Subversive Conversations*

Nemanja Antic† Archishman Chakraborty‡ Rick Harbaugh§

This version: July, 2023

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 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.

†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 signals and common interests must exchange 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 consider the possibility of subversion. The players subvert when they share enough information to take the same ideal decision they would have 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 signals 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

[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," and to always "bookend" discussions within their proper contexts. Attorney-client privilege can be legally trumped by the "crime-fraud exception" as shown by a recent case involving the handling of classified documents by a former president.

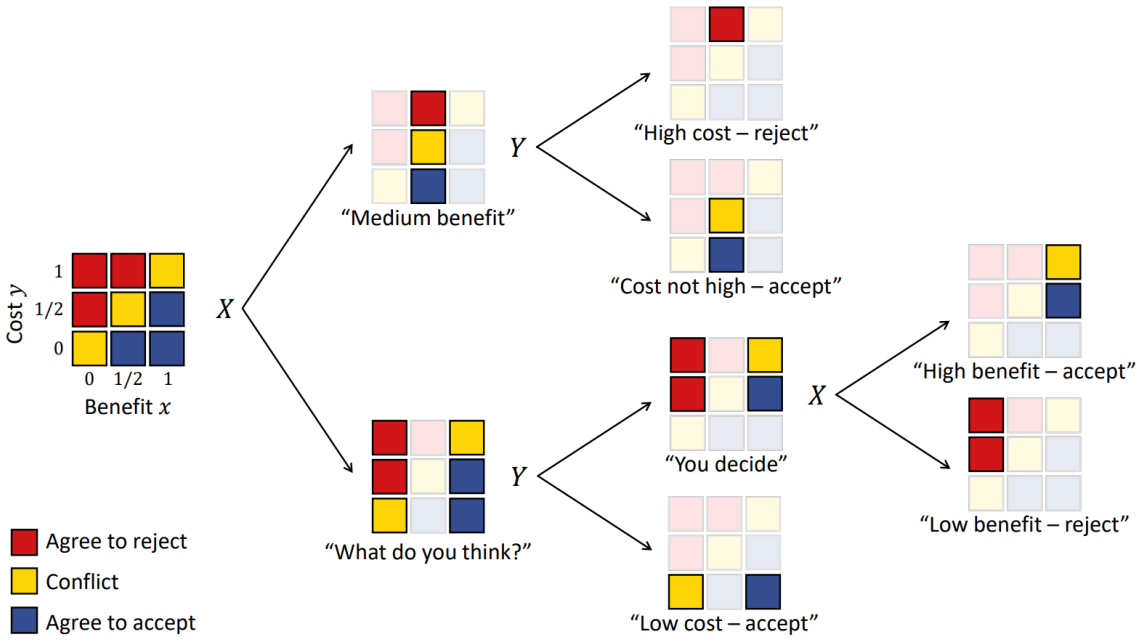

Figure 1: Conversation tree

being good. Priors are uniform on x and y. Communication is unverifiable "cheap talk".

Figure 1 depicts the following conversation which dynamically pools and separates types as it progresses, allowing the managers to determine if the project is truly bad while concealing from the public whether it is good or only mediocre. If the benefit is medium, manager X says so 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 but only Y knows whether the project is truly good or just 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 says nothing further. If instead the cost is medium or high then Y pools this information to gauge what X now thinks. 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 good or mediocre, but only Y knows which and reveals 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 achieve their first-best outcome, there is no incentive to deviate from this strategy. As long as the managers follow the subversive communication protocol, 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. In a wide range of

situations, we show that the committee can take its favored decisions as if it is free to choose as it pleases. This is true, for instance, when the choice is between two ex ante identical alternatives, such as hiring one of two job market candidates, and the players are biased in favor of one. Although their information is dispersed, they can use a conversation to achieve the same first best 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 sufficiently unfavorable news, while also waiting to reveal good news in order to create a favorable context in case her partner has bad news.

Our results imply that accountability and regulatory compliance may be difficult to ensure, even for transparent organizations. In the 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] 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] This is better than maintaining deniability by avoiding decision-relevant information from subordinates, which can lead to inefficient outcomes (Garicano and Rayo, 2016).

Hiring and personnel committees are often accused of bias toward candidates from similar backgrounds. Even though anti-discrimination laws exist and committee communications can be subpoenaed, our results imply this alone is unlikely to eliminate biased decisions. Transparency and greater stakeholder representation on boards are also concerns in the ongoing debate on corporate governance. Since management controls information flows and the deliberative process, such measures 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

[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."

[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 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."

state surveillance can be successful if they are careful about how they speak.

We allow the observer to fully understand the meaning of messages in the same way as the players, which rules out encryption of messages to coordinate on the optimal decision. Communication may be insecure because encryption might be cracked or bypassed; or encryption may be impractical because the conversation has to be in public and take place in ordinary language. Encryption is not necessary for our existence results since, unlike most of 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.

In particular, subversion in our model requires that the "conflict" zone be no larger than the "agree to accept" zone as in Figure 1.[5] This non-negative slack condition captures a commonality of interest between the observer and the players. 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 that the committee wants to do so. With dispersed information the committee has to figure out its own reasons publicly and because of this non-negative slack is not sufficient for subversion. The players have to come up with a conversation that is sufficiently informative for them but conceals enough from the observer.

As noted above, subversion is always possible when the committee is biased toward one of two ex ante identical alternatives. The same is true if the committee's bias is in favor of the alternative that is ex ante more likely to be better, and even when there is uncertainty about the magnitude or sign of the bias. More generally, we study changes to a game under which a given subversive conversation in the original game remains subversive in the changed game. The set of games identified by these invariance properties describe the robustness features of the given conversation; and the conversation is a "universal" subversion for that set of games.

When is subversion impossible? As seen in the opening example, a conversation must separate bad states while pooling mediocre and good states. If this cannot be done while taking the committee's optimal decisions with positive probability and meeting the deniability constraint, the conversation is at a "dead-end". For finite type spaces we find that all conversational dead-ends reach one of three irreducible dead-ends. Each has a binary type space for each player and observer preferences that are less flexible than the committee's. For instance, the committee may require only one of the two signals to be good in order to accept the proposal while the observer requires both of them to be good. For an irreducible dead-end, any attempt at subversion while maintaining deniability results only in continuation games that are all self-similar to the original game. No progress can be made, no matter how long the conversation.

To understand if any given game admits a subversive conversation, we employ a belief-based

[5]Indeed, the same condition must obtain at every stage of a subversive conversation, implying that the observer has no incentive to intervene and overrule the players at any stage of the conversation.

approach that tracks how the alternating statements by X and Y affect beliefs on the x or y dimension respectively. If the conversation is subversive, such a bimartingale must converge to a set of terminal beliefs consistent with subversion. Using the characterization of biconvexity in Aumann and Hart (1986), we show that this convergence occurs if and only if priors lie in (a version of) the biconvex hull of the set of subversive terminal beliefs. This provides a necessary and sufficient condition for the existence of subversive conversations in finite environments.

The rest of the paper is organized as follows. We develop and analyze the model in Sections 2 through 4. We then review the literature in Section 5 and conclude in Section 6. Appendices A and B contain proofs of results not contained in the main text, as well as additional results.

2 A model of subversion

2.1 Players, preferences and information

A committee is composed of two players, X and Y (both "she"). Player X privately observes a signal (or type) $x \in \mathcal{S}_X \subseteq \mathbb{R}$, 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 a state of the world. We assume players' types are independent and let P and Q denote the cumulative distribution of x and y respectively. Let $G = P \times Q$ denote the joint distribution which we allow to be continuous, discrete, or a mixture of the two.

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]

Since neither player is fully informed of the state $s \in \mathcal{S}$, they need to communicate to determine their optimal decision. Time is discrete, with rounds indexed by $t = 1, 2, ...$ 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 subset of her types. Let $m_t \in M$ denote a message sent in round t, $d_t \in D = \{A, R, N\}$ 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 and $d^t \in D^t$ a history of decisions, each of length t. Let $h^t = (m^t, d^t) \in H^t = M^t \times D^t$ denote a history. Let $\mathcal{H}$ be the set of all histories of arbitrary length. Let $\omega \in \Omega = [0, 1]$ denote the draw of a (uniformly distributed) random

[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.

variable that will represent any randomization by the players. A *protocol* for the committee, $\xi \equiv (\sigma, \alpha) : \mathcal{S} \times \mathcal{H} \times \Omega \to M \times D$, has two components, a *conversation* σ and an *action plan* α. A conversation specifies a (possibly random) message as a function of the state and history, $\sigma(s, h^t, \omega) \in M$, while an action plan specifies a (possibly random) decision, $\alpha(s, h^t, \omega) \in D$. We will write $\sigma \equiv (\sigma_X, \sigma_Y)$ and $\alpha \equiv (\alpha_X, \alpha_Y)$, where $\xi_i \equiv (\sigma_i, \alpha_i)$ is measurable with respect to player i's information in the rounds where $i \in \{X, Y\}$ makes a move. Thus, $\xi = (\sigma, \alpha)$ is a (mixed) behavior strategy profile with $\xi_i = (\sigma_i, \alpha_i)$ the strategy of player i.

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

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

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 $\Gamma = \{\mathcal{A}, \mathcal{C}, \mathcal{R}; P, Q\}$.

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

[9] We show why this assumption is without loss of generality later in the paper.

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 $H^t_A(\xi) = \{h^t \in H^t \mid d_t = A,\, d_{t'} = N,\, t' < t\}$ 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 $\Pr[\mathcal{A} \mid h^t] \geq 1/2$, for $h^t \in H^t_A(\xi)$. 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

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

When the committee accepts the proposal after a history h^t, (DC) 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. The deniability constraint is similar to the "balance of probabilities" burden of proof faced by courts in U.S. civil cases.[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 end 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, so that Bayes' Rule cannot be used to evaluate (DC). In such cases, we use (generalized) probability densities to compute posteriors. 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 $H^t_A(\xi)$ but also when we integrate over all histories that belong to any measurable subset of $H^t_A(\xi)$.

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.[11] 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

[10] 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 benchmark where a single expert holds both pieces of information, non-negative slack is necessary and sufficient for the expert to get her first-best decisions while meeting deniability.

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.[12] 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.[13]

2.2 Biased committee, unbiased observer

Consider the following environment where the committee chooses between two ex ante identical choices but is biased toward one. Let $\mathcal{S} = [0,1]^2$, $\mathcal{A} = \mathcal{L} \equiv \{(x,y) \in \mathcal{S} \mid y \leq x\}$ and $P = Q$. We interpret x as a benefit and y as a cost of the proposal from the observer's perspective since he prefers to accept the proposal if and only if y does not exceed x. Since $\mathcal{C} \subseteq \mathcal{L}^c$, the committee is biased relative to the observer—it may favor accepting the proposal even when the cost exceeds the benefit. The biased committee model is the set of games Γ with $\mathcal{A} = \mathcal{L}$ and $P = Q$.

We can restrict attention to uniform priors without loss of generality. Since any random variable is a transformation (via the quantile function) of a uniformly distributed random variable, the quantiles can be interpreted as the true underlying types. If the common cdf $P = Q$ is invertible then this is a one-to-one transformation and $\mathcal{A} = \mathcal{L}$ in the transformed space. If the cdf has atoms, many quantiles map into the same (atomic) type. But this is immaterial since all these quantiles have the same payoffs as the atom. The only difference is that we may have $\mathcal{A} \supseteq \mathcal{L}$ after the transformation. With priors taken to be uniform, a biased committee game Γ is fully specified by the sets $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$ with $\mathcal{A} \supseteq \mathcal{L}$. 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 $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$ that defines a biased committee game. 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

[12] 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.

[13] 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 optimal decision rule, identifying the ones that maintain deniability.

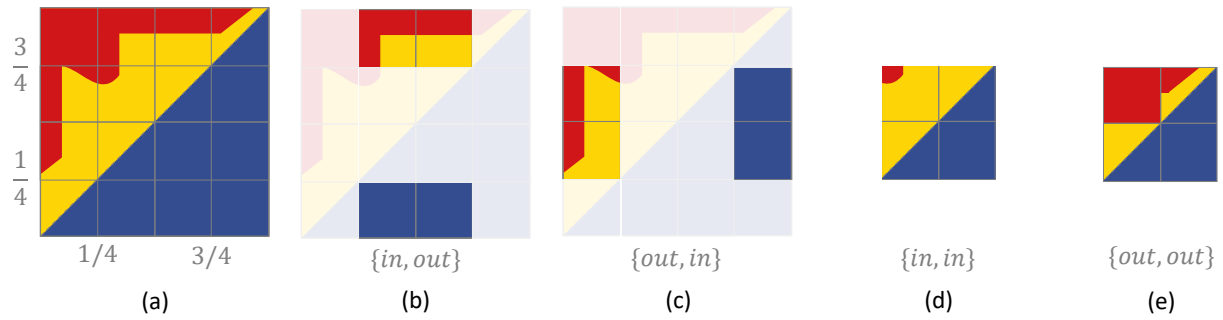

Figure 2: Recursive conversation for the biased committee model

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 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. Notice the importance of the history of prior messages that create a suitable context before a player can safely take a decision, i.e., meet 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 and $y = 3/4$. 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 before.

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.

Nothing in this argument depends on the particular properties of the game in Figure 2(a). Thus, Proposition 1 asserts the existence of a *universal* subversive conversation for the biased committee model. Because $\mathcal{A} \cup \mathcal{C}$ and $\mathcal{R}$ 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. When the committee encounters similar games frequently, this robustness feature simplifies communication design within organizations.

The conversation of Figure 2 is universal 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 $\Gamma = \{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$ is also a fine subversive conversation for any other game $\Gamma' = \{\mathcal{A}', \mathcal{C}', \mathcal{R}'\}$ 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. Since the residual set of states $\mathcal{S}\left(h^t\right)$ reached after history h^t is the same for both games, given the fixed conversation, we must have $\mathcal{C}' \cap \mathcal{S}\left(h^t\right) \subseteq \mathcal{C} \cap \mathcal{S}\left(h^t\right)$ and $\mathcal{A}' \cap \mathcal{S}\left(h^t\right) \supseteq \mathcal{A} \cap \mathcal{S}\left(h^t\right)$. As (DC) is met in the original game Γ, it is also met in the new game Γ' that has a smaller conflict set and larger agreement set.

Since she is fully informed in a fine subversion, the player who takes the decision can always tailor her decision to suit her preferences. The conversation before the decision allows her to be fully informed and still meet deniability. We generalize the subset property in Section 2.4. To see how it works in the biased committee model, consider the "worst case scenario" game with the largest possible conflict set $\mathcal{C} = \mathcal{L}^c$ and $\mathcal{A} = \mathcal{L}$, so that the players want to accept the proposal regardless of the state. The conversation described in Figure 2 is a fine subversion for this game. By the subset property, it is also subversive in every other biased committee game.

2.3 Unbiased committee, biased observer

To see how the committee can obtain its ideal outcome in other economic settings, consider a model where x and y describe the benefit and cost of the proposal from the perspective of the committee, while the observer is biased against the proposal relative to the players. Assume without loss of generality the common cdf is uniform and $\mathcal{S} = [0,1]^2$, i.e., we have made the quantile transformation described in Section 2.2. Let $\mathcal{R} = \mathcal{U} \equiv \{(x,y) \in \mathcal{S} \mid y \geq x\}$ and $\mathcal{A} = \mathcal{L}_b \equiv \{(x,y) \in \mathcal{S} \mid y < x - b\}$ where the parameter $b > 0$ represents the bias of the observer relative to the committee and $\mathcal{C} = \{(x,y) \in \mathcal{S} \mid x - b \leq y < x\}$ is the zone of conflict where the unbiased committee favors acceptance and the biased observer favors rejection.To ensure non-negative slack (a necessary condition for subversion), we need $b \leq b^* \equiv 1 - 1/\sqrt{2}$.

Call the set of games with these properties the biased observer model.[14] Notice that in the absence of any information from the committee, the biased observer would like to reject the proposal, unlike in the biased committee model where he would be willing to accept it. 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.*

[14]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.

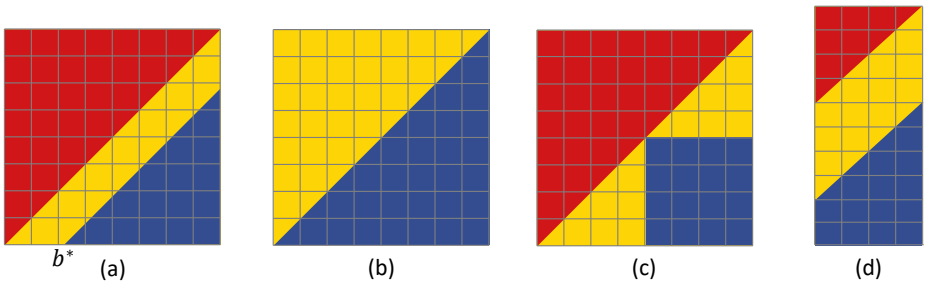

Figure 3: Biased observer model (a) and continuation games (b)-(d)

In the proof of Proposition 2, we construct a fine subversion for the "worst case" biased observer game with $b = b^*$ that is depicted in Figure 3(a). By the subset property, it is also subversive for any game with $\mathcal{R} \supseteq \mathcal{U}$ and $\mathcal{A} \supseteq \mathcal{L}_{b^*}$, i.e., it is a universal subversion for the biased observer model. As described in the Appendix, this conversation 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. These three games can each be solved in a similar way as the biased committee model of the previous section.

Panel (b) of Figure 3 is the worst case biased committee game, described in Section 2.2 and solved in Figure 2. The game 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 depicted in panel (c) 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 between costs and benefits. As shown in the proof of Proposition 2, these two continuation games in panels (c) and (d) can be solved via a recursive conversation like the one described in Figure 2.

The constructive, geometric approach we employ has the benefit of identifying additional economically interesting situations, beyond the biased committee and biased observer models, where subversion is also possible in similar ways. In the next section we provide some invariance properties of subversive conversations that allow us to further extend our existence results by showing how different games are connected by their subversive conversations.

2.4 Invariance properties

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 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

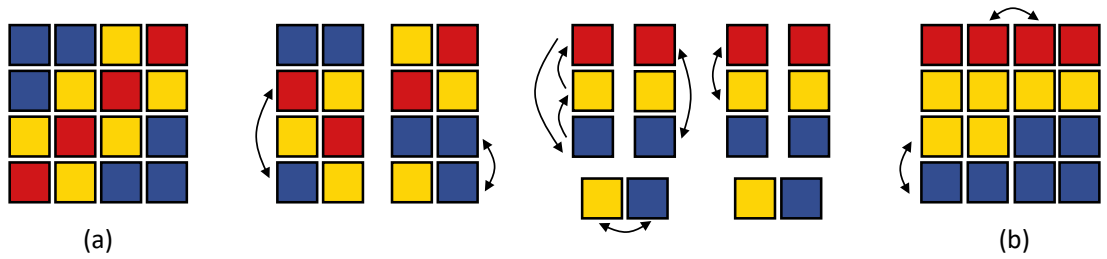

Figure 4: Relabeling property

established the existence of universal subversions in the models of Sections 2.2 and 2.3. 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. 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.[15]

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 two 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, as shown in the figure. These permutations are inessential relabelings of the residual type spaces at every continuation game generated by the fixed conversation. We now formally define this *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 $i = X, Y$, following history h^t, so that $\mathcal{S}\left(h^t\right) = \mathcal{S}_X\left(h^t\right) \times \mathcal{S}_Y\left(h^t\right)$ is the residual state space generated by σ after history h^t. This also defines a continuation game $\Gamma\left(h^t\right)$, using intersections of $\mathcal{S}\left(h^t\right)$ with the sets $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$.

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.[16] 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 permuations. 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 original priors F are invertible, an example of such a rescaling is a transformation to quantiles. The rescaling described in the context of Figure 2 and used to establish Proposition 1 is another example.

Relabeling property. A subversive conversation σ for Γ is also subversive for Γ' obtained from

[15] 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.

[16] 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.

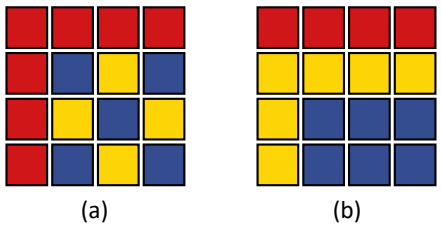

Figure 5: Decision-measurability property

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.

Figure 5(a) 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 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. The decision-measurability property 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).

We apply these invariance properties throughout the paper. Each subversive conversation for a given game defines a larger set of games where subversion is possible in the same way. The original conversation is then a universal subversion for the model defined by this larger set.[17]

[17] Subversive conversations are also invariant to interchanging the roles of the two players (rotations of the state space) as well as the cardinal specification of the players' payoffs.

2.5 Extensions

In this section we briefly discuss some extensions of our model. First we consider settings with non-iid priors, including correlation. Next, we consider cases where the sign or magnitude of the conflict between the players and the observer is uncertain.

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, then the transformation to quantiles via the relabeling property yields $\mathcal{A} \supseteq \mathcal{L}$ so that by the decision measurability property we obtain Proposition 1. Alternatively, suppose P and Q are tail-symmetric.[18] 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. Looked at in another way, 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 exactly this symmetry to prove Proposition 1.

What if x and y are not statistically independent? Suppose 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. It allows us to interpret x as a benefit and y as a cost of accepting the proposal.

Proposition 3 *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.*

The proof follows from these observations. A conversation truncates the state space into smaller continuation games which inherit $\mathcal{A}$-monotonicity. The monotone likelihood ratio properties from the proposition statement are also inherited by them. It is easy to see that decision measurability extends to correlated priors. At the decision stage of the conversation, since g' puts higher weight on larger values of the benefit x and smaller values of the cost y, relative to g, both conditions of the decision measurability are satisfied under g', exactly because they were met under g. So (DC) will obtain and the same conversation will continue to be subversive.

We note that statistical independence of x and y is, in a sense, the most interesting case since neither player has an advantage over the observer when it comes to decoding a message sent by the other player. Whatever Y learns about x, so does the observer. In contrast, correlation may

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

give the players an advantage over the observer, effectively allowing encryption. For instance, if x and y are either both less than $1/2$ or both greater than $1/2$, the players can publicly convey information about each other's signals that will not be accessible to the observer.

Uncertain preferences. So far we have assumed the sets $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$ are common knowledge. What would happen if, for instance, it was common knowledge the players were playing a biased committee game but the exact specification of $\mathcal{R}$ was the private information of the committee (and that of $\mathcal{A}$ the private information of the observer)? Such uncertainty about preferences would have no effect on our results because we establish the existence of a universal subversion for the biased committee model. The players could play according to this conversation, regardless of their private information, and deniability will be met when they take decisions. The same is true of the biased observer model. More generally, the invariance properties implicitly describe a model for which a given conversation is universal and thus provide a description of the uncertainties that the conversation is robust to.

The preceding discussion assumes the direction of conflict is common knowledge—if the committee prefers to reject the proposal then the observer prefers to reject it too. We now relax this assumption in a variant of the biased committee model in which the players may be more or less inclined than the observer to accept the proposal. Consequently, there are two kinds of deniability constraints, one where the proposal is accepted and another where it is rejected.

Let $\mathcal{S} = [0,1]^2$, $P = Q$, and denote by $\mathcal{A}$ and $\mathcal{R}$ the acceptance and rejection zones where everyone agrees on the optimal decision. Let $\mathcal{C}_0$ be the set of states where the committee prefers to reject the proposal while the observer prefers to accept it, with $\mathcal{C}_1$ be the zone of conflict where the committee prefers to accept but the observer prefers to reject. Like the biased committee model, we assume that the states where the observer likes to accept, $\mathcal{A} \cup \mathcal{C}_0 = \mathcal{L}$. But unlike the biased committee model, we now allow $\mathcal{C}_0$ to be non-empty. The committee prefers to accept when the state is in $\mathcal{A} \cup \mathcal{C}_1$ and we assume this set is of the form $\{(x,y) \mid y \leq C(x)\}$ for some continuous function $C(.)$. So $y = C(x)$ forms the border between the states where the committee prefers one decision or another and we will refer to it as the *committee's decision line*.

Proposition 4 *Assume $\mathcal{A} \cup \mathcal{C}_0 = \mathcal{L}$ and $P = Q$. A subversive conversation exists for any increasing committee decision line $y = C(x)$.*

Proposition 4 is illustrated by Figure 6. Relative to the observer, the committee is biased in favor of acceptance when $C(x) > x$ and biased in favor of rejection when $C(x) < x$. In Figure 6, the committee's decision line crosses the diagonal $y = x$ at the point (x', y'). Note that X knows the direction of conflict based on her own information x. She can reveal this fact at the beginning of the game, e.g., by disclosing whether $x \leq x'$ or $x > x'$. In the former case, Y can reject the proposal if $y > y'$; and if $y \leq y'$ the residual state space is an instance of

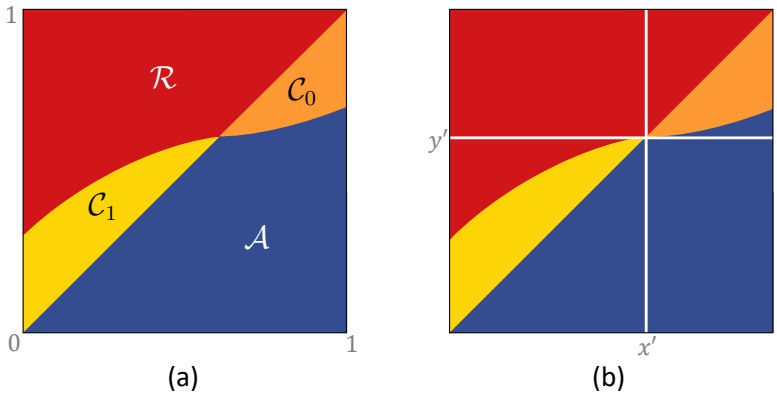

Figure 6: Uncertain direction of conflict

the biased committee model (once appropriately rescaled) and so subversion 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" (or row permuted) version of the biased committee model, with the labels "accept" and "reject" reversed. By Proposition 1 subversion is possible in this subspace as well. It is easy to see that the same argument holds as long as $C(x)$ is increasing, 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.

More generally, in any subversive conversation for any game, the player taking the decision must know the direction of conflict. In particular, she knows either that the state does not belong to $\mathcal{C}_0$ (if she is about to accept the proposal) or that the state does not belong to $\mathcal{C}_1$ (if she is about to reject it). Since this information will be inferred from the decision itself, the player can reveal it (if it has not already been revealed) and still meet deniability. So subversion is possible under uncertainty about the direction of conflict 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. For this reason, in the rest of the paper we study the case where the direction of conflict is common knowledge at the outset.

3 Conversational dead-ends

A subversive conversation implements the committee's ideal decision rule with probability one in finite time. To do so, it is necessary to maintain non-negative slack in every continuation game generated by the conversation. In this section, we describe games where subversion is impossible in a strong sense—every protocol that maintains non-negative slack in every continuation game and meets the deniability constraint allows the committee to take its optimal decisions with zero probability. We call these games *dead-ends* and provide a characterization. To do so we restrict

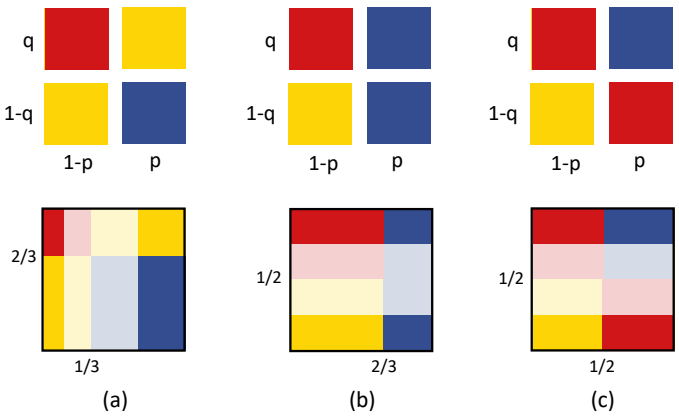

Figure 7: Self-similar dead-ends

attention to finite type problems in which $\mathcal{S}_X$ and $\mathcal{S}_Y$ are both finite.

Figure 7 depicts three types of games, in each of which each player has binary signals. We assume that the prior probabilities p and q are such that the problem is *tight*, i.e., $\Pr[\mathcal{A}] = \Pr[\mathcal{C}]$. Following the approach introduced in Section 2.2, we can perform a quantile transformation so that x and y are uniformly distributed on $[0,1]^2$. Randomized messages in each original 2×2 game correspond to partitions of the quantile space $[0,1]^2$. The lower panels of Figure 7 depict these transformed games for specific choices of p and q that ensure tightness.

For the game in panel (a), the players would like to accept the proposal if at least one signal is "good", while the observer requires both signals to be so. For panel (b), the players have the same preferences, whereas the observer is willing to accept only if X's signal is good. In panel (c) the observer has the same preferences as panel (a), while the players only need the two signals to "match". We show first that the conflict between the committee and the observer precludes the existence of a subversion conversation in each of these three games.

Suppose by way of contradiction that a subversive protocol exists for game (a). Player X can only take a decision in round 1 for types $x > 1/3$ because she knows her optimal decision only for these states. If she takes a decision for some positive measure subset of this part of the state space, there must be negative total slack in the continuation game on the part of the state space where X does not take a decision. Subversion is then impossible in that continuation game. So X cannot take a decision with positive probability in round 1. She can only partition the state space into two or more elements, each of which is tight. Such a two-element partition is depicted in the lower left panel of Figure 7 by the shaded and unshaded regions.

By the relabeling property, each partition element can be rescaled back to the original game. But for each such element a decision cannot be taken with positive probability because this was not possible for the original game. Continuing the argument, we conclude that a decision can never be taken with positive probability. Since a subversive protocol must take a decision with probability one in finite time, this yields a contradiction. The same argument extends

to the games (b) and (c), regardless of which player starts the conversation. In fact, not only is it impossible to implement the committee's optimal decision rule in a.s.-finite time but any protocol that maintains non-negative slack in each continuation game (a necessary property of a subversive protocol) can never take optimal decisions with positive probability. Thus, each game of Figure 7 is a dead-end.

If a subversion exists in any game, then it cannot generate a dead-end with positive probability as a continuation game. The games depicted in Figure 7 are dead-ends that have an additional property—any protocol that decomposes them into smaller games while maintaining non-negative slack results in continuation games that are copies of the original game. They are *self-similar dead-ends.* Our next result shows self-similar dead-ends underlie all possible dead-ends.

Proposition 5 *Any finite type game* Γ *that is a dead-end can be partitioned via a conversation into self-similar dead-ends.*

The examples depicted in Figure 7 are the only possible self-similar dead-ends, up to some transformations. Any other ones are either permutations (the relabeling property applied to the null history) or rotations (see footnote 17) of these games, or they can be obtained by switching the sets $\mathcal{A}$ and $\mathcal{C}$. By decision measurability, such a switch does not affect the possibility of subversion since any dead-end must be tight. As such, Proposition 5 shows that the three examples of Figure 7 constitute the irreducible set of non-existence examples in which no progress is possible. However, it does not provide conditions that allow us to determine a priori whether a particular game allows for subversion or not. We turn to this question now.

4 A necessary and sufficient condition

In this section we provide a necessary and sufficient condition on the priors for the existence of subversive conversations. We use a belief-based approach and apply the techniques and results of Aumann and Hart (1986). To do so, we restrict attention to *finite problems*, 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, $\widetilde{\mu}_X^t = \widetilde{\mu}_X^{t-1}$ a.s. for t even, and $\widetilde{\mu}_Y^t = \widetilde{\mu}_Y^{t-1}$ a.s. for t odd, $t \geq 1$. Thus, $\widetilde{\mu}$ is a bimartingale (Aumann and Hart, 1986).[19] Conversely, given a bimartingale $\widetilde{\mu}$ with expectation μ^0, one can derive a protocol $\xi = (\sigma, \alpha)$ from it, using Bayes Rule recursively.[20]

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 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, $\mu[\mathcal{R}^c] = 0$ 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.[21]

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 main task in this section 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 apply the techniques of Aumann and Hart (1986). Any $B \subset \Delta(\mathcal{S}_X) \times \Delta(\mathcal{S}_Y)$ is 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 $B = ns_{\mathcal{T}}(B)$. By Theorem 4.3 in Aumann and Hart (1986), we have the following necessary and sufficient condition for the existence of a subversive conversation.

Proposition 6 *Consider a finite problem and fix $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$. A subversive protocol exists if and only if the prior $\mu^0 \in bico^{\#}(\mathcal{T})$.*

[19] 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.

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

[21] 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 $\mathcal{R}$.

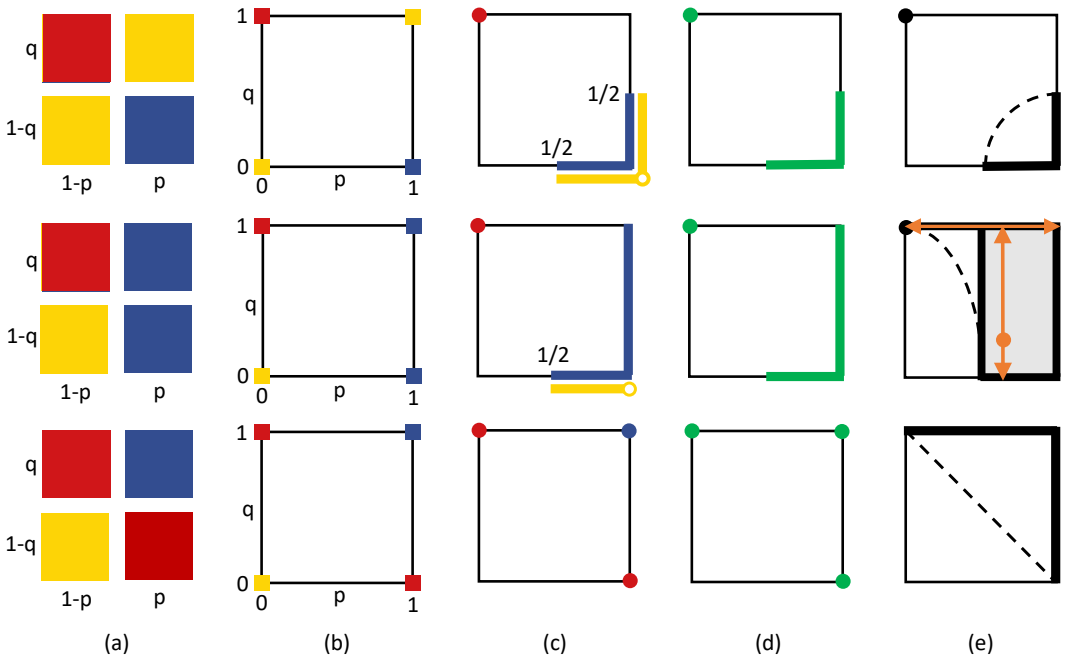

Figure 8: Biconvex hulls of subversive terminal beliefs

Figure 8 illustrates Proposition 6. Panel (a) presents the non-existence examples we introduced in Figure 7 but now allowing for strictly positive (or negative) total slack. The pair $(p, q) \in [0, 1]^2$ pins down any posterior belief $\mu \in \Delta(\mathcal{S}_X) \times \Delta(\mathcal{S}_Y)$, as depicted in panel (b). Panel (c) depicts the set of possible subversive terminal beliefs $\mathcal{T}(s)$, $s \in \mathcal{S}$.

To see how these sets are constructed, consider the top example of Figure 8 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, for $s \in \mathcal{C}$ that is in the bottom row, $\mathcal{T}(s)$ must then be of the form $(p, 0)$ with $p \in [1/2, 1)$, as depicted in the figure; and for the state $s \in \mathcal{C}$ that is in the top row, $\mathcal{T}(s)$ must be of the form $(1, q)$ with $q \in (0, 1/2]$. Similarly, 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 $q \in [0, 1/2]$. The terminal belief sets $\mathcal{T}(s)$ for the other examples are constructed analogously.

Panel (d) depicts the union $\mathcal{T} = \cup_{s \in \mathcal{S}} \mathcal{T}(s)$, while panel (e) depicts its biconvex hull $bico(\mathcal{T})$.[22] In general, $bico(\mathcal{T}) \subseteq bico^{\#}(\mathcal{T})$, but for these examples $bico^{\#}(\mathcal{T}) = bico(\mathcal{T})$.[23] By Proposition

[22] In general, $bico(\mathcal{T})$ 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).

[23] To show this, it suffices to find a bounded biconvex function $f : \Delta(\mathcal{S}_X) \times \Delta(\mathcal{S}_Y) \to \mathbb{R}$ that separates $bico(\mathcal{T})$ from its complement (see Proposition 4.1 in Aumann and Hart, 1986). For the top example of Figure 8, we refer the reader to Example 3.4 in Aumann and Hart (1986) for the relevant separating function (subject to a rotation). For the middle example, $f(p, q) = \max[\frac{1}{2} - p, 1 - q]$ is such a function. For the last example, we may take $f(p, q) = \max[1 - p, 1 - q]$.

6, a subversive conversation exists for these examples if and only if $\mu^0 \in bico^{\#}(\mathcal{T}) = bico(\mathcal{T})$.

For the top example, 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 total slack and the problem is a self-similar dead-end, as shown in Section 3. Starting from any interior point with positive total slack that lies below the zero slack curve, the posteriors can never enter $bico^{\#}(\mathcal{T})$ with probability one and so subversion is not possible. The players can take optimal decisions for a subset of the states, while maintaining non-negative slack in the complementary set, approaching (or reaching) some self-similar dead-end (a point on the zero slack curve). Similar remarks apply to the bottom example where subversion is also impossible for any interior priors since $bico^{\#}(\mathcal{T})$ has an empty interior.[24]

For the middle example, $bico^{\#}(\mathcal{T})$ contains interior priors. Panel (e) shows a bimartingale that starts within $bico^{\#}(\mathcal{T})$ and reaches $\mathcal{T}(s)$ for each s. In this conversation, Y first reveals her type, then X takes the appropriate decision. Of course, subversion is impossible for priors outside $bico^{\#}(\mathcal{T})$ since posteriors cannot enter this set with probability one.

Proposition 6 follows from the results of Aumann and Hart (1986). The only input is our specification of the sets of possible terminal beliefs $\mathcal{T}(s)$. Since the decision rule is fixed for a subversion and incentive constraints have no bite, it is relatively straightforward to specify these sets. The same approach 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 $\mathcal{T}^{fine}(s)$ 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 $\mu = (\mu_X, \mu_Y) \in \mathcal{T}^{fine}(s)$, $s = (x, y) \in \mathcal{S}$. Proposition 6 can then be amended to conclude that a fine subversion is possible if and only if $\mu^0 \in bico^{\#}(\mathcal{T}^{fine})$, where $\mathcal{T}^{fine} = \cup_{s \in \mathcal{S}} \mathcal{T}^{fine}(s)$.[25]

A similar belief-based approach can be used to pursue other extensions of our model. It is easy to see how the terminal belief sets $\mathcal{T}(s)$ can be amended in order to allow for alternative burdens of proof different from the balance of probabilities notion that underlies (DC). This approach can also be used beyond subversion, for instance to characterize all possible decision rules that can be implemented via conversations, or to characterize the optimal decision rule when a problem has negative slack and subversion is impossible. If one assumes the players commit to a conversation, the bimartingales approach of Aumann and Hart (1986) applies. Absent commitment, incentive constraints will have a role to play. One would then need to adapt the dimartingales approach

[24] For this example, $\mathcal{T}(s)$ is empty for $s \in \mathcal{C}$ and so subversion is only possible for priors that put zero weight on the row or column that contains s, as the figure shows.

[25] For the 2×2 type examples of Figure 8, this distinction between $T(s)$ and $T^{fine}(s)$ is immaterial, as panel (c) of the figure shows. But it matters when the space of posterior beliefs $\Delta(\mathcal{S}_X) \times \Delta(\mathcal{S}_Y)$ is higher dimensional.

introduced by Aumann and Hart (2003) to two-sided private information and take into account active incentive constraints. 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). Our focus on first-best outcomes for the senders implies that there is never an incentive to deviate. Hence our results on existence and non-existence of subversive conversations apply equally when commitment is possible, as in the Bayesian persuasion literature (Kamenica and Gentzkow, 2011).

The players in our model have the same preferences, are each informed on one dimension, and need to share enough information to reach a decision while being scrutinized by an observer. Unlike the team-theoretic literature on organizations that focuses on internal communication frictions (Bolton and Dewatripont, 1994; Dessein and Santos, 2006), external scrutiny is the driving force of our model.[26] Krishna and Morgan (2001) and Battaglini (2003) show how multiple senders with the same information and different preferences can be made to reveal all information to a receiver. This is the opposite of our setup which also differs from the case of a single sender with multidimensional information making credible tradeoffs across dimensions (Chakraborty and Harbaugh, 2007, 2010).

Forges (1990a) and Aumann and Hart (2003) 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. We study polite talk for which Aumann and Hart (2003) and Krishna and Morgan (2004) do not present results under incomplete information. Our model of back-and-forth communication by two informed players also differs from Golosov, Skreta, Tsyvinski and Wilson (2014), and Chen, Goltsman, Horner, and Pavlov (2017).

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. Their focus is on the optimal design of the committee by an uninformed principal. They introduce a notion of agreement between committee members that they call consensus, which takes into account information revealed by the play of the game. In this respect, plausible deniability is similar, although we impose an additional

[26] So we have multiple audiences (Farrell and Gibbons, 1989), one of which has private information (Watson, 1996).

requirement that an outside observer must also consent after observing the messages. Absent this requirement, there is no role for a back-and-forth conversation in their setting.

To provide a necessary and sufficient condition for subversion, we employ the belief-based approach (Forges, 1990a, 1990b; Aumann and Hart, 2003; Lipnowski and Ravid, 2020) that is surveyed by Forges (2020).[27] Because of our focus on subversion, incentive constraints play no role. This allows us to use the characterization of bimartingales and biconvexity in Aumann and Hart (1986). As we discuss in Section 4, the same approach can be directly applied to analyze alternative notions of subversion, and extended to non-subversive decision rules if one assumes commitment to the communication protocol. Absent commitment, when incentive constraints have a role to play, a belief-based approach that extends the analysis of dimartingales in Aumann and Hart (2003) to two-sided private information is likely to provide further insight.

Information is decentralized in our model. In contrast, Glazer and Rubinstein (2004) consider persuasion of a receiver by a single speaker informed about multiple aspects of a decision problem when the receiver can request evidence on one aspect. Assuming instead the sender has the choice of what evidence to reveal, Antic and Chakraborty (2023) apply Hall's Marriage Theorem to provide a necessary and sufficient condition for subversion. The matching of conflict and agreement points in these papers is reminiscent of strategic argumentation by a single expert to an uninformed receiver (Dziuda, 2011).

Within the large political economy literature, this paper is most directly related to studies 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 the receiver, when the conflict between the two groups is large enough. Our comparison of secure versus insecure communication is related to this transparency versus opacity distinction. In our different environment, we identify conditions when the committee can subvert even under transparency.

The cryptography literature on information-theoretic security studies related problems of hiding information from third parties who have opposing preferences to the players and infinite computing resources. Close to our motivation, the literature on "deniable encryption" (Beaver, 1996; Canetti et al., 1997) consider situations where the third party can force the players to decrypt the ciphertext. Plausible deniability is ensured if the players can generate an innocuous decryption. Our notion of deniability is more demanding since the third party understands the equilibrium meaning of all messages and is not fooled by an alternate decryption.

More closely related are secure multiparty computation protocols, e.g., mental poker (Shamir, Rivest and Adleman, 1979), and secret sharing algorithms (Shamir, 1979). Shamir's secret

[27] A key contribution of the belief-based approach 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 deniability constraint along with the requirement of public messages limits what can be done with both approaches.

sharing designs how information (i.e., shares/parts of a secret) are split between players to ensure that any one player's information reveals nothing about others' information.[28] In a two player environment, if one player publicly reveals her share of the secret, an observer will not be able to update his beliefs about the secret while the other player will have complete information. We do not give the players the freedom to design how information is partitioned between them and so more intricate communication protocols are needed for the players to subvert.

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, although no player has all the information necessary for the decision, 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 (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. We show that when communication is scrutinized, 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 a back and forth conversation can create a sequence of contexts that allows sufficient information to be shared between the players to take the right 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 $\mathcal{A} \supseteq \mathcal{L}$:

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

[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.

- 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 $\mathcal{A} \supseteq \mathcal{L}$. 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 2.4. 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 $\mathcal{L} \subseteq \mathcal{A}$. 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 $\Gamma_\square$, followed by the one in Figure 3(d), which we denote $\Gamma_\lozenge$. Recall that the game in Figure 3(b), which we denote $\Gamma_\triangle$, 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* $\Gamma_\square = \{\mathcal{A}_\square, \mathcal{C}_\square, \mathcal{R}_\square\}$, *where* $\mathcal{S} = [0,1]^2$, $\mathcal{A}_\square = \{(x,y) \in \mathcal{S} \mid x \geq 1/2, y \leq 1/2\}$, $\mathcal{R}_\square = \{(x,y) \in \mathcal{S} \mid y \geq x\}$ *and* $C_\square = \mathcal{S} \setminus \{\mathcal{A}_\square \cup \mathcal{R}_\square\}$.

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 $\Gamma_\square$ 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 $\Gamma_\square$, see panel (d), otherwise Y says "reveal" and X reveals x in round 3. ■

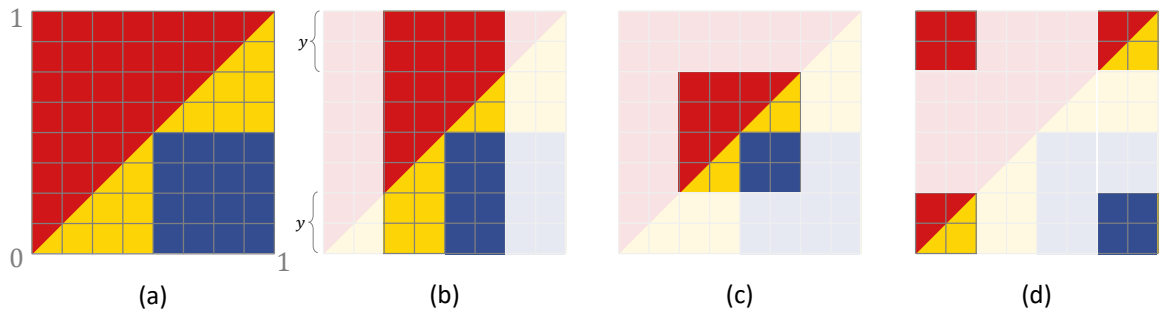

Figure 9: Recursive conversation for game $\Gamma_\square$

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

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 $\Gamma_\diamondsuit$ 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 $\Gamma_\diamondsuit$ through a set of bijections; see Figure 10(e). ■

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

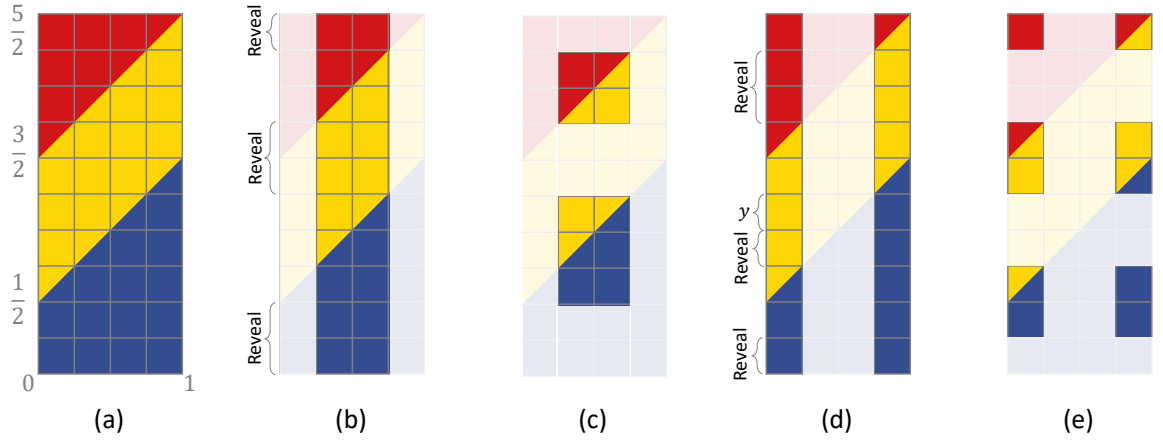

Figure 10: Recursive conversation for game $\Gamma_\diamondsuit$

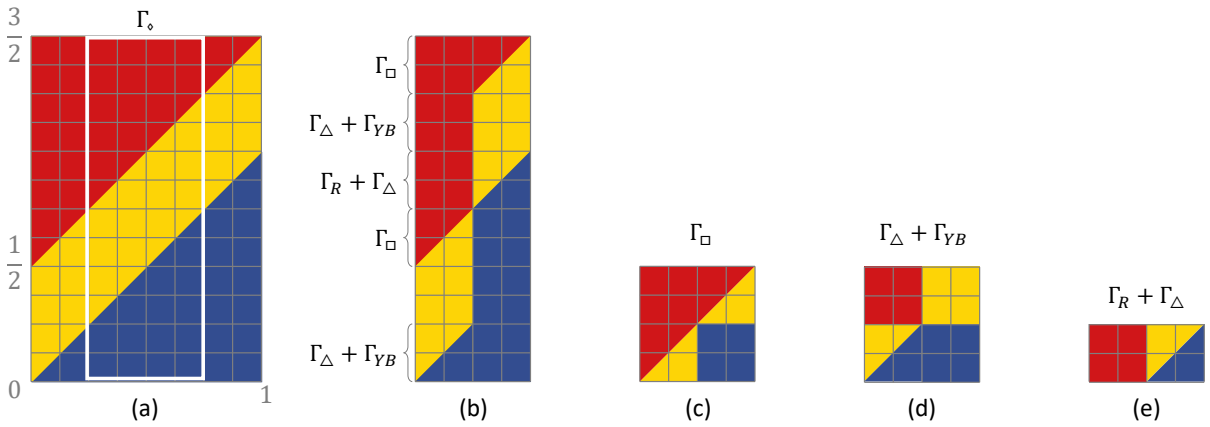

Figure 11: Recursive conversation for game $\Gamma^{\ell}_{\Diamond}$, with $\ell = 3/2$

Proof: Observe that $\Gamma^{5/2}_{\Diamond} = \Gamma_{\Diamond}$, 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 "$\Gamma_{\Diamond}$" since pasting the remaining states together yields $\Gamma_{\Diamond}$.

Consider $\ell < 5/2$. Panel (a) of Figure 11 shows the case $\ell = 3/2$. In period 1, X says "$\Gamma_{\Diamond}$" 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 residual game is a rescaled $\Gamma_{\Diamond}$, 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 "$\Gamma_{\square}$" 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 $\Gamma_{\square}$ 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 "$\Gamma_{\triangle} + \Gamma_{YB}$"; this is depicted in panel (d). After this X says "$\Gamma_{\triangle}$" 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 $\Gamma_{\triangle}$, 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 "$\Gamma_R + \Gamma_{\triangle}$" in round 2, X reveals the state if $x \leq \frac{5}{8} - \frac{1}{4}\ell$ or says "$\Gamma_{\triangle}$"; 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 $b = b^*$, 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 a. If $x \in (2b - k, 2b + k)$, in round 1 player X says "case a" and after Y rejects the proposal where appropriate, game $\Gamma^{\ell}_{\Diamond}$ remains, with $\ell = d/k + 3/2$. This is solved by Lemma 3.

Case b. 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 "$\Gamma_{\triangle}$" if $x \in [b, b + d)$, the second column in panel (b), after which $\Gamma_{\triangle}$ is played once Y rejects the proposal for

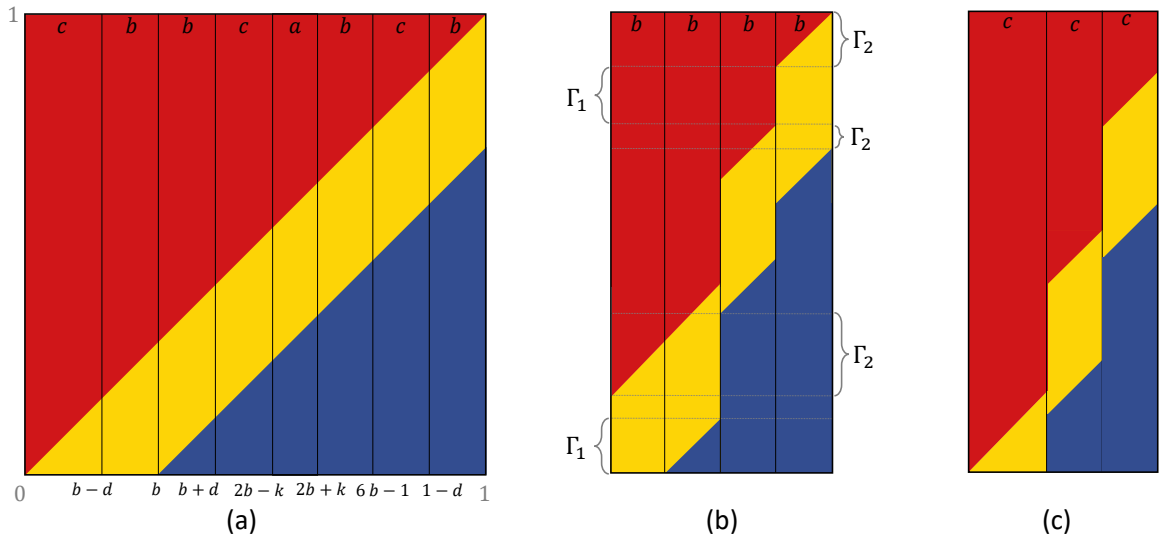

Figure 12: Construction for biased observer model

$y \in [6b-1, 1-d]$. 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 "$\Gamma_\square$" if $x \in [0, b+k] \cup [1-2b+d, 6b-1] \cup [1-d, 1]$ and $\Gamma_\square$ 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 "$\Gamma_\triangle$" if $x \in [2b+k-d, 3b+k]$ after which $\Gamma_\triangle$ 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 "$\Gamma_\Diamond^{3/2}$" and game $\Gamma_\Diamond^{3/2}$ is solved by Lemma 3.

Case c. 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:

"$\Gamma^\ell_\diamondsuit$" 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 $\Gamma^\ell_\diamondsuit$ 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 $x \in [\tau_{n-1} - \Delta\omega_n, \tau_{n-1}] \cup \left[\, x^r_n, x^r_{n-1}\right]$ and the state space is shown in **??**(b). In round 2, player Y (i) perfectly reveals $y \leq \tau_{n-1} - \Delta\omega_n$, (ii) says "$\Gamma_\square$" if $y \in [\tau_{n-1} - \Delta\omega_n, \tau_{n-1}] \cup \left[\, x^r_n, x^r_{n-1}\right]$ after which $\Gamma_\square$ is played, (iii) says "$\Gamma_R + \Gamma_\triangle$" 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 "$\Gamma_\triangle$" after which $\Gamma_\triangle$ is played, (iv) says "pass" (see panel (c) of the figure) after which X reveals x.

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

Proof of Proposition 3. Follows from the discussion in the text. ■

Proof of Proposition 4. Follows from the discussion in the text. ■

Proof of Proposition 5. We start with the following claim.

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

Proof of claim: Assume by way of contradiction that $\Pr[\mathcal{A} \mid \mathcal{S}, \mathcal{A} \cup \mathcal{C}] > \Pr[\mathcal{C} \mid \mathcal{S}, \mathcal{A} \cup \mathcal{C}]$. 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.

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 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 $\Pr[\mathcal{A} \mid \mathcal{S}, \mathcal{A} \cup \mathcal{C}] = \Pr[\mathcal{C} \mid \mathcal{S}, \mathcal{A} \cup \mathcal{C}]$, can be (maximally) reduced to a union of self-similar dead-ends from figure 7. 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 7, which describes all 2×2 dead-ends. ■

Proof of Proposition 6. 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

In this section we show that a universal subversive conversation requires at least four periods in the biased committee model and give an example of one that takes exactly four periods. We end with a short discussion.

B.1 The necessity of a conversation in the biased committee model

Proposition 7 *Any universal subversive conversation in the biased committee model requires at least four rounds.*

Proof. Assume by way of contradiction that there exists a three-round universal 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 instances of the model. While our positive results use pure strategies, we allow for mixed strategies in this impossibility proof.

A universal 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 $\{\sigma(x, h^0, \omega) : x \in [0,1/2), \omega \in [0,1]\}$. 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. ■

B.2 A four round conversation in the biased committee model

The following conversation takes four rounds and is a universal subversive conversation in the biased committee model. In round 1, Player X says "x is in $\{a, 1-a\}$" if $x = a$ or $x = 1-a$, for some $a \in [0,1/2]$. In round 2, Y perfectly reveals her type y if $y \in (a, 1-a)$; otherwise she says "pass". In round 3, X takes a decision if Y revealed her type y in round 2. Otherwise, X perfectly reveals her type x in round 3, following which Player Y takes a decision in round 4.

If player X takes a decision to accept in round 3, everyone knows that $x \in \{a, 1-a\}$ for some $a \leq 1/2$, as well as the exact value of $y \in (a, 1-a)$. We must have $a < 1/2$ since otherwise $(a, 1-a)$ is empty. Since $(1-a, y) \in \mathcal{A}$ in the biased committee model as $y < 1-a$, the worst-case situation is that $(a, y) \in \mathcal{C}$ when the proposal is accepted. But (a, y) and $(1-a, y)$ are equally likely, so (DC) is satisfied.

If player Y takes a decision to accept the proposal in round 4, from the history of the above conversation everyone knows the exact value of $x \in \{a, 1-a\}$ and the fact that $y \notin (a, 1-a)$. The residual part of $\mathcal{A}$ is at least the interval $[0, a]$, while the residual part of $\mathcal{C}$ is at most the interval $[1-a, 1]$. Since these two sets are of equal measure, (DC) is met.

B.3 Discussion

We note that the same conversation structure works for the three games in panels (b)-(d) in Figure 3. All of these games start with player X in round 1 pooling $x = a$ and $x = 1-a$, for some $a \in [0,1/2]$. We then have a game where one player's type (player X) is binary and the acceptance and rejection sets are monotonic. To solve these games, a slight generalization of the above is needed (in particular for game $\Gamma_\square$), that we now provide.

Let $x \in \mathcal{S}_X = \{l, r\}$, $l < r$, and let $\mathcal{S}_Y \subseteq \mathbb{R}$. 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 stick" and $\{r\} \times \mathcal{S}_Y$ as the "right stick".

Lemma 4 *Fix* $\Pr[\mathcal{C}] \leq \Pr[\mathcal{A}]$ *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 stick has non-negative slack.*

Proof. Suppose the right stick 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 stick has non-negative slack. If the left stick also has non-negative slack, X can just reveal x and Y can take a decision. So suppose the left stick has negative total 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 sticks 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 stick) 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. ■

Since the pooling $x = a$ and $x = 1 - a$ results in non-negative slack in the right stick, all of these games can be solved using the above conversation as well. The same argument however does not apply for the biased observer model (panel (a) in Figure 3). This is because for a sufficiently close to $1/2$ the binary state game will not satisfy $\Pr[\mathcal{C}] \leq \Pr[\mathcal{A}]$. In fact, no construction where pairs of X types are pooled together is possible for the biased observer model. Because the problem is tight, (DC) must hold with equality for all pooled pairs. However, for values of x close to 1 there is no other stick with sufficiently high conflict to pair it with which ensures zero slack in the deniability constraint.

The conversations in this appendix have the property that player X uses a continuum of messages in the first period. While Bayes Rule cannot be directly applied, this is not a problem as it can be applied via taking the appropriate limits (e.g., using the density). Our admissibility restriction on subversive protocols ensures that the law of iterated expectations continues to hold and hence arbitrary finite approximations of the protocol will still be subversive. The recursive conversation we present in the main text uses a a finite number of messages, until a player perfectly reveals her type and the updating is degenerate.

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] Antic, Nemanja and Archishman Chakraborty, "Selected Facts." *mimeo.*

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

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

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

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

[8] Bolton, Patrick, and Mathias Dewatripont. 1994. "The firm as a communication network." *The Quarterly Journal of Economics* 109(4): 809-839.

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

[10] Chakraborty, Archishman, and Rick Harbaugh. 2007. "Comparative Cheap Talk." *Journal of Economic Theory* 132(1): 70-94.

[11] Chakraborty, Archishman and Rick Harbaugh. 2010. "Persuasion by Cheap Talk." *American Economic Review* 100(5): 2361–2382.

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

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

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

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

[16] Dessein, Wouter. and Tano Santos, 2006. "Adaptive organizations." *Journal of political Economy* 114(5): 956-995.

[17] Dziuda, Wioletta. 2011. "Strategic Argumentation." *Journal of Economic Theory* 146(4): 1362–1397.

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

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

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

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

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

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

[24] Glazer, Jacob and Ariel Rubinstein. 2004. "On Optimal Rules of Persuasion." *Econometrica* 72(6): 1715–1736.

[25] Golosov, Mikhail, Vasiliki Skreta, Aleh Tsyvinski, and Andrea Wilson. 2014. "Dynamic Strategic Information Transmission." *Journal of Economic Theory* 151:304–341.

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

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

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

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

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

[31] 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.

[32] Lipnowski, Elliot and Doron Ravid. 2020. "Cheap Talk with Transparent Motives." *Econometrica* 88(4): 1631–1660.

[33] 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.

[34] Rivest, Ronald L., Adi Shamir, and Leonard Adleman. 1978. "A Method for Obtaining Digital Signatures and Public-key Cryptosystems." *Communications of the ACM* 21(2): 120–126.

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

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

[37] 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.

[38] Watson, Joel. 1996. "Information Transmission when the Informed Party is Confused." *Games and Economic Behavior* 12(1): 240–254.

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