}\left(h^t\right) = \mathcal{S}_X\left(h^t\right) \times \mathcal{S}_Y\left(h^t\right)$. This defines a continuation game $\Gamma\left(h^t\right)$, using intersections of $\mathcal{S}\left(h^t\right)$ with the sets $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$, and posteriors derived from P, Q using σ and Bayes' Rule.

Fix $\Gamma\left(h^t\right)$. An *admissible relabeling* of $\mathcal{S}_i\left(h^t\right)$ is a measure-preserving bijection $\rho_i(\cdot|h^t) : \mathcal{S}_i\left(h^t\right) \to \widehat{\mathcal{S}}_i\left(h^t\right) \subseteq \mathbb{R}$ whose inverse is also measure-preserving.[18] The set $\widehat{\mathcal{S}}_i\left(h^t\right)$ must have the same cardinality as $\mathcal{S}_i\left(h^t\right)$. If $\widehat{\mathcal{S}}_i\left(h^t\right) = \mathcal{S}_i\left(h^t\right)$ the relabeling delivers a *permutation* of $\mathcal{S}_i\left(h^t\right)$. Figure 4 is a discrete type example of such permutations. If $\widehat{\mathcal{S}}_i\left(h^t\right) \neq \mathcal{S}_i\left(h^t\right)$, $\rho_i(\cdot|m^t)$ is a *rescaling* of $\mathcal{S}_i\left(m^t\right)$. As mentioned before, when the priors are invertible, an example of such a rescaling is a transformation to quantiles. The rescalings employed in Figures 2 through 13 are other examples.

Relabeling property. A subversive conversation σ for Γ is also subversive for Γ' obtained from Γ via relabelings $\rho_i(\cdot|h^t)$ of $\mathcal{S}_i\left(h^t\right)$, $i = X, Y$, at every history h^t generated by σ.

Continuing to assume uniform priors, now consider Figure 5(a) which admits a subversive conversation where Y moves first and rejects the proposal if her type corresponds to the top row and otherwise turns the conversation over to X, who then accepts or rejects the proposal to attain the committee's ideal outcome. Looking at Figure 5(b), this same conversation works, although the action plan is different since X never rejects. Indeed, if we look back at Figure 4(b), the same conversation is subversive there as well. These examples illustrate the decision-measurability property that we now define formally.

Fix a subversive conversation σ for a game $\Gamma = \{\mathcal{A}, \mathcal{C}, \mathcal{R}; P, Q\}$. Let $\mathcal{S}\left(h^t, d_{t+1}\right)$ be the residual state space after a history h^t that is followed by a decision $d_{t+1} \in \{A, R\}$ and consider

[18] Since σ is fixed, we do not denote the dependence of $\mathcal{S}_i\left(h^t\right)$ (or $\rho_i(\cdot|h^t)$) on σ, in order to avoid clutter.

any other game $\Gamma = \{\mathcal{A}', \mathcal{C}', \mathcal{R}'; P', Q'\}$ with the same state space $\mathcal{S}$.

Decision-measurability property. A subversive conversation σ for Γ is also subversive for Γ' if at every history h^t followed by a decision $d_{t+1} \in \{A, R\}$, (i) $\mathcal{R}' \cap \mathcal{S}\left(h^t, d_{t+1}\right)$ is measurable with respect to the information set of the player who takes the decision in round $t+1$ and (ii) $\mathcal{A}' \cap \mathcal{S}\left(h^t, d_{t+1}\right)$ is at least as large in measure (induced by $\{P', Q'\}$) as $\mathcal{C}' \cap \mathcal{S}\left(h^t, d_{t+1}\right)$.

The decision-measurability property allows us to modify the residual state space $\mathcal{S}\left(h^t, d_{t+1}\right)$ in a way that permits the player taking the decision to (i) still be able to determine the committee-optimal decision and (ii) still meet the deniability constraint (DC) if she accepts the proposal. The decision(s) are allowed to be different across the two games but the conversation that precedes each decision is unchanged. It generalizes the subset property of fine subversive conversations. Since the player who takes the decision is fully informed in a fine subversion, condition (i) is automatically satisfied, while the subset conditions $\mathcal{A}' \supseteq \mathcal{A}$ and $\mathcal{C}' \subseteq \mathcal{C}$ are sufficient (but not necessary) for condition (ii). More generally, the decision stages of a fine subversion, taken together, partition the state space into one-dimensional sets. Decision-measurability allows us to modify preferences within each of these sets in any way we wish, subject only to meeting condition (ii). The same conversation will remain subversive after the modifications.

The decision measurability property also implies that the set of subversive conversations is invariant to switching the sets $\mathcal{A}$ and $\mathcal{C}$ in games where $\Pr[\mathcal{A}] = \Pr[\mathcal{C}]$, i.e., there is zero ex ante slack. We will refer to this as the recoloring property. In addition, the set of subversive conversations is invariant to interchanging the roles of the two players (i.e., rotations of the state space) as well as the cardinal specification of the players' payoffs. We apply the invariance properties throughout the paper. In particular, we used them to establish the existence of robust subversive conversations in the biased committee and and biased observer models of Section 3.

4 Limits of subversion

We turn now to analyzing the boundaries of where subversion is possible and where it is not. We first characterize conversational dead-ends. Next we employ the belief based approach of Aumann and Hart (1986) to provide a necessary and sufficient condition for existence in finite environments. Finally, we provide a discussion that links the belief-based approach to the invariance properties of a subversive conversation.

4.1 Conversational dead-ends

A necessary condition for subversion is that the players maintain non-negative slack not only ex ante but in every continuation game. A dead-end is a game where this requirement makes it

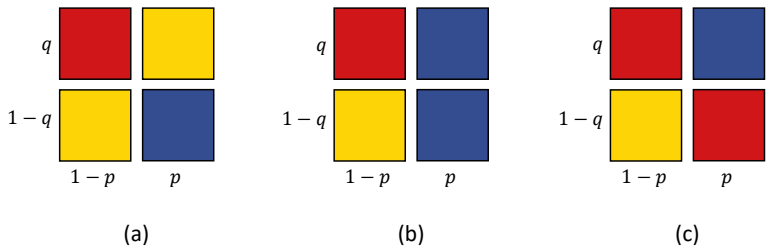

Figure 6: All self-similar dead-ends

impossible for the committee to take its ideal decisions with positive probability. The committee must avoid a dead-end as a continuation game in order to subvert.

A self-similar dead-end has the additional property that neither player can even convey any information to the other. The committee can deliberate for ever but not make any progress towards subversion. Figure 6 shows three binary type games, each with interior priors $p, q \in (0, 1)$ such that we have zero slack, $\Pr[\mathcal{A}] = \Pr[\mathcal{C}]$. We show in this section that subversion is impossible in each of these three games. In fact, they describe all possible self-similar dead-ends, key end games that have to be avoided for subversion to be possible in any larger game.

Consider the game in Figure 6(a) first. In this game, the players favor accepting the project as long as either the benefit x is high or the cost y is low. The observer requires both the benefit to be high and the cost to be low. Since there is zero total slack by assumption, and every continuation game must have non-negative slack for subversion to be possible, using the law of iterated expectations we see that in any subversive protocol any move (message or decision) by X in round 1 must create a continuation game with zero slack. Using this, we show via contradiction that subversion is impossible in Figure 6(a). In fact, it is a self-similar dead-end.

Notice first that in any subversive protocol X cannot cannot take her ideal decision with positive probability in round 1 when she knows the benefit is low and the state is in the left column, because she does not know her ideal decision in this case. She only knows her ideal decision (to accept the proposal) when the benefit is high. Conditional on this type there is positive slack. So, if X accepts the proposal with positive probability in round 1 when x is high, she leaves negative slack in at least some continuation games that arise after she does not take a decision. Subversion is impossible in such a continuation game, implying X cannot take a decision in round 1. She can only send messages that maintain zero slack in every continuation game. This is possible only if every message sent by X in round 1 is sent by both of her types with the same probability, i.e., X can only be uninformative in round 1. Since the game is symmetric, Y faces the same situation in round 2, following any uninformative round 1 message, and so she must also be uninformative in round 2. Continuing this logic, we conclude that neither player can ever take their ideal decision, or even convey any information to the other, and so Figure 6(a) is a self-similar dead-end.

For the game in Figure 6(b) the observer's preferred decision is not affected by Player Y's information, whereas in Figure 6(c) the players prefer acceptance when their signals match but the observer only prefers acceptance when the matching signals take a particular value. Identical arguments to those used above can be used to show that both of these games are also self-similar dead-ends when the interior priors p and q are such that there there is zero slack. The next result shows that in finite type environments the three games of Figure 6 describe, essentially, all possible dead-ends.[19]

Proposition 3 *Any finite type game Γ that is a dead-end has zero slack and can be partitioned via a conversation into self-similar dead-ends. There are three kinds of self-similar dead-ends, each a binary type game with interior priors, zero slack and preferences as in Figure 6.*

4.2 Necessary and sufficient condition for existence

We now provide necessary and sufficient conditions on priors for the existence of subversive conversations, using the techniques of Aumann and Hart (1986). To do so, we restrict attention to *finite environments*, i.e., those where the type spaces $\mathcal{S}_X$ and $\mathcal{S}_Y$ as well as the message space M are all finite. Let p and q denote the priors on x and y derived, respectively, from the cumulative distributions P and Q.

Let $\mu = (\mu_X, \mu_Y)$ be a typical element of $\Delta(\mathcal{S}_X) \times \Delta(\mathcal{S}_Y)$. Given any protocol $\xi = (\sigma, \alpha)$, let $\mu(h^t) = (\mu_X(h^t), \mu_Y(h^t)) \in \Delta(\mathcal{S}_X) \times \Delta(\mathcal{S}_Y)$ denote the observer's posterior beliefs over states, derived using Bayes Rule from a history h^t generated by ξ. Let $\widetilde{\mu}^t = (\widetilde{\mu}^t_X, \widetilde{\mu}^t_Y) \in \Delta[\Delta(\mathcal{S}_X) \times \Delta(\mathcal{S}_Y)]$ be the random variable (with realizations $\mu(h^t)$, $h^t \in H^t$) that describes the possible round t beliefs of the observer.

By the law of iterated expectations, $\widetilde{\mu} \equiv \{\widetilde{\mu}^t\}_{t\in\mathbb{N}}$ is a (bounded) martingale that has expectation equal to the prior $\mu^0 \equiv (p, q)$. Further, since X speaks only in odd rounds, and Y only in even rounds, $\widetilde{\mu}^t_X = \widetilde{\mu}^{t-1}_X$ a.s. for t even, and $\widetilde{\mu}^t_Y = \widetilde{\mu}^{t-1}_Y$ a.s. for t odd, $t \geq 1$. Thus, $\widetilde{\mu}$ is a bimartingale (Aumann and Hart, 1986).[20] Conversely, given a bimartingale $\widetilde{\mu}$ with expectation μ^0, one can derive a protocol $\xi = (\sigma, \alpha)$ from it, using Bayes Rule recursively.[21]

For each $s \in \mathcal{S}$, define

$$\mathcal{T}(s) = \begin{cases} \{\mu \in \Delta(\mathcal{S}_X) \times \Delta(\mathcal{S}_Y) \mid \mu[s] > 0,\ \mu[\mathcal{R}] = 1\} & \text{if } s \in \mathcal{R}, \\ \{\mu \in \Delta(\mathcal{S}_X) \times \Delta(\mathcal{S}_Y) \mid \mu[s] > 0,\ \mu[\mathcal{R}] = 0,\ \mu[\mathcal{A}] \geq 1/2\} & \text{if } s \in \mathcal{R}^c. \end{cases}$$

The set $\mathcal{T}(s)$ describes the possible posterior beliefs of the observer at the time a decision is

[19] These are the only kinds of self-similar dead-ends up to the invariance properties of Section 3.3. Other tight games obtained by row/column permutations, rotations and recoloring of these three are also self-similar dead-ends, and for the same reasons.

[20] If the protocol takes a non-null decision $d_t \in \{A, R\}$ in round t after some history h^{t-1}, with $d_{t'} = N$ for all $t' < t$, we assume the bimartingale $\widetilde{\mu}$ is constant afterwards, i.e., equals $\mu^t(\{h^{t-1}, m_t, d_t\})$ forever after.

[21] This pins down ξ only on the path of play, which is all that is necessary given our focus on subversion.

taken, given that the committee has played according to some subversive protocol ξ, and given that the realized state is s. Since beliefs are correct, we must have $\mu[s] > 0$. Since the protocol is subversive, $\mu[\mathcal{R}] = 1$ when $s \in \mathcal{R}$ and the decision is to reject the proposal; while $\mu[\mathcal{R}] = 0$ when $s \in \mathcal{R}^c$ and the decision is to accept the proposal, with $\mu[\mathcal{A}] \geq 1/2$ in order to meet the deniability constraint. We can assume, without loss of generality, that each $\mathcal{T}(s)$ is non-empty.[22]

Let $\mathcal{T} = \cup_{s \in \mathcal{S}} \mathcal{T}(s)$. If a bimartingale $\widetilde{\mu}$ with expectation μ^0 is derived from a subversive protocol ξ, then its limiting distribution must belong to $\mathcal{T}$ and the limit is reached a.s. in finite time. Conversely, if a bimartingale $\widetilde{\mu}$ with expectation μ^0 has a limiting distribution in $\mathcal{T}$ that is reached a.s. in finite time, then one can construct a subversive protocol from it, as we show below. Our objective is to characterize the priors μ^0 that can give rise to a subversive protocol represented by a bimartingale $\widetilde{\mu}$ with expectation μ^0.

To this end, we recall some definitions from Aumann and Hart (1986). Call $B \subset \Delta(\mathcal{S}_X) \times \Delta(\mathcal{S}_Y)$ a biconvex set if each of its μ_X- and μ_Y-sections is a convex set. The biconvex hull of $B \subset \Delta(\mathcal{S}_X) \times \Delta(\mathcal{S}_Y)$, denoted by $bico(B)$, is the smallest biconvex set containing B. A real valued function $f(\mu_X, \mu_Y)$ defined on a biconvex set $B \subset \Delta(\mathcal{S}_X) \times \Delta(\mathcal{S}_Y)$ is a biconvex function if it is convex in each argument μ_X and μ_Y separately. Given $Z \subset B \subset \Delta(\mathcal{S}_X) \times \Delta(\mathcal{S}_Y)$, with B biconvex, the point $a \in B$ is (strongly) separated from Z with respect to B if there is a bounded biconvex function f on B such that $f(a) > \sup\{f(z) \mid z \in Z\}$. Denote by $ns_Z(B)$ the set of points $a \in B$ that cannot be separated from Z by any biconvex function. Let $bico^{\#}(\mathcal{T})$ denote the largest (in terms of set inclusion) set $B \subset \Delta(\mathcal{S}_X) \times \Delta(\mathcal{S}_Y)$ that satisfies $B = ns_{\mathcal{T}}(B)$. Using Theorem 4.3 in Aumann and Hart (1986), we have the following necessary and sufficient condition for the existence of a subversive conversation.

Proposition 4 *Consider a finite environment and fix $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$. A subversive protocol exists if and only if the prior $\mu^0 \in bico^{\#}(\mathcal{T})$.*

Figure 7 depicts the three examples of Figure 6 that we will use to illustrate Proposition 4. The pair $(p, q) \in [0, 1]^2$ pins down any posterior belief $\mu \in \Delta(\mathcal{S}_X) \times \Delta(\mathcal{S}_Y)$, including those with positive (or negative) slack, as depicted in panel (b). Panel (c) depicts the set of possible subversive terminal beliefs $\mathcal{T}(s)$, $s \in \mathcal{S}$, for each example of panel (a).

To see how these terminal belief sets are constructed, consider example (i) of Figure 7 first. For $s \in \mathcal{R}$, $\mathcal{T}(s)$ must equal the singleton point $(p, q) = (0, 1)$, since $\mu[\mathcal{R}] = 1$ for any $\mu \in \mathcal{T}(s)$. For $s \in \mathcal{R}^c$, we must have $\mu[s] > 0$, $\mu[\mathcal{R}] = 0$ and $\mu[\mathcal{A}] \geq 1/2$ for any $\mu \in \mathcal{T}(s)$. Since μ is a product measure, this implies that for $s \in \mathcal{C}$ that is in the bottom row, any $\mu \in \mathcal{T}(s)$ must attach zero probability to the top row and at least $1/2$ probability to the right column, i.e., it must be of the form $(p, 0)$ with $p \in [1/2, 1)$, as depicted in the figure. Similarly, for $s \in \mathcal{C}$ that

[22] If subversion is possible, $\mathcal{T}(s)$ cannot be empty for any s on which priors put positive weight. So we can focus on priors that put zero weight on either the row or column containing s. If $\mathcal{T}(s)$ is empty, we must have $s \in \mathcal{C}$ and there cannot exist a product set containing s that has a non-empty intersection with $\mathcal{A}$ but not $\mathcal{R}$.

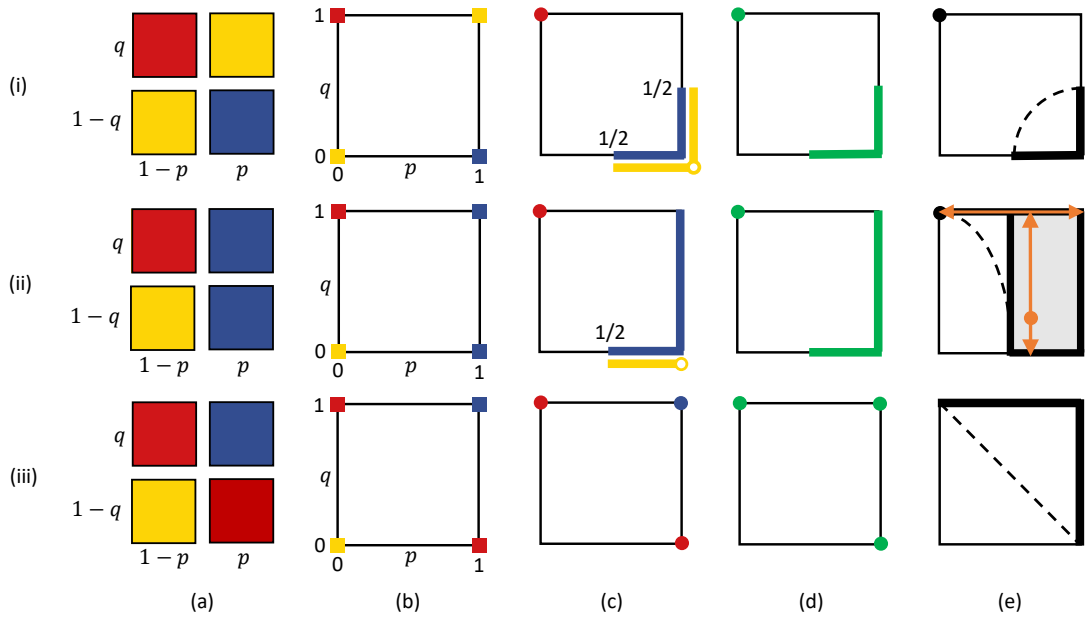

Figure 7: Biconvex hulls of subversive terminal beliefs

is in the top row, $\mu \in \mathcal{T}(s)$ must be of the form $(1, q)$ with $q \in (0, 1/2]$. For the remaining state $s \in \mathcal{A} \subset \mathcal{R}^c$ in the bottom right, $\mu \in \mathcal{T}(s)$ must either be of the form $(p, 0)$ with $p \in [1/2, 1]$, or of the form $(1, q)$ with $q \in [0, 1/2]$, in order to attach zero probability to $\mathcal{R}$ and at least one-half probability to $\mathcal{A}$. The terminal belief sets $\mathcal{T}(s)$ for examples (ii) and (iii) are constructed analogously and depicted in the figure. For completeness, we list these sets in Appendix B.4.

Panel (d) depicts the union $\mathcal{T} = \cup_{s \in \mathcal{S}} \mathcal{T}(s)$, while panel (e) depicts its biconvex hull $bico(\mathcal{T})$, for each of the three examples.[23] In general, $bico(\mathcal{T}) \subseteq bico^{\#}(\mathcal{T})$, but for these examples $bico^{\#}(\mathcal{T}) = bico(\mathcal{T})$.[24] By Proposition 4, a subversive conversation exists for these examples if and only if $\mu^0 \in bico^{\#}(\mathcal{T}) = bico(\mathcal{T})$.

For the top example (i) of Figure 7, notice that $bico^{\#}(\mathcal{T}) = \mathcal{T}$ and so subversion is impossible for any interior priors $(p, q) \in (0, 1)^2$. The dotted curve in panel (e) depicts interior priors for which there is zero slack and the problem is a self-similar dead-end, as shown by Proposition 3. Starting from any interior point with arbitrarily large positive slack that lies below this zero slack curve, the committee can accept the proposal for a subset of the states, while maintaining non-negative slack in the complementary set, approaching (or reaching) some point on the zero

[23] The set $bico(\mathcal{T})$ can be constructed iteratively, by first including all points that lie on a "horizontal" or "vertical" line joining elements of $\mathcal{T}$, then including all points that lie on any horizontal/vertical line joining the new points obtained in the previous step, and so on. The condition $\mu^0 \in bico(\mathcal{T})$ is necessary and sufficient for a finite subversive protocol to exist. See Proposition 2.1 and Remark 2.4 in Aumann and Hart (1986).

[24] To show $\mu_0 \notin bico^{\#}(\mathcal{T})$, it is enough to find a bounded biconvex function $f : \Delta(\mathcal{S}_X) \times \Delta(\mathcal{S}_Y) \to \mathbb{R}$ that separates μ_0 from $\mathcal{T}$. Since $\mathcal{T} \subseteq bico(\mathcal{T})$, to show $bico^{\#}(\mathcal{T}) = bico(\mathcal{T})$ it then suffices to find such a function separating $bico(\mathcal{T})$ from its complement (Proposition 4.1 in Aumann and Hart, 1986). For example (i) of Figure 7, see Example 3.4 in Aumann and Hart (1986) for the relevant separating function (subject to a permutation). For example (ii), $f(p, q) = (\frac{1}{2} - p)(1 - q)$ is such a function. For the last example, we may take $f(p, q) = (1 - p)(1 - q)$.

slack curve, a self-similar dead-end where beliefs get stuck. At no point in the process can the players maintain non-negative slack and also be sure that the optimal decision is to reject the proposal, i.e., they cannot reach $\mathcal{T}(s)$ for $s \in \mathcal{R}$.

Similar remarks apply to example (iii) where subversion is also impossible for any interior priors since $bico^{\#}(\mathcal{T})$ has an empty interior.[25] For example (ii), $bico^{\#}(\mathcal{T})$ contains interior priors. Panel (e) shows a bimartingale that starts within $bico^{\#}(\mathcal{T})$ and reaches $\mathcal{T}(s)$ for each s. In this protocol, Y reveals her type, then X takes the appropriate decision. Subversion is not possible if instead X first reveals her type. Subversion is also impossible for priors outside $bico^{\#}(\mathcal{T})$ as posteriors cannot enter $bico^{\#}(\mathcal{T})$ with probability one.

4.3 Discussion

The belief-based approach of Proposition 4 utilizes the characterization of bimartingales and biconvex hulls in Aumann and Hart (1986). Since the decision rule is fixed for a subversion and incentive constraints have no bite, the only input needed is the specification of the sets $\mathcal{T}(s)$, for each $s \in \mathcal{S}$, and their union $\mathcal{T}$. These terminal belief sets are determined in turn by preferences that pin down how the state space $\mathcal{S}$ is partitioned into the three sets $\mathcal{A}$, $\mathcal{C}$ and $\mathcal{R}$.

This method can be used to generate additional results on subversion. For instance, one can ask for necessary and sufficient conditions for the existence of fine subversions, in view of their robustness properties. Recall for a fine subversion decision making is always fully informed, which is possible if and only if some player perfectly reveals her type before a decision is taken. This implies that the set $\mathcal{T}_f(s)$ of possible terminal beliefs for $s \in \mathcal{S}$ in a fine subversion will differ from $\mathcal{T}(s)$ only in the additional requirement that either μ_X be degenerate on x, or μ_Y be degenerate on y, for any $\mu = (\mu_X, \mu_Y) \in \mathcal{T}_f(s)$, $s = (x, y) \in \mathcal{S}$. Proposition 4 can then be amended to conclude that a fine subversion exists if and only if $\mu^0 \in bico^{\#}(\mathcal{T}_f)$, where $\mathcal{T}_f = \cup_{s \in \mathcal{S}} \mathcal{T}_f(s)$. Similarly, it is easy to see how the terminal belief sets can be amended in order to allow for alternative burdens of proof different from the balance of probabilities notion that underlies (DC).

The belief-based approach fixes preferences and generates a biconvex hull that contains all the priors for which a subversive protocol exists. In contrast, the constructive approach of Sections 3 fixes priors and, starting from a subversive conversation for a given game, uses the invariance properties to identify a set of games where the acceptance, rejection and conflict sets (and priors) may be different and yet the same conversation is subversive. Since subversion is possible in the same way for this set of games, the constructive approach identifies the robustness properties of a given conversation.

Each of the two approaches can be used to shed light on the other. Given a prior, the (set of)

[25] Since $\mathcal{T}(s)$ is empty for $s \in \mathcal{C}$ subversion is only possible in this example for priors that put zero weight on the row or column that contains s. The biconvex hull constructed in panel (e) confirms this.

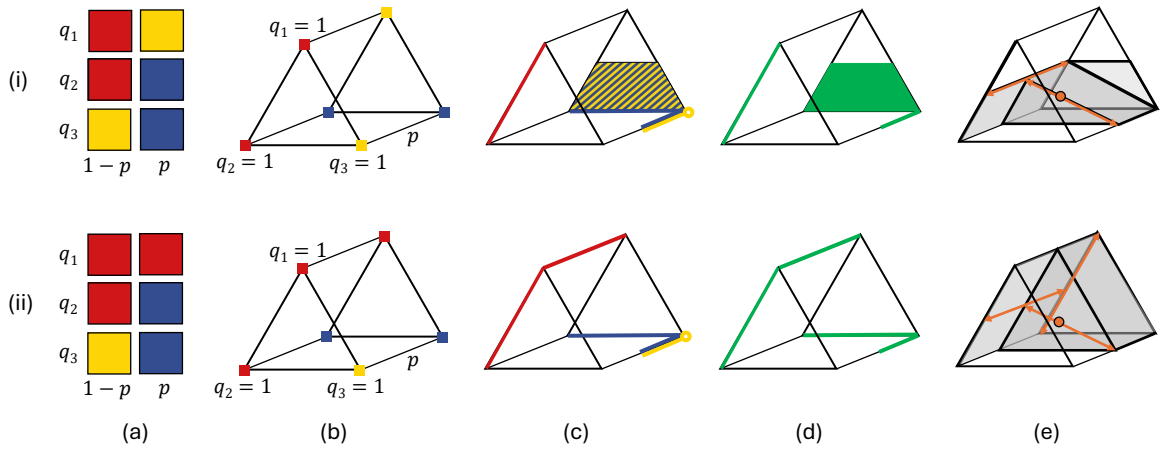

Figure 8: Biconvex hulls and the subset property

games identified by the invariance properties applied to a fixed subversive conversation must all generate biconvex hulls that contain that prior. For instance, the invariance properties allow us to conclude that all iid priors belong to the biconvex hull generated by any (finite type) biased committee game because the same conversation is subversive in all these games. More generally, if $\{\mathcal{A},\mathcal{C},\mathcal{R}\}$ and $\{\mathcal{A}',\mathcal{C}',\mathcal{R}'\}$ satisfy $\mathcal{A} \subseteq \mathcal{A}'$ and $\mathcal{R} \subseteq \mathcal{R}'$, then the biconvex hulls for fine subversions that they generate must also have the same subset relation, $bico^{\#}(\mathcal{T}_f) \subseteq bico^{\#}(\mathcal{T}'_f)$, using the subset property of fine subversions introduced in Section 3.1, a special case of the decision measurability property of Section 3.3.

Figure 8 shows an example of two games with the subset property. The state space and preferences are depicted in panel (a). Panel (b) shows the space of beliefs $\Delta(\mathcal{S}_X) \times \Delta(\mathcal{S}_Y)$ equal to the prism $\Delta^1 \times \Delta^2$, where Δ^n is the n-simplex. In panel (c) we construct the terminal belief sets for each $s \in \mathcal{S}$ in this prism, using the definition of $\mathcal{T}(s)$ and also the fact that beliefs must be a product measure. Panel (d) shows $\mathcal{T}$ and panel (e) its biconvex hull. Notice that the terminal belief sets for the two games are not subset ordered but their biconvex hulls are.[26]

Panel (e) shows two subversive bimartingales, one for each game, starting from uniform priors for x and y. In the associated conversation, that is common to both games, Y first reveals whether the state belongs to the bottom row or not and, in the latter case, X reveals the column. The associated action plans specify that a player takes the committee-optimal decision, that may differ across the two games, immediately after the other player reveals her type. As can be seen using the information contained in the other panels, posterior beliefs reach the terminal belief sets $\mathcal{T}(s)$, for each $s \in \mathcal{S}$, and each game, whenever a decision is taken. Indeed, this is true for every subversive conversation one can construct for the game of Figure 8(i), for each prior in

[26] In order to avoid clutter, we do not show every $\mathcal{T}(s)$ in Figure 8, instead listing them all in Appendix B.4. Nevertheless, panel (c) reveals that each of these sets only contain beliefs that are degenerate on either the X or Y dimension so that $\mathcal{T} = \mathcal{T}_f$ and a subversion exists in each game if and only if a fine subversion does. It can also be shown that $bico(\mathcal{T}) = bico^{\#}(\mathcal{T})$ for each game (details available upon request).

its biconvex hull. The same conversation will also be subversive for the game in Figure 8(ii), for that prior. These invariance properties underlie the robust subversions constructed for the biased committee and biased observer models of Section 3.

These results also show when encryption-like phenomena are necessary for subversion. When priors lie inside the biconvex hull, encryption is not necessary. The players can do as well with a conversation as when they can guarantee the encryption is completely secure. When priors lie outside the biconvex hull, secure encryption is necessary for subversion, given our assumption that the observer sees their final decision. Encryption is also sufficient in this case, as long there is ex ante non-negative slack.

Since incentives constraints do not play a role for subversion, using the bimartingales approach of Aumann and Hart (1986) to only track beliefs is enough to obtain necessary and sufficient conditions for existence. Similar techniques can be used to pursue extensions that go beyond subversion, although we will then need to fully specify the cardinal preferences of the committee, as opposed to only their ordinal preferences. For instance, one may wish to characterize the optimal decision rule that can be implemented via conversations when subversion is impossible. Under the assumption the players can commit to a protocol, one can still use bimartingales and only track beliefs. Absent commitment, incentive constraints will have a role to play. We would then need to adapt the dimartingales approach introduced by Aumann and Hart (2003) to two-sided private information and take into account not only beliefs but also expected payoffs and active incentive constraints. Similar remarks apply to the case where there are conflicts of interest within the committee in addition to those with the outside observer. We leave these interesting questions for future research.

5 Related literature

This paper considers communication of multi-dimensional information over multiple rounds by two senders. We analyze this communication as cheap talk (Crawford and Sobel, 1982; Green and Stokey, 2007) with multiple audiences (Farrell and Gibbons, 1989). The players in our model have the same preferences, are each informed on one dimension, and need to share enough information to reach their ideal decision while being scrutinized by an outside observer. This is the opposite of the questions asked by Krishna and Morgan (2001) and Battaglini (2002) who show how multiple senders with the same information and different preferences can be made to reveal all information to a receiver. Our focus on first-best outcomes for the senders implies they never have an incentive to deviate. Hence our results on existence and non-existence of subversive conversations apply equally when commitment is possible, as in the belief-based approach of the Bayesian persuasion literature (Kamenica and Gentzkow, 2011).

Forges (1990a), Aumann and Hart (2003), Krishna and Morgan (2004) and Chen, Golts-

man, Horner, and Pavlov (2017) consider communication over multiple rounds when one player is informed and the rounds are used either for communication by that player or for adjoint lotteries via simultaneous messages.[27] They show how extended communication expands the set of equilibrium outcomes. The same is true in our setting, although we restrict attention to polite talk and focus on the particular case of subversion where incentive constraints play no role. This allows us to use the characterization of biconvex hulls in Aumann and Hart (1986) in order to obtain necessary and sufficient conditions on priors that makes subversion possible in finite environments. The invariance properties of a subversive conversation that we provide show how the same conversation is subversive across a wide range of preferences and priors.

Matthews and Postlewaite (1995) ask if polite talk can create contexts that allow information to be safely shared, when it could not be earlier. They provide examples where multi-round sequential communication between two players, in the presence of a third, leads to different decisions compared to one-shot communication. In our setting, the presence of the third party modifies the process of communication but ultimately has no effect on decisions. Chakraborty and Yilmaz (2017) consider a cheap talk game between two experts with different information and possibly different preferences. They introduce a notion of agreement between committee members called consensus that takes into account information revealed by the play of the game. Plausible deniability requires the outside observer to also consent after observing the messages.

Within the political economy literature, this paper is most directly related to models of committee deliberations. Gradwohl and Feddersen (2018) study the effect of transparency in a cheap talk model with multiple senders who have common interests and correlated binary signals (see also Wolinsky, 2002; Feddersen and Gradwohl, 2020). They show that transparency may prevent any information transmission, hurting the senders and receiver, when conflict between the two groups is large. Our comparison of secure versus insecure communication is related to this transparency versus opacity distinction. In our environment with a richer space of signals for each player, we identify conditions when the committee can subvert even under transparency.

The cryptography literature on information-theoretic security analyzes related problems of how to both hide and share information. Secret sharing protocols (Shamir, 1979) analyze how information (i.e., shares/parts of a secret) can be split between players to ensure that any one player's information reveals nothing about others' information. If one of two players publicly reveals her share of the secret, an observer will not learn anything about the secret while the other player will have complete information.[28] Since we do not give the players the freedom to design how information is partitioned between them, back-and-forth communication is needed for subversion. Yao's (1982) millionaires' problem considers how two millionaires can figure

[27] See Forges (2020) for a survey. A key contribution of this literature is showing when equilibria attainable by use of a mediator can be implemented by cheap talk over multiple stages (e.g., Forges, 1990b). In our context the requirement of public communication limits what can be done.

[28] He, Sandomirskiy, and Tamuz (2021) consider the related problem of how to design such "private-private" information structures which give information about a decision-relevant state to the players.

out who is wealthier, without revealing their exact wealths to each other.[29] In our model, the players are happy to reveal their types to each other, and indeed one player must fully reveal its type to the other player in fine subversions which are robust. The deniable encryption literature (Beaver, 1996; Canetti et al., 1997) considers situations where plausible deniability is ensured if the encrypting scheme can generate an innocuous decryption. We require that the observer understand the equilibrium meaning of all messages and is willing to accept the innocuous interpretation (e.g., the state belongs to $\mathcal{A}$) only if its Bayesian posterior is large enough.

In a subversion, the players effectively hold decision making power, even if legal authority lies with the observer. This echoes the distinction between formal and real authority drawn by Aghion and Tirole (1997). In our setting, the ability to manage the process of communication may give the players effective authority. Because information is dispersed, the circumstances under which delegating formal authority is optimal for the organization (see, e.g., Dessein, 2002; Alonso, Dessein, and Matouschek, 2008) remains an open question.

6 Conclusion

This paper analyzes a common situation in information transmission between players with similar interests—their communications may be overheard by outsiders with different interests. When communication is scrutinized by an outside observer who understands the equilibrium meaning of messages, the process of communication matters. Different communication protocols that all implement the same optimal decisions from the perspective of the players can differ in what information is revealed publicly.

We show how a back and forth conversation can create a sequence of contexts that allows sufficient information to be shared between the players to to take their optimal decision, while also concealing enough information to withstand scrutiny. Even if the conversation is public, or private but leaked with some chance, the exact reason for the decision remains uncertain. The players thereby maintain deniability that their decision was influenced by bias rather than just the facts, while still taking the same optimal decisions they would in the absence of scrutiny.

A Proofs

Proof of Proposition 1. Let the action plan α be defined as follows: player i takes a non-null decision, $\alpha_i \neq N$, if and only if the other player has perfectly revealed her type in the previous round; player i accepts the proposal if $(x, y) \in \mathcal{A} \cup \mathcal{C}$ and rejects it otherwise.

Fix $z \in (0, 1/2)$. The conversation is the same for each game $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$ with $\mathcal{A} = \mathcal{L}$:

- In round 1, X either says m_1: "$x \in [z, 1-z]$" or m_1': "$x \notin [z, 1-z]$".

[29] See Grigoriev, Kish and Shpilrain (2017) for a probabilistic formulation.

- In round 2, Y either says m_2:"$y \in [z, 1-z]$" or m_2': "$y \notin [z, 1-z]$".
- After history m_1, m_2': the residual state space is $[z, 1-z] \times [z, 1-z]^c$ and player X perfectly reveals x in round 3.
- After history m_1', m_2: the residual state space is $[z, 1-z]^c \times [z, 1-z]$, player X passes in round 3 and player Y perfectly reveals y in round 4.
- Otherwise, the residual state space is a game with $\mathcal{A} = \mathcal{L}$. In both of these cases the conversation restarts in a rescaled state space. After history m_1, m_2: the residual state space is $[z, 1-z] \times [z, 1-z]$. Rescale the state space to make it the unit box using the bijections $x' = (x - z) / (1 - 2z)$ and $y' = (y - z) / (1 - 2z)$; see Section 3.3. We now have an instance of the biased committee model and the conversation continues as above. After history m_1', m_2': the residual state space is $[z, 1-z]^c \times [z, 1-z]^c$ and using similar bijections we obtain another instance of the same model in which the conversation proceeds as above.

For each type of each player, a decision is taken with probability at least $\min[2z, 1-2z]$ in each recursion of the above process. It follows the conversation is a.s.-finite. It remains to show that a decision to accept will meet (DC). Consider the history m_1, m_2', after which X perfectly reveals x. If Y accepts the proposal, all (x, y) with $x \geq z$, $y \leq z$ belong to $\mathcal{A}$. Conditional on m_1, m_2', x, the residual part of $\mathcal{A}$ is $\{x\} \times [0, z)$ whereas the residual part of $\mathcal{C}$ is a subset of $\{x\} \times (1-z, 1]$ and the latter is weakly smaller (in the induced measure on $\mathbb{R}^1$). After history m_1', m_2 identical arguments establish (DC) will be met after an acceptance. ■

Proof of Proposition 2.

We first provide a subversive conversation for the game in Figure 3(c), which we denote Γ_c, followed by the one in Figure 3(d), which we denote Γ_d. Recall that the game in Figure 3(b), which we denote Γ_b, is solved by Proposition 1. Throughout, we consider uniform priors and describe conversations up to the point where one player reveals her type. After this the other player takes the subversive decision.

Lemma 1 *There exists a subversive conversation for* $\Gamma_c = \{\mathcal{A}_c, \mathcal{C}_c, \mathcal{R}_c\}$, *where* $\mathcal{S} = [0,1]^2$, $\mathcal{A}_c = \{(x, y) \in \mathcal{S} \mid x \geq 1/2, y \leq 1/2\}$, $\mathcal{R}_c = \{(x, y) \in \mathcal{S} \mid y \geq x\}$ *and* $C_c = \mathcal{S} \setminus \{\mathcal{A}_c \cup \mathcal{R}_c\}$.

Proof: The state space is shown in Figure 9(a). Player X in round 1 says "in" if $x \in [1/4, 3/4]$ and "out" otherwise. If X says "in", Y perfectly reveals $y \leq 1/4$ or $y \geq 3/4$, see panel (b), otherwise if $y \in [1/4, 3/4]$ we obtain Γ_c using the bijections $x' = 2x - 1/2$ and $y' = 2y - 1/2$, as shown in panel (c). If X says "out" in round 1, Y says "out" if $y \notin [1/4, 3/4]$ and we obtain a rescaled Γ_c, see panel (d), otherwise Y says "reveal" and X reveals x in round 3. ■

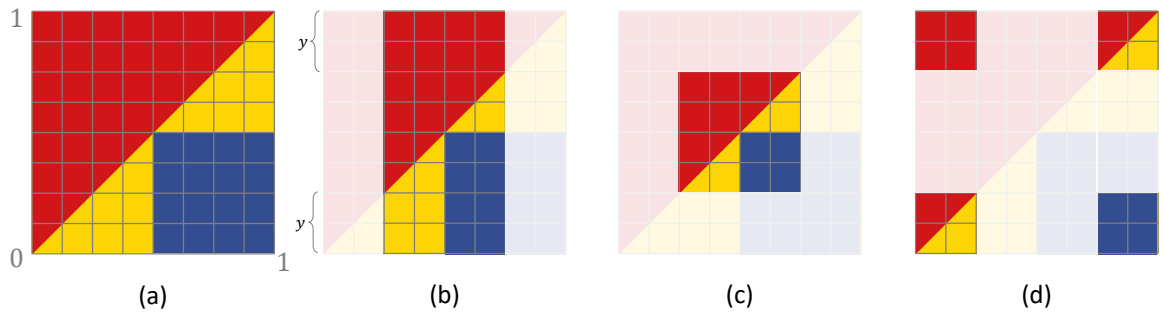

Figure 9: Recursive conversation for game Γ_c

Lemma 2 *There exists a subversive conversation for* $\Gamma_d = \{\mathcal{A}_d, \mathcal{C}_d, \mathcal{R}_d\}$, *where* $\mathcal{S} = [0,1] \times [0,5/2]$, $\mathcal{A}_d = \{(x,y) \in \mathcal{S} \mid y \leq x + 1/2\}$, $\mathcal{R}_d = \{(x,y) \in \mathcal{S} \mid y \geq x + 3/2\}$, *and* $C_d = \mathcal{S} \setminus \{\mathcal{A}_d \cup \mathcal{R}_d\}$.

Proof: The state space is shown in Figure 10(a). In round 1, player X reports "in" if $x \in [1/4, 3/4]$ and "out" otherwise. After player X says "in", player Y says "reveal" if $y \in [0,1/2] \cup [5/4, 7/4] \cup [9/4, 5/2]$ or says "pass" otherwise, as shown in panel (b). If Y said "reveal" in round 2, X reveals x. If Y says "pass", we can obtain Γ_d by the bijections $x' = 2x - 1/2$ and (i) $y' = 2y - 1$ if $y \leq 5/4$, or (ii) $y' = 2y - 2$ if $y > 5/4$; these bijections paste together the remaining pieces of the state space, as shown in Figure 10(c). If player X says "out" in round 1, player Y perfectly reveals the state if $y \in (1, 5/4]$, says "reveal" if $y \in [0, 1/4] \cup [3/4, 1] \cup [7/4, 9/4]$, or says "pass", as shown in panel (d). If Y said "reveal" player X reveals x in round 3, while if Y said "pass" in round 2, we obtain Γ_d through a set of bijections; see Figure 10(e). ■

Lemma 3 *There exists a subversive conversation for* $\Gamma_d^\ell = \{\mathcal{A}_d^\ell, \mathcal{C}_d^\ell, \mathcal{R}_d^\ell\}$ *if* $\ell \geq 3/2$, *where* $\mathcal{S} = [0,1] \times [0,\ell]$, $\mathcal{A}_d^\ell = \{(x,y) \in \mathcal{S} \mid y \leq x + \frac{1}{2}\ell - \frac{3}{4}\}$, $\mathcal{R}_d^\ell = \{(x,y) \in \mathcal{S} \mid y \geq x + \ell - 1\}$ *and* $C_d^\ell = \mathcal{S} \setminus \{\mathcal{A}_d \cup \mathcal{R}_d\}$.

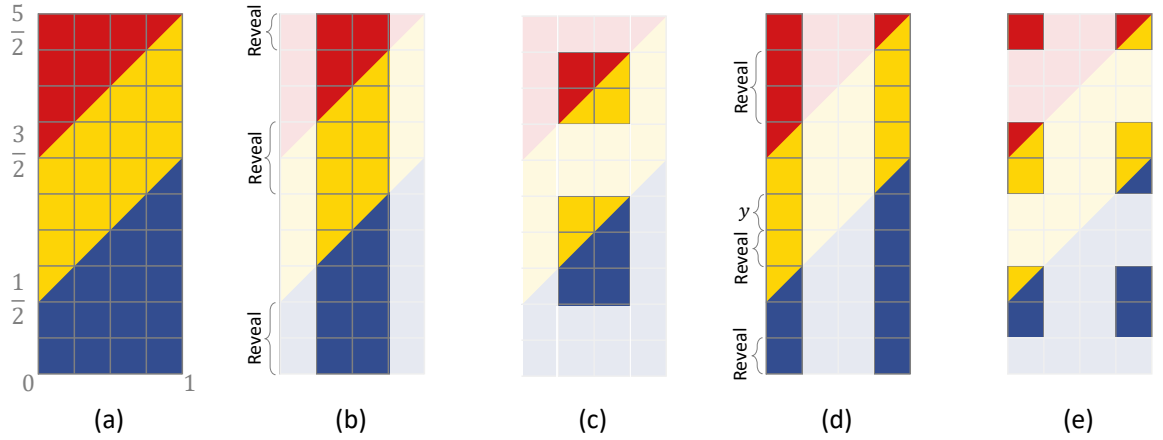

Figure 10: Recursive conversation for game Γ_d

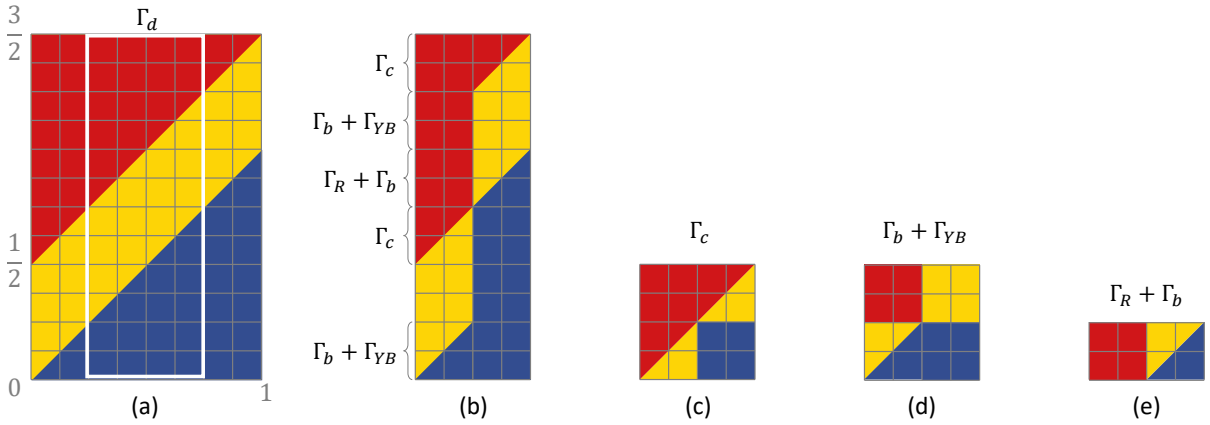

Figure 11: Recursive conversation for game Γ_d^ℓ, with $\ell = 3/2$

Proof: Observe that $\Gamma_d^{5/2} = \Gamma_d$, which is already solved. Consider $\ell > 5/2$ then player Y can say "reveal" if $y \in [0, \ell - 5/2] \cup \left[\frac{1}{2}\ell + \frac{1}{4}, \ell - 1\right]$ after which player X reveals the state and player Y takes a decision, or say "Γ_d" since pasting the remaining states together yields Γ_d.

Consider $\ell < 5/2$. Panel (a) of Figure 11 shows the case $\ell = 3/2$. In period 1, X says "Γ_d" if $x \in \left[\frac{5}{8} - \frac{1}{4}\ell, \frac{3}{8} + \frac{1}{4}\ell\right]$ and after y rejects the proposal for high y, the continuation game is a rescaled Γ_d, as seen in panel (a). Otherwise, X says "pass" and we are left with the state space in panel (b) of Figure 11. In round 2 Y says "Γ_c" if $y \in \left[\ell - 1, \frac{3}{4}\ell - \frac{3}{8}\right] \cup \left[\frac{5}{4}\ell - \frac{5}{8}, \ell\right]$ and Γ_c is solved as above; see panel (c). If $y \in \left(\frac{1}{2}\ell - \frac{3}{4}, \frac{1}{4}\ell - \frac{1}{8}\right) \cup \left(\frac{3}{2}\ell - \frac{5}{4}, \frac{5}{4}\ell - \frac{5}{8}\right)$ player Y says "$\Gamma_b + \Gamma_{YB}$"; this is depicted in panel (d). After this X says "Γ_b" if $x \leq \frac{5}{8} - \frac{1}{4}\ell$ and after Y takes the reject decision for sufficiently high y, we have a rescaled Γ_b, otherwise X reveals x. If $y \in \left(\frac{3}{4}\ell - \frac{3}{8}, \frac{1}{2}\ell + \frac{1}{4}\right)$ player Y says "$\Gamma_R + \Gamma_b$" in round 2, X reveals the state if $x \leq \frac{5}{8} - \frac{1}{4}\ell$ or says "Γ_b"; see panel (e). Finally, Y perfectly reveals $y \in \left[\frac{1}{4}\ell - \frac{1}{8}, \frac{1}{2}\ell - \frac{1}{4}\right]$ (these states are unlabeled in panel (b) of the figure) and says "pass" otherwise, after which X reveals x. ■

Proof of the proposition.

We provide a fine subversion for the case $\mathcal{A} = \mathcal{L}_b$, where there is zero total slack. Figure 12(a) shows how player X in round 1 partitions the state space into cases a, b and c.

Throughout, the cutoffs and sequence of messages described below have been chosen so that (DC) is always met, as can also be verified using the accompanying figures. Let $k = 7b - 2 \geq 0$ and $d = 1 - 3b \geq 0$.

Case i. If $x \in (2b - k, 2b + k)$, in round 1 player X says "case a" and after Y rejects the proposal where appropriate, game Γ_d^ℓ remains, with $\ell = d/k + 3/2$. This is solved by Lemma 3.

Case ii. If $x \in [b - d, b + d) \cup [6b - 1 - d, 6b - 1) \cup (1 - d, 1]$, X says "case b". In round 2, Y says:

"Γ_1" if $y \in [0, d] \cup (6b - 1, 1 - d]$; see Figure 12(b). In round 3 X says "Γ_b" if $x \in [b, b + d)$, the second column in panel (b), after which Γ_b is played once Y rejects the proposal for

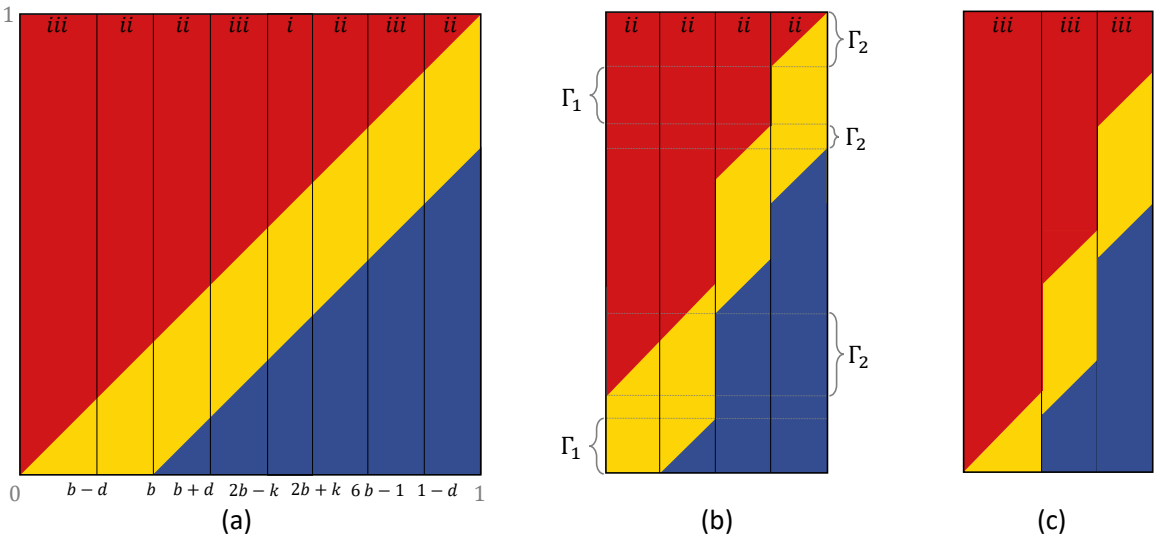

Figure 12: Construction for biased observer model

$y \in [6b-1, 1-d]$. Furthermore, in round 3 X perfectly reveals $x \in (1-d, 1]$. Otherwise, Y reveals y in round 4.

"y" perfectly revealing $y \in [d, b-d] \cup [5b-1, 1-d-b]$, (and X takes a decision).

"Γ_2" if $y \in [b-d, b+k] \cup [1-b, 1-b+k] \cup [1-d, 1]$; see Figure 12(b). Then X says "Γ_c" if $x \in [0, b+k] \cup [1-2b+d, 6b-1] \cup [1-d, 1]$ and Γ_c is played. Else Y perfectly reveals y.

"Γ_3" if $y \in [b+d, 5b-1] \cup [1-d-b, 2b+k]$, then X says "Γ_b" if $x \in [2b+k-d, 3b+k]$ after which Γ_b is played or X passes in which case Y reveals y in round 4.

"pass" for all other y. In round 3, X perfectly reveals if $x \in [0, b+k] \cup [1-b, 3b+k]$ or else says "$\Gamma_d^{3/2}$" and game $\Gamma_d^{3/2}$ is solved by Lemma 3.

Case iii. Figure 12(c) shows the remaining states. Player X in round 1 partitions these into a countable number of similar pieces, indexed by n (this could instead be done recursively). Let τ_n be the width and height of the triangular conflict area of the left-most piece, w_n be the common width of the other two pieces. Let the x-coordinate of the right edge of the right-most piece be x_n^r and x-coordinate of the right edge of the middle piece be x_n^m. We have $\tau_0 = b-d$, $x_0^r = 3b$, $x_0^m = 2-5b$, $w_0 = d$. For $n \geq 1$ define:

$$\begin{aligned} \tau_n &= (b-d)^n, & \Delta\tau_n &= \tau_n - \tau_{n-1}, \\ w_n &= d\,(b-d)^{n-1}, & \Delta w_n &= w_n - w_{n-1}, \\ x_n^r &= x_{n-1}^r - \Delta w_n, & x_n^m &= x_{n-1}^m - \Delta w_n. \end{aligned}$$

Player X in round 1 pools $x \in [\tau_n, \tau_{n-1}] \cup \left[x_n^m, x_{n-1}^m\right] \cup \left[x_n^r, x_{n-1}^r\right]$ by revealing the message "n" for $n \geq 1$, which exhausts case c.

We now solve the game for arbitrary n. In round 2, Y perfectly reveals the value of $y \in [x^m_{n-1} - b, x^m_{n-1} - b + w_n) \cup (x^r_{n-1} + b, 1]$; these decisions will meet (DC) because the two right columns both have width Δw_n. For any $n \geq 1$, in round 3, X says:

"Γ^ℓ_d" if $x \in [\, x^m_n, x^m_n + 2\Delta w_n - \Delta\tau_n]$, and after Y rejects the proposal for $y > x^m_{n-1} - \Delta\tau_n + \Delta w_n$, through a set of bijections the game Γ^ℓ_d is played for $\ell = (x^m_n + 2\Delta w_n - \Delta\tau_n) \,/\, (2\Delta w_n - \Delta\tau_n)$. This is shown in panel (a) of the figure.

"Γ_4" if $x \in [\tau_{n-1} - \Delta w_n, \tau_{n-1}] \cup \left[\, x^r_n, x^r_{n-1}\right]$. In round 2, player Y either (i) perfectly reveals $y \leq \tau_{n-1} - \Delta w_n$, (ii) says "$\Gamma_c$" if $y \in [\tau_{n-1} - \Delta w_n, \tau_{n-1}] \cup \left[\, x^r_n, x^r_{n-1}\right]$ and then Γ_c is played, (iii) says "$\Gamma_R + \Gamma_b$" if $y \in \left[\, x^r_n - b, x^r_{n-1} - b\right]$ after which X perfectly reveals $x \leq \tau_{n-1}$ or says "Γ_b" after which Γ_b is played, or (iv) says "pass" after which X reveals x.

"Γ_5" if $x \in [\tau_n, \tau_{n-1} - \Delta w_n] \cup \left[x^m_{n-1} - \Delta\tau_n + \Delta w_n, x^m_{n-1}\right]$. In round 2, player Y (i) perfectly reveals $y \leq \tau_n$, (ii) pools $y \in [\tau_n, \tau_{n-1} - \Delta w_n] \cup \left[x^m_{n-1} - \Delta\tau_n + \Delta w_n, x^m_{n-1}\right]$ after which Γ_c is played, (iii) pools $y \in \left[x^m_{n-1} - \Delta\tau_n + \Delta w_n - b, x^m_{n-1} - b\right]$ after which X perfectly reveals $x \leq \tau_{n-1}$ or says "Γ_b" and Γ_b is played, (iv) says "pass" after which X reveals x. ■

Proof of Proposition 3. We first show that any dead-end has zero slack.

Claim: If Γ is a dead-end, then $\Pr[\mathcal{A}] = \Pr[\mathcal{C}]$.

Proof of claim: Assume by way of contradiction that $\Pr[\mathcal{A}] > \Pr[\mathcal{C}]$. Then, by the law of iterated expectations, there exists $x \in \mathcal{S}_X$ which occurs with positive probability, such that $\Pr[\mathcal{A} \mid x, \mathcal{A} \cup \mathcal{C}] > \Pr[\mathcal{C} \mid x, \mathcal{A} \cup \mathcal{C}]$. Let $X^* \subset \mathcal{S}_X$ be the set of all such x. Observe that X^* is measurable since it is a subset of a finite set.

Consider the following protocol ξ for game Γ: Player X perfectly reveals $x \in X^*$ with probability $p \in (0, 1]$ and passes otherwise. If $\Pr[\mathcal{A} \mid (\mathcal{S}_X \setminus X^*) \times \mathcal{S}_Y, \mathcal{A} \cup \mathcal{C}] = \Pr[\mathcal{C} \mid (\mathcal{S}_X \setminus X^*) \times \mathcal{S}_Y, \mathcal{A} \cup \mathcal{C}]$, then $p = 1$. Else, $\Pr[\mathcal{A} \mid (\mathcal{S}_X \setminus X^*) \times \mathcal{S}_Y, \mathcal{A} \cup \mathcal{C}] < \Pr[\mathcal{C} \mid (\mathcal{S}_X \setminus X^*) \times \mathcal{S}_Y, \mathcal{A} \cup \mathcal{C}]$ and by the intermediate value theorem there exists a $p \in (0, 1)$ so that $\Pr[\mathcal{A} \mid \mathcal{S}(\xi), \mathcal{A} \cup \mathcal{C}] = \Pr[\mathcal{C} \mid \mathcal{S}(\xi), \mathcal{A} \cup \mathcal{C}]$, where $\mathcal{S}(\xi)$ denotes the remaining state space after X speaks , according to the protocol ξ and does not perfectly reveal a state in X^*. If X reveals $x \in X^*$, player Y can implement the subversive decision rule and satisfy (DC). Since decisions are taken with positive probability, Γ could not have been a dead-end. This proves the claim. ■

We are left to show that a dead-end Γ with $\Pr[\mathcal{A}] = \Pr[\mathcal{C}]$, can be (maximally) reduced to a union of self-similar dead-ends from figure 6. By the law of iterated expectations, there must exist two types $x, x' \in \mathcal{S}_X$ and a $p \in (0, 1]$ so that

$$\begin{aligned} & p \Pr[\mathcal{A} \mid x \times \mathcal{S}_Y, \mathcal{A} \cup \mathcal{C}] + \Pr[\mathcal{A} \mid x' \times \mathcal{S}_Y, \mathcal{A} \cup \mathcal{C}] \\ = \;& p \Pr[\mathcal{C} \mid x \times \mathcal{S}_Y, \mathcal{A} \cup \mathcal{C}] + \Pr[\mathcal{C} \mid x' \times \mathcal{S}_Y, \mathcal{A} \cup \mathcal{C}], \end{aligned}$$

and observe that the obedience constraint also has no slack on the complementary states (after state x' and a p-weight of state x are removed). Since $\mathcal{S}_X$ is finite, we can repeat this process to partition the types of Player X into pairs. Repeating this argument for Player Y, leaves us with a set of dead-ends at which each player has two possible types. It is easy to check that these must be the self-similar dead-ends of Figure 6, which describes all 2×2 dead-ends. ■

Proof of Proposition 4. Theorem 4.3 in Aumann and Hart (1986) states that a bimartingale $\widetilde{\mu}$ with expectation μ^0 has a limiting distribution in a set $\mathcal{T}$, that it reaches a.s. in finite time, if and only if $\mu^0 \in bico^{\#}(\mathcal{T})$.

Necessity follows immediately from how we construct $\widetilde{\mu}$ from ξ. For suppose $\widetilde{\mu}$ is derived from a subversive protocol ξ. Since for each $s \in \mathcal{S}$ the observer's beliefs belong to $\mathcal{T}(s)$ the first time a non-null decision is taken (and constant thereafter), the limiting distribution of $\widetilde{\mu}$ must be in $\mathcal{T} = \cup_{s \in \mathcal{S}} \mathcal{T}(s)$ and this limit must be reached a.s. in finite time.

In the other direction, assume we are given a bimartingale $\widetilde{\mu}$ with expectation $\mu^0 \in bico^{\#}(T)$ that has a limiting distribution in $\mathcal{T}$, reached a.s. in finite time. Pick a round t and history h^t such that $\widetilde{\mu}^t = \mu(h^t) \in \mathcal{T}$. Since $\mathcal{T} = \cup_{s \in \mathcal{S}} \mathcal{T}(s)$, either $\mu(h^t)[\mathcal{R}] = 1$, or $\mu(h^t)[\mathcal{R}] = 0$ and $\mu(h^t)[\mathcal{A}] \geq 1/2$. It follows that if $\mu(h^t)[s] > 0$ for some $s \in \mathcal{S}$, then $\mu(h^t) \in \mathcal{T}(s)$. To see this, suppose first that $s \in \mathcal{R}$ in which case we must have $\mu(h^t)[\mathcal{R}] = 1$, since otherwise $\mu(h^t)[\mathcal{R}] = 0$, contradicting $\mu(h^t)[s] > 0$ for $s \in \mathcal{R}$. Similarly, if $s \in \mathcal{R}^c$ we must have $\mu(h^t)[\mathcal{R}] = 0$ and $\mu(h^t)[\mathcal{A}] \geq 1/2$, since otherwise $\mu(h^t)[\mathcal{R}] = 1$, contradicting $\mu(h^t)[s] > 0$ for $s \in \mathcal{R}^c$. It follows that $\widetilde{\mu}$ reaches $\mathcal{T}(s)$ a.s. in finite time, for every s for which $\mu^0[s] > 0$. We conclude that the protocol ξ derived from $\widetilde{\mu}$ is subversive.■

B Additional Results

B.1 Short conversations

This section first shows a necessary and sufficient condition for subversion in games where player X's type is binary and priors are uniform. This allows us to give a four-round robust subversive conversation for the biased committee model, in which X creates a binary type continuation game in round 1 and then the conversation proceeds as described in the first result. Next we show that four rounds are the minimum required for a robust subversive conversation. Finally, we discuss how to use this four-round conversation in the continuation games of the biased observer model depicted in Figure 3.

Let $x \in \mathcal{S}_X = \{l, r\}$, $l < r$, and let $\mathcal{S}_Y \subseteq \mathbb{R}$. Assume that the priors P and Q are uniform and suppose $\mathcal{A}$ and $\mathcal{R}$ are monotonic: (i) $(x, y) \in \mathcal{A}$ implies $(x', y') \in \mathcal{A}$ for all $x' \geq x$ and $y' \leq y$ and (ii) $(x, y) \in \mathcal{R}$ implies $(x', y') \in \mathcal{R}$ for all $x' \leq x$ and $y' \geq y$. We refer to $\{l\} \times \mathcal{S}_Y$ as the "left type" and $\{r\} \times \mathcal{S}_Y$ as the "right type".

Lemma 4 *For a game with binary X types and uniform priors,* $\Pr[\mathcal{C}] \leq \Pr[\mathcal{A}]$ *and $\mathcal{A}$, $\mathcal{R}$ monotonic, a subversive conversation exists if and only if* $\Pr[(\{r\} \times \mathcal{S}_Y) \cap \mathcal{C}] \leq \Pr[(\{r\} \times \mathcal{S}_Y) \cap \mathcal{A}]$*, i.e., the right type has non-negative slack.*

Proof. Suppose the right type has negative slack, $\Pr[(\{r\} \times \mathcal{S}_Y) \cap \mathcal{C}] > \Pr[(\{r\} \times \mathcal{S}_Y) \cap \mathcal{A}]$, but a subversion exists. Player X must pool $x = r$ and $x = l$ in round 1. Since a decision must be taken with probability 1 in finite time, Y must pool types in $proj_Y Z$, for some positive measure $Z \subset (\{r\} \times \mathcal{S}_Y) \cap \mathcal{C}$. By monotonicity of $\mathcal{A}$ the acceptance zone is non-decreasing in x, and so $(\{l, r\} \times proj_Y Z) \cap \mathcal{A} = \emptyset$. Thus, (DC) cannot hold after such a history, a contradiction.

Suppose the right type has non-negative slack. If the left type also has non-negative slack, X can just reveal x and Y can take a decision. So suppose the left type has negative slack and let $q = \Pr[(\{l\} \times \mathcal{S}_Y) \cap \mathcal{C}] - \Pr[(\{l\} \times \mathcal{S}_Y) \cap \mathcal{A}] > 0$. Since $\Pr[\mathcal{C}] \leq \Pr[\mathcal{A}]$, $\mathcal{R}$ is monotonic and both types have the same measure, it follows that $\Pr[(\{r\} \times \mathcal{S}_Y) \cap \mathcal{A}] - \Pr[(\{l\} \times \mathcal{S}_Y) \cap \mathcal{A}] \geq q$. By the monotonicity of $\mathcal{A}$ and $\mathcal{R}$, there exists a set, $Z \subset (\{l\} \times \mathcal{S}_Y) \cap \mathcal{C}$, (the part of the conflict zone just above the acceptance zone of the left type) such that the measure of Z is at least q and so that $\{r\} \times proj_Y Z \subset \mathcal{A}$. Thus, Y can perfectly reveal $y \in proj_Y Z$, after which X accepts. Otherwise Y passes, X perfectly reveals x and Y takes the players' ideal decision. ■

In the biased committee model, we can have player X pool types $x = a$ and $x = 1 - a$ in the first period. For any biased committee game, this results in non-negative slack in the right type and monotonicity will be satisfied. Starting from the second round, the conversation proceeds as in the proof of Lemma 4 . Since this is a fine conversation that is subversive for the worst case where $\mathcal{R}$ is empty, it yields a robust subversion that takes four rounds. Proposition 5 shows that this is the shortest possible robust subversion for the biased committee model.

Proposition 5 *Any robust subversive conversation in the biased committee model requires at least four rounds.*

Proof. Assume by way of contradiction that there exists a three-round robust subversive conversation. It must work for conflict sets of the form $\mathcal{C}(b) = \left\{(x, y) \in [0,1]^2 \,|x \leq y \leq x + b\right\}$, $b \in (0, 1)$, as they are instances of the model.

A robust subversive protocol can depend on b only at the decision stage. So before a decision is taken, the other player must perfectly reveal her type. Otherwise there would not be sufficient information to take a decision for every $b \in (0, 1)$.

Player X cannot perfectly reveal $x \in [0, 1/2)$ in round 1 and satisfy (DC) for $b \geq 1/2$. Thus for a positive measure of X's types a decision cannot be taken in round 2. For X to take a decision in round 3, Y must reveal $y \in [0, 1]$ in round 2 for all histories which follow the set of messages $\{\sigma(x, h^0, \omega_1) : x \in [0, 1/2)\}$. However, Y cannot do so and meet deniability, since perfectly revealing $y = 1$ results in only the singleton point $(1, 1) \in \mathcal{A}$ and a positive measure of

states in $\mathcal{C}$. A similar argument applies for y sufficiently close to 1. Thus a decision cannot be taken in round 3 for a positive measure of X's types, which is a contradiction. ■

Unlike the biased committee model, the same four-round construction cannot work for the biased observer model (panel (a) in Figure 3). This is because for a sufficiently close to $1/2$ the continuation game with binary types $\{a, 1-a\}$ for X will not have non-negative slack, $\Pr[\mathcal{A}] \geq \Pr[\mathcal{C}]$. In fact, no construction where pairs of X types are pooled together is possible for the biased observer model. Because the problem has zero slack, (DC) must hold with equality for all pooled pairs. But, for values of x close to 1 there is no other type with sufficiently high conflict to pair it with and ensure zero slack in the deniability constraint. The construction we employ to prove Proposition 2, reduces the problem to the three continuation games depicted in panels (b)-(d) of Figure 3 for each of which the short conversation described above applies.

B.2 Non-iid priors

Consider again the biased committee model for which $\mathcal{A} = \mathcal{L}$ and $P = Q$. What if priors were not identical but still independent? The invariance properties help us identify robustness in this dimension. For instance, if P first-order stochastically dominates Q, we can still employ the transformation to quantiles via the relabeling property to obtain $\mathcal{A} \supseteq \mathcal{L}$. The decision measurability (specifically, the subset) property then gives Proposition 1. Alternatively, suppose P and Q are tail-symmetric.[30] From panels (b) and (c) of Figure 2 it is easy to see that the conversation goes through unchanged since both conditions of the decision measurability property are satisfied in the continuation games where decisions are taken. Alternatively, the relabeling property allows us to transform the tail-symmetric priors P and Q to uniform priors. The set $\mathcal{A}$ will then inherit a similar "symmetry" property in the transformed space of quantiles: either $(x, y) \in \mathcal{A}$ or $(1-x, 1-y) \in \mathcal{A}$. We exploit this symmetry to prove Proposition 1.

Next, suppose that x and y are not statistically independent. Assume they admit a strictly positive joint density $g(x, y)$ and denote by $g(x|y)$ and $g(y|x)$ the conditional densities. Fix the sets $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$ such that $\mathcal{A}$ is monotonic, that is, if $(x, y) \in \mathcal{A}$ then $(x', y') \in \mathcal{A}$ for $x' \geq x$ and $y' \leq y$. This property obtains in both the biased committee and biased observer models.

Proposition 6 *Fix $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$ with $\mathcal{A}$ monotonic. If a subversive conversation exists when priors are given by a joint density $g(x, y)$ then the same conversation is subversive if instead priors are given by a joint density $g'(x, y)$ that satisfies (a) $\frac{g(y|x)}{g'(y|x)}$ is non-decreasing in y, for all x, and (b) $\frac{g'(x|y)}{g(x|y)}$ is non-decreasing in x, for all y.*

Proof: Fix a subversive conversation under prior g. The conversation truncates the state space into smaller continuation games which inherit both $\mathcal{A}$-monotonicity and the monotone

[30] A cdf F on $[0, 1]$ is tail-symmetric if $F(x) + F(1-x) = 1$ for each $x \in [0, 1]$.

likelihood ratio properties from the statement of the proposition. To see that decision measurability extends to correlated priors, take any history h^t which leads to a decision $d_{t+1} \in \{A, R\}$. If the prior is g' instead of g, in the continuation game after history h^t the player taking the decision is still able to separate $\mathcal{R}$ states from the rest andtake the same decision $d_t = R$, since preferences over states have not changed, only the prior has.

We only need to check (DC) is met following an accept decision under priors g', for a (product) set of states that is a subset of $\mathcal{A} \cup \mathcal{C}$, given this constraint was met under g. The assumed likelihood ratio conditions imply that the ratio of joint distributions $\frac{g'(x,y)}{g(x,y)}$ is non-decreasing in y for any fixed x and non-increasing in x for any fixed y. This implies that g' puts less weight on the state with the lowest x-value and highest y-value, which must be in $\mathcal{C}$ by $\mathcal{A}$-monotonicity (if $\mathcal{C}$ is non-empty in this continuation game). Starting at this extreme (the lowest x and highest y values) as x increases or y decreases, g' puts increasingly more mass on those states relative to g. Thus by $\mathcal{A}$-monotonicity, prior g' puts more overall weight on $\mathcal{A}$ than g. Since (DC) was met for g, it must also hold for prior g'. Thus, the same conversation will continue to be subversive.■

B.3 Uncertain direction of bias

Is subversion possible when the direction of bias is not common knowledge? To allow for this possibility we consider problems that, in addition to the independent priors P and Q, are described by four (measurable) sets $\{\mathcal{A}, \mathcal{C}, \mathcal{D}, \mathcal{R}\}$ that partition the state space $\mathcal{S}$. As before, $\mathcal{A}$ is the set of states where both the committee and the observer would prefer to accept the proposal, while $\mathcal{R}$ is the set where all parties would agree to reject it. Also as before, $\mathcal{C}$ is the set of states where the committee is biased in favor of accepting the proposal while the observer would prefer to reject it. The new set $\mathcal{D}$ allows for the remaining possibility—it is the set of states where the committee would like to reject but the observer would prefer to accept.

Figure 13 provides a variant of the biased committee model where $\mathcal{D}$ is non-empty and so the direction of the committee's bias is not common knowledge at the outset. The committee prefers to accept when the state is in $\mathcal{A} \cup \mathcal{C}$ where $\{(x, y) \mid y \leq C(x)\}$ for some continuous function $C(.)$, and prefers to reject otherwise. The observer is impartial and he prefers one alternative over the other depending on whether the state is above or below the diagonal $y = x$, i.e., $\mathcal{A} \cup \mathcal{D} = \mathcal{L}$. So the committee is biased in favor of acceptance when $C(x) > x$ and biased in favor of rejection when $C(x) < x$. We continue to assume $P = Q$ is invertible.

Note that the committee's decision line $C(x)$ is increasing and crosses the diagonal $y = x$ at the point (x', y'). This implies X knows the the direction of conflict at the beginning of the game based on her own type x. Suppose she reveals it by disclosing whether $x \leq x'$ or $x > x'$. In the former case, Y can reject the proposal if $y > y'$; while if $y \leq y'$ the residual state space is an instance of the biased committee model (once appropriately rescaled) and so subversion

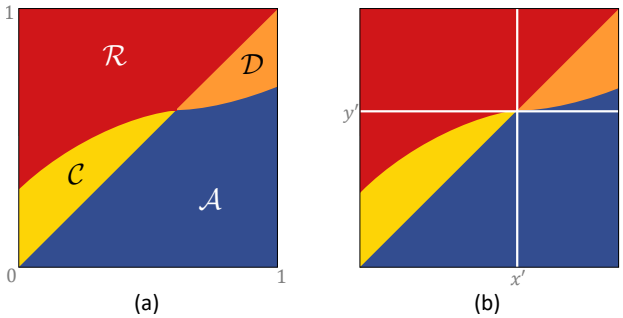

Figure 13: Uncertain direction of bias.

is possible in this subspace via Proposition 1. On the other hand, if X reveals $x > x'$, then Y can accept the proposal if $y < y'$. Otherwise, when $y \geq y'$, the residual state space is just an "upside down" version of the biased committee model (i.e., a "row" permutation, with the labels "accept" and "reject" reversed). By Proposition 1 subversion is possible in this subspace as well. The same argument holds for any increasing $C(x)$, no matter how many times $C(x)$ intersects the observer's decision line. The players can adapt the initial partitioning of the state space to take into account all the intersection points.

With uncertainty about the direction of bias, there are two deniability constraints, one where the proposal is accepted and another where it is rejected, unlike the model introduced in Section 2 where we assumed $\mathcal{D}$ was empty so there was only one deniability constraint. We can use the belief based approach to provide a necessary and sufficient conditions for existence of subversive protocols in this case. To do so, we restrict attention to finite environments, as in Section 4. Next, for each $s \in \mathcal{S}$, we define the subversive terminal belief set $\mathcal{T}_2(s)$ under two kinds of bias as follows:

$$\mathcal{T}_2(s) = \begin{cases} \{\mu \in \Delta(\mathcal{S}_X) \times \Delta(\mathcal{S}_Y) \mid \mu[s] > 0,\, \mu[\mathcal{A} \cup \mathcal{C}] = 0, \mu[\mathcal{R}] \geq 1/2\} & \text{if } s \in \mathcal{R} \cup \mathcal{D}, \\ \{\mu \in \Delta(\mathcal{S}_X) \times \Delta(\mathcal{S}_Y) \mid \mu[s] > 0, \mu[\mathcal{R} \cup \mathcal{D}] = 0, \mu[\mathcal{A}] \geq 1/2\} & \text{if } s \in \mathcal{A} \cup \mathcal{C}. \end{cases}$$

The definition takes into account both kinds of deniability constraints that the players have to meet in a subversion. Let $\mathcal{T}_2 = \cup_{s \in \mathcal{S}} \mathcal{T}_2(s)$.

Proposition 7 *Consider a finite environment and fix $\{\mathcal{A}, \mathcal{C}, \mathcal{D}, \mathcal{R}\}$. A subversive protocol exists if and only if the prior $\mu^0 \in bico^{\#}(\mathcal{T}_2)$.*

The result is an extension of Proposition 4 and its proof follows from identical arguments. Notice from the definition of the terminal belief sets that the direction of bias will be inferred from the decision itself, namely that the state does not belong to $\mathcal{D}$ (following an acceptance) or that the state does not belong to $\mathcal{C}$ (following a rejection). Indeed, the direction of bias must be known by the player about to take a decision. She can reveal this information at this point, if it has not already been revealed earlier. For the construction of Figure 13, the direction

of bias was revealed at the outset. In general, subversion is possible under uncertainty about the direction of bias if and only if it is possible when some player resolves the uncertainty at some stage during the conversation. Uncertainty about the direction of conflict does not create additional strategic issues with respect to existence.

B.4 Terminal belief sets

This section lists all the terminal belief sets $\mathcal{T}(s)$ for the examples of Figures 7 and 8, using the fact that beliefs must be a product measure. We will refer to the columns (X's types) as Left (l) and Right (r) and the row's (Y's types) as Top (t), Middle (m) and Bottom (b) and describe the terminal beliefs in terms of the pair (p, q).

Example (i) of Figure 7:

Table 1: Example (i) of Figure 7

	$x = l$	$x = r$
$y = t$	$\{p = 0, q = 1\}$	$\{p = 1, 0 < q \leq \frac{1}{2}\}$
$y = b$	$\{\frac{1}{2} \leq p < 1, q = 0\}$	$\{p \geq \frac{1}{2}, q = 0\} \cup \{p = 1, q \leq \frac{1}{2}\}$

Example (ii) of Figure 7:

Table 2: Example (ii) of Figure 7

	$x = l$	$x = r$
$y = t$	$\{p = 0, q = 1\}$	$\{p = 1, q > 0\}$
$y = b$	$\{\frac{1}{2} \leq p < 1, q = 0\}$	$\{p \geq \frac{1}{2}, q = 0\} \cup \{p = 1, q < 1\}$

Example (iii) of Figure 7:

Table 3: Example (iii) of Figure 7

	$x = l$	$x = r$
$y = t$	$\{p = 0, q = 1\}$	$\{p = 1, q = 1\}$
$y = b$	$\emptyset$	$\{p = 1, q = 0\}$

Example (i) of Figure 8:

Table 4: Example (i) of Figure 8

	$x = l$	$x = r$
$y = t$	$\{p = 0, q_1 > 0, q_3 = 0\}$	$\{p = 1, 0 < q_1 \leq \frac{1}{2}\}$
$y = m$	$\{p = 0, q_2 > 0, q_3 = 0\}$	$\{p = 1, q_1 \leq \frac{1}{2}, q_2 > 0\}$
$y = b$	$\{\frac{1}{2} \leq p < 1, q_3 = 1\}$	$\{p \geq \frac{1}{2}, q_3 = 1\} \cup \{p = 1, q_1 \leq \frac{1}{2}, q_3 > 0\}$

Example (ii) of Figure 8:

Table 5: Example (ii) of Figure 8

	$x = l$	$x = r$
$y = t$	$\{p = 0, q_1 > 0, q_3 = 0\} \cup \{p < 1, q_1 = 1\}$	$\{p > 0, q_1 = 1\}$
$y = m$	$\{p = 0, q_2 > 0, q_3 = 0\}$	$\{p = 1, q_1 = 0, q_2 > 0\}$
$y = b$	$\{\frac{1}{2} \leq p < 1, q_3 = 1\}$	$\{p \geq \frac{1}{2}, q_3 = 1\} \cup \{p = 1, q_1 = 0, q_3 > 0\}$

References

[1] Aghion, Philippe, and Jean Tirole. 1997. "Formal and Real Authority in Organizations." *Journal of Political Economy* 105(1): 1–29.

[2] Alonso, Ricardo, Wouter Dessein, and Niko Matouschek. 2008. "When Does Coordination Require Centralization?" *American Economic Review* 98(1): 145–179.

[3] Aumann, Robert J. and Sergiu Hart. 1986. "Bi-Convexity and Bi-Martingales." *Israel Journal of Mathematics* 54(2): 159–180.

[4] Aumann, Robert J. and Sergiu Hart. 2003. "Long Cheap Talk." *Econometrica* 71(6): 1619–1660.

[5] Battaglini, Marco. 2002. "Multiple Referrals and Multidimensional Cheap Talk." *Econometrica* 70(4): 1379–1401.

[6] Beaver, Donald. 1996. "Plausible Deniability." *1st International Conference on the Theory and Applications of Cryptology*, Pragocrypt'96, 272–288.

[7] Canetti, Rein, Cynthia Dwork, Moni Naor, and Rafail Ostrovsky. 1997. "Deniable Encryption." *Advances in Cryptology–CRYPTO'97 Proceedings* 17: 90–104.

[8] Chakraborty, Archishman and Bilge Yilmaz. 2017. "Authority, Consensus, and Governance." *Review of Financial Studies* 30(12): 4267–4316.

[9] Chen, Yi, Maria Goltsman, Johannes Hörner, and Gregory Pavlov. 2017. "Straight Talk." working paper.

[10] Crawford, Vincent P. and Joel Sobel, 1982. "Strategic Information Transmission." *Econometrica* 50(6): 1431–1451.

[11] Dessein, Wouter. 2002. "Authority and Communication in Organizations." *Review of Economic Studies* 69(4): 811–838.

[12] Farrell, Joseph and Robert Gibbons. 1989. "Cheap Talk with Two Audiences." *American Economic Review* 79(5): 1214–1223.

[13] Feddersen, Timothy, and Ronen Gradwohl. 2020. "Decentralized Advice." *European Journal of Political Economy* 63.

[14] Forges, Françoise. 1990a. "Equilibria with Communication in a Job Market Example." *Quarterly Journal of Economics* 105(2): 375–398.

[15] Forges, Françoise. 1990b. "Universal Mechanisms." *Econometrica* 58(6): 1341–1364.

[16] Forges, Françoise. 2020. "Games with Incomplete Information: From Repetition to Cheap Talk and Persuasion." *Annals of Economics and Statistics* 137: 3–30.

[17] Garicano, Luis, and Luis Rayo. 2016. "Why Organizations Fail: Models and Cases." *Journal of Economic Literature* 54(1): 137–92.

[18] Gradwohl, Ronen, and Timothy Feddersen. 2018. "Persuasion and Transparency." *Journal of Politics* 80(3): 903–915.

[19] Green, Jerry R., and Nancy L. Stokey. 2007. "A Two-Person Game of Information Transmission." *Journal of Economic Theory* 135(1): 90–104.

[20] Grigoriev, Dima, Laszlo B. Kish, and Vladimir Shpilrain. 2017. "Yao's millionaires' problem and public-key encryption without computational assumptions." *International Journal of Foundations of Computer Science*, 28(4): 379–389.

[21] He, Kevin, Fedor Sandomirskiy, and Omer Tamuz. 2021. "Private Private Information." *arXiv preprint arXiv*:2112.14356.

[22] Kamenica, Emir and Matthew Gentzkow. 2011. "Bayesian Persuasion." *American Economic Review* 101(6): 2590–2616.

[23] Krishna, Vijay and John Morgan. 2001. "A Model of Expertise." *Quarterly Journal of Economics* 116(2): 747–775.

[24] Krishna, Vijay and John Morgan. 2004. "The Art of Conversation: Eliciting Information from Experts through Multi-Stage Communication." *Journal of Economic Theory* 117(2): 147–179.

[25] Matthews, Steven A. and Andrew Postlewaite, 1995. "On Modeling Cheap Talk in Bayesian Games." In John O. Ledyard (ed.) *The Economics of Informational Decentralization: Complexity, Efficiency and Stability.* Springer, Boston, MA, 347–366.

[26] Shamir, Adi. 1979. "How to Share a Secret." *Communications of the ACM* 22 (11): 612–613.

[27] Shannon, Claude. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27(3): 379–423.

[28] Silberman, Alan H and Leah R. Bruno. 2017. "Sunk By Your Own Torpedoes! How Emails and Memos Can Lead to Antitrust and Other Litigation Issues." Presentation, Dentons.com.

[29] Wolinsky, Asher. 2002. "Eliciting Information from Multiple Experts." *Games and Economic Behavior* 41(1): 141–160.

[30] Yao, Andrew C. 1982. "Protocols for Secure Computations." *23rd Annual Symposium on Foundations of Computer Science*. 1: 160–164.