Subversive Conversations*

Nemanja Antic† Archishman Chakraborty‡ Rick Harbaugh§

This version: April, 2022.

Abstract

Two players with common interests exchange information to make a decision. Their communication is scrutinized by an observer with different interests who understands the meaning of all messages and may object to the 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 provide a theory of conversations based on deniability in the face of possible public outrage.

JEL Classification: C72, D71, D72, D82.

Keywords: conversations, deniability, subversion, cheap talk.

*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, 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.

†Managerial Economics and Decision Sciences Department, Kellogg School of Management, Northwestern University, Evanston, Illinois; nemanja.antic@kellogg.northwestern.edu.

‡Finance Department, Syms School of Business, Yeshiva University, New York, New York; archishman@yu.edu.

§Business Economics and Public Policy Department, Kelley School of Business, Indiana University, Bloomington, Indiana; 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 could be accessible to other stakeholders. Communications between parties to a merger may be subpoenaed by anti-trust regulators. 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. Activists organizing under state surveillance need to watch what they say. Scientists trying to persuade a skeptical public must be careful about their private communications.[1]

When communication is public or exposure is a concern, can people still share enough information to determine their best action, or does decision-making suffer? We study this problem of communication under scrutiny. Two players with private signals and common interests must share information to decide whether to accept a proposal or not. An uninformed observer with misaligned interests sees the players' messages and decisions ex post. 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 maintain plausible deniability and avoid controversy, protests, interventions or penalties that may be imposed on them by the observer.

We consider the possibility of subversion. The players subvert when they do as well as they would if their communication was not scrutinized, while ensuring the observer never judges the decision to be against his own interest. Since the players have the same preferences, they must share enough information to determine their ideal decision while concealing enough information to maintain plausible deniability. Because of this deniability constraint, the players cannot immediately reveal their signals to each other so the problem of subversion is not straightforward.

We show how back-and-forth conversations must be used by the players in order to maintain deniability. As the conversation progresses, the players share increasingly detailed information, but only once the proper context has been created by previous statements. Over several rounds, the conversation allows the players to determine their preferred decision, while also hiding enough information to prevent any objections to the players' determination. The process of communication is important for subversion to be possible. A subversive conversation is an indirect mechanism that allows the players to get their first best outcomes, even in full view of an observer with different interests.

[1] Recent prominent exposures include the leak of Climategate emails by a server breach, subpoena of private documents in the VW Dieselgate and Purdue Pharma settlements, whistleblowing by a government employee that led to 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.

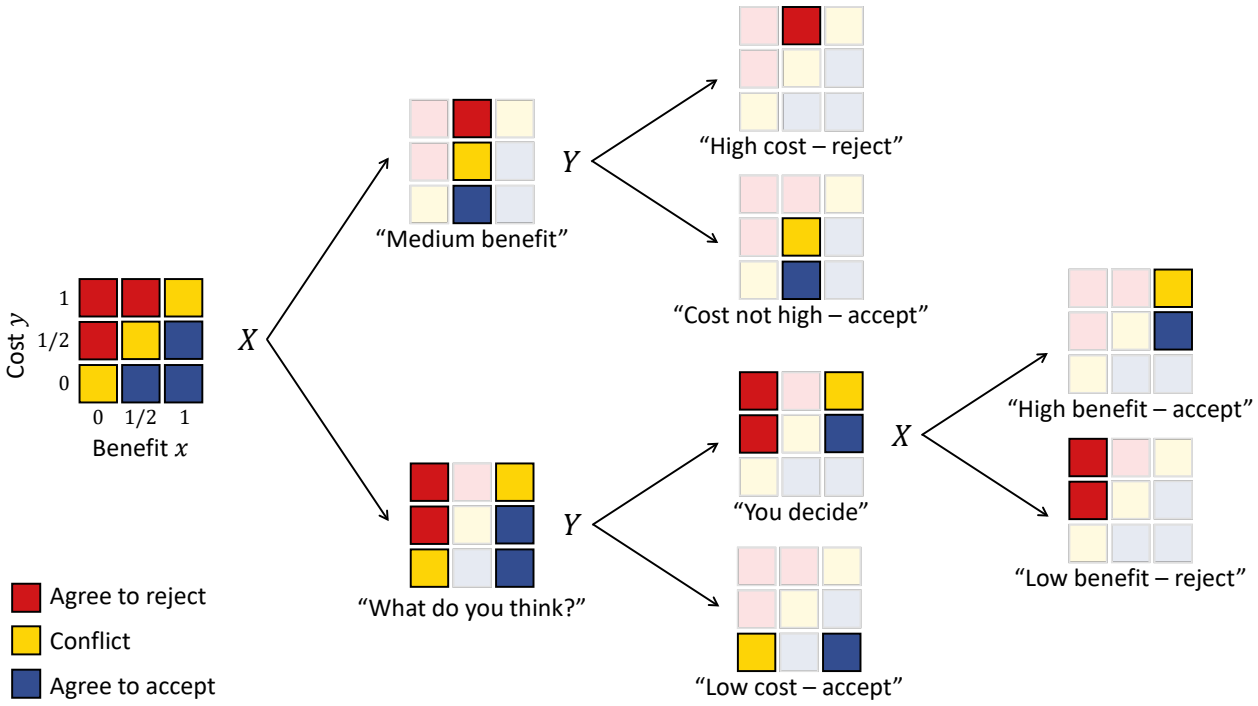

Figure 1: A conversation tree.

To see how a conversation can be subversive, suppose a committee of two managers is evaluating whether to accept or reject a new mining operation. The project has both environmental costs and economic benefits and the public (the observer) cares relatively more about the environment than the firm does. Manager X privately knows the project's (economic) benefit $x \in \{0, \frac{1}{2}, 1\}$ while manager Y privately knows the (environmental) cost $y \in \{0, \frac{1}{2}, 1\}$. The two managers have the same preferences and would like to undertake the project if the net benefits are such that the project is good ($x - y > 0$) or mediocre ($x - y = 0$), but not if it is bad ($x - y < 0$). The public favors a project that has at least an even chance of being good. Like all other examples in the paper, we assume uniform iid priors on x and y. The situation is depicted in Figure 1. The game has partial common interests since both the players and the public support good projects and oppose bad projects, but preferences for mediocre projects are in conflict.

The following conversation dynamically pools and separates types as the conversation progresses, allowing the managers to share enough information to determine whether the project is truly bad while concealing from the public whether it is good or only mediocre. In the first round of the conversation manager X speaks. If the benefit is $\frac{1}{2}$ then X just says so as seen in the first upper branch of the tree in Figure 1. In the second round manager Y then rejects the proposal if the cost is 1, in which case everyone opposes the project; or accepts it if the cost is 0 or $\frac{1}{2}$, in which case everyone knows the project is not bad, but only Y knows whether the project is truly good

or just mediocre. The pooled message shares enough information that the managers support the project, while hiding enough information from the public that they have no reason to oppose it.

If the benefit is either 0 or 1 then in the first round X passes the conversation over to the other manager. In the second round Y then accepts the project if the cost is 0, at which point the public favors the project if the benefit is 1 and opposes it if the benefit is 0, but only X knows which is the case and strategically says nothing further. Since there is an equal chance the project is good or mediocre, the public will not object. If instead the cost is $\frac{1}{2}$ or 1 then Y pools this information and tries to gauge what X now thinks. Having learned that the cost is not 0, in the third round X now rejects if the benefit is 0 and accepts if the benefit is 1. In the former case everyone agrees the project should be rejected. In the latter case the project is good or mediocre, but only manager Y knows which, so the public will not object to the decision.

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 this dynamic pooling and separating allows them to achieve their first-best outcome, there is no incentive to deviate from this strategy. As long as they follow the subversive communication protocol, any ex post scrutiny of their exchanges will not reveal that the managers knowingly acted against the public interest.[2]

In this paper we examine problems like in Figure 1 but with richer information structures and more general types of conflict between the players and the observer. For the players to be able to subvert, it is necessary (but not sufficient) that there be non-negative total slack—it must be at least as likely as not that the observer prefers to accept the proposal given that the players do. This condition is necessary and sufficient for a fully informed expert to subvert. Such an expert knows both pieces of information and so she does not need to reveal any information publicly to take her ideal decision. When the fully informed expert can subvert, she provides an upper bound on what the players can do.

We provide two economically natural classes of models, similar to the example above, where the players achieve this upper bound by using back and forth conversations. We establish two invariance properties of subversive conversations that can be used to extend existence results to an even wider class of models, including situations where preferences are not common knowledge. We also identify cases where subversion is impossible for the players, including situations where it is impossible even for the fully informed expert. In the latter case, the players do at least as well as the fully informed expert, often by using back and forth conversations. Throughout, our arguments are geometric.

Our results have implications for communication design within organizations when interests

[2]Since X cannot reveal $x = 0$ in round 1 and hope to avoid objections, it follows that a conversation is necessary for subversion in this example. Because the example is symmetric, a similar subversive conversation exists in which Y starts.

within the hierarchy diverge. For instance, a common problem in hiring committees is their tendency to self-replicate due to a bias toward candidates from similar backgrounds. To combat this problem, many institutions have implemented policies such as documentation of hiring explanations and auditing of committee communications. Our results imply that to eliminate bias in hiring it may be necessary to go beyond transparency. Transparency and greater stakeholder representation are also common concerns in the ongoing debate on corporate governance in the U.S. and other countries. Our results imply that worker representation in corporate boards may not be enough to prevent management interests from being fully served when management controls information flows and the deliberative process.

Instead of different levels of the organization having different preferences, they may share the same preferences but have a conflict with the public or an outside authority. Concerns about accountability to outsiders may lead executives to forego gathering information from subordinates, resulting in inefficiencies for the organization.[3] We show that suitable communication protocols can yield plausible deniability for leadership without any efficiency costs for the organization even when dispersed information is exchanged in full view. In the example above, if the managers follow the protocol and report their recommendation to the executive then any ex post scrutiny of their deliberations will not find that the executive had sufficient grounds to intervene and stop the project. Guiding questions by a supervisor with aligned interests can also prevent coordination failures, misunderstandings or untimely revelations. Organizations may encounter similar decision problems frequently. So they have an incentive to design communication protocols that will work for every realization of the state, together with rules and systems that ensure their implementation.[4] As a result, accountability and regulatory compliance may be more difficult to ensure, even for seemingly transparent organizations, than might otherwise be expected.

Our results offer insight into the role of secure versus insecure communication in commercial, governmental, and national security contexts. If the players could encrypt their messages they could communicate their exact information to each other privately and coordinate on their optimal decision.[5] Encoded communication is often prohibited by public policy or by internal corporate policy, or its usefulness is limited by subpoenas or FOIA laws. We assume that encryption is impractical or prohibited, i.e., that the observer understands the meaning of messages in the same way as the players. This amounts to assuming that in situations where the players are allowed to

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

[4] 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).

[5] In practice even "end-to-end" encryption fails if communication devices themselves can be corrupted.

use coded language, they can maintain deniability even if the code is somehow cracked and the intended meaning of messages deciphered by the observer. Encryption is not necessary because, unlike in the cryptography literature, we assume sufficient commonality of interest between all relevant parties.

The themes in this paper also relate to long-standing questions in political economy, especially those concerning the functioning of democracies. The Oxford English Dictionary defines subversion as "The undermining of the power and authority of a system or institution, e.g., the ruthless subversion of democracy". The observer in our model can be the median voter, on a usual Hoteling line, who has an ideological conflict with the players. The players can undermine the will of the majority but they do not need to use coercion. Instead, they manufacture consent via their control over the process of information exchange. Divergent interests between technocratic experts and the broader public have become a central concern for many policy issues, from free trade to Brexit to climate change. We show that the ability of experts to engage in fact-based decision making is less affected by public interference than might be expected. If open government, sunshine laws and other transparency regulations force deliberations out from behind closed doors, careful technocrats can still persuade the public, although the form of their communication may become more roundabout.[6]

The rest of the paper is organized as follows. Section 2.1 sets up our general model of subversion and Section 2.2 describes a baseline version of it. In the baseline model, the players subvert regardless of the conflict with the observer. They do as well as a fully informed expert. In sections 2.3 and 2.4 we extend this result to a large class of preferences and priors. Section 2.5 provides two invariance properties of subversive conversations that can be used to further extend our existence and robustness results. Section 3 describes situations where subversion is impossible and what can happen when it is. Section 4 reviews the literature. Section 5 contains our concluding remarks as well as a number of open questions suggested by this research. Appendix A contains proofs of results not contained in the main text, while Appendix B presents additional results on quick conversations and on subversion with correlated types.

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 $y \in \mathcal{S}_Y \subseteq \mathbb{R}$. We assume x and y are independent. Let $G(\cdot)$ denote the joint cumulative distribution function of (x, y) and let $\mathcal{S} \subseteq \mathbb{R}^2$

[6] The convoluted patterns of bureacratic 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."

denote the support of G. The probability measure represented by G can be continuous, discrete, or a mixture of the two.

The two players have common interests and they have 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 $u(x, y) \in \mathbb{R}$. Let $\mathcal{R} = \{(x, y) \in \mathcal{S} \mid u(x, y) < 0\}$ be the (measurable) set where the committee prefers to reject the proposal with $\mathcal{S} - \mathcal{R} \equiv \mathcal{R}^c$ its complement in $\mathcal{S}$ where it prefers to accept it.[7]

Since neither player is fully informed of the state $(x, y) \in \mathcal{S}$, they need to communicate with each other to determine their optimal decision. The players communicate through cheap talk. Time is discrete, with successive rounds indexed by $t = 1, 2, ...$ As long as a decision has not been taken earlier, and t is odd, player X may take a decision to either accept or reject the proposal, or she may not take any decision. Regardless, she sends a cheap talk message to the other player. Player Y does the same when t is even. Let M be the set of possible messages with $m_t \in M$ denoting a message sent in round t. Let $d_t = A$ denote a decision to accept the proposal and $d_t = R$ a decision to reject it in round t, with $d_t = N$ denoting the null (or no) decision. The game terminates as soon as a player takes a (non-null) decision.[8]

Let m^0 be the null history of messages, m^t a message history of length t, and M^t the set of such histories. Let $\mathcal{M}$ be the set of all histories of messages of arbitrary length. A *protocol* for for the committee has two components, a conversation and an action plan. A *conversation* is a map $\sigma \equiv (\sigma_X, \sigma_Y) : \mathcal{S} \times \mathcal{M} \to M$, where σ_i is measurable with respect to player i's information in the rounds where $i \in \{X, Y\}$ makes a move. An *action plan* is a map $\alpha \equiv (\alpha_X, \alpha_Y) : \mathcal{S} \times \mathcal{M} \to \{A, R, N\}$, where α_i is measurable with respect to player i's information in the rounds where $i \in \{X, Y\}$ makes a move. A protocol (σ, α) is a pure strategy profile, with (σ_i, α_i) the strategy of player i. This decomposition of strategy profiles into conversations (messaging stages) and action plans (decision stages) helps describe our results succinctly.[9]

We focus on *subversions*. In a subversion, the committee rejects the proposal if $(x, y) \in \mathcal{R}$ and accepts it if $(x, y) \in \mathcal{R}^c$. This is their first best optimal decision rule. So neither player has

[7] The possibility that the players could be indifferent has no bearing on our results and so we assume they always strictly prefer one action or the other. The decision taken by the committee is a collective action, although our results extend to some cases of individual decisions taken by each committee member, such as problems of pure coordination.

[8] We set payoffs to zero if they never take a decision. As is standard in cheap talk games, messages have no intrinsic cost or benefit and the message space M is rich enough that information transmission is constrained only by incentives. 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.

[9] We restrict attention to pure strategies since we can reformulate mixed strategies as pure strategies of a transfomed game. Our main results on existence do not rely on mixed strategies and they obtain under sequential communication, or *polite talk* in the language of Aumann and Hart (2003). In cheap talk games, anything that can be done with sequential communication can also be done with simultaneous communication (but not vice versa) by making one or the other player babble in every round.

an incentive to deviate from a protocol that implements this rule and it does not matter if the players can commit to the protocol or not.[10] As described so far, it is easy to create a protocol that implements such an action plan. Player X can simply communicate the value of x to Y who will then know the exact state (x, y). So she can take the committee-optimal decision. But we suppose that the players face a constraint. Their conversation and decision making will be observed ex post by another agent who may have a conflict of interest with the committee. We call this agent the observer ("he"). Because of the conflict of interest, the observer may have an incentive to object to the committee's decisions. The players must communicate in a manner that ensures that 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 from rejecting the proposal is set equal to zero. His payoff from accepting it is $v(x, y) = 1$ if (x, y) belongs to some (measurable) set $\mathcal{A} \subseteq \mathcal{S}$, with $v(x, y) = -1$ otherwise. So the observer prefers to accept the proposal if $(x, y) \in \mathcal{A}$ and reject it otherwise. We suppose that $\mathcal{A} \subseteq \mathcal{R}^c$ so that whenever the observer prefers to accept the proposal so does the committee (but not vice versa). Let $\mathcal{C} = \{\mathcal{A} \cup \mathcal{R}\}^c$ denote the set of states where the committee prefers to accept the proposal but the observer does not. This is the zone of conflict between the committee and the observer. In contrast, the sets $\mathcal{A}$ and $\mathcal{R}$ denote the acceptance and rejection zones, in each of which all parties agree on the decision.

Following a decision $d_{t+1} = R$ to reject the proposal the observer will infer $(x, y) \in \mathcal{R}$. In the rejection zone $\mathcal{R}$ all parties agree that rejection is best, so the observer never objects to such a decision. On the other hand, if the committee takes a decision $d_{t+1} = A$ to accept the proposal after a history of messages m^t, the observer will not want to object to the decision if and only if $\Pr[\mathcal{A} \mid m^t, d_{t+1} = A] \geq 1/2$. Since a decision to accept leads the observer to infer that $(x, y) \in \mathcal{R}^c = \mathcal{A} \cup \mathcal{C}$, we can rewrite the last inequality as

$$\Pr[\mathcal{A} \mid m^t, \mathcal{A} \cup \mathcal{C}] \geq \Pr[\mathcal{C} \mid m^t, \mathcal{A} \cup \mathcal{C}]. \tag{DC}$$

When the committee accepts the proposal, (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. So the committee maintains deniability that it was acting in the observer's interest. The deniability constraint is similar to the "balance of probabilities" burden of proof faced by courts in U.S. civil cases. When (DC) is met, the balance of probabilities favors "acquittal", i.e., allowing the committee to accept the proposal. So the committee will also be acquitted under the more demanding "reasonable doubt" burden of proof used for criminal trials.

Definition 1 *A subversive protocol (σ, α) implements the committee's optimal decision rule in*

[10] Our default assumption is of no commitment and our equilibrium notion is perfect Bayesian equilibrium (i.e., strategies are sequentially rational, with beliefs derived via Bayes Rule if possible and unrestricted if not).

finite time with probability one, while satisfying the deniability constraint (DC). A conversation σ is subversive if (σ, α) is a subversive protocol for some action plan α.

We conclude this section with a few remarks on the model. Condition (DC) is equivalent to saying that the measure of the residual part of $\mathcal{A}$, after deleting the states ruled out by the observed history of messages, is at least as large as the measure of the residual part of $\mathcal{C}$. These residual sets, after some states have been deleted, may in fact be subsets of $\mathbb{R}^1$, or even finite, and it may not be possible to apply Bayes' Rule. In all cases where the residual state space has zero measure in $\mathbb{R}^2$ we will follow standard practice and use generalized probability densities to compute posterior probabilities. A continuum of types can also create measure theoretic paradoxes that contradict the law of iterated expectations. To rule these out, we impose an admissibility restriction. For a given subversive protocol (σ, α), let $H_A^t(\sigma, \alpha) = \{m^t \in M^t \mid \alpha(s, m^t) = A,\ s \in \mathcal{S}\}$ be the set of all message histories m^t that terminate in round $t+1$ with a decision $d_{t+1} = A$ to accept the proposal. The deniability constraint (DC) must hold for each element of $H_A^t(\sigma, \alpha)$. Our admissibility restriction says that it must also hold when we integrate over all message histories that belong to any measurable $H \subset H_A^t(\sigma, \alpha)$.

We have assumed that the committee's deliberations are observed ex post. When (DC) is met the observer will approve of every committee decision. By the law of iterated expectations, the observer will have no incentive to object even after observing histories where the committee has not yet made a decision. So if the committee can subvert under ex post scrutiny, it can also subvert when its deliberations are observed contemporaneously. Indeed, we can allow the observer (and not the committee) to have formal authority over decisions, with the committee simply recommending a decision.

The only substantive restriction on the observer in our model is that he cannot design the procedural rules of committee deliberations. For instance, he cannot restrict the length and form of communication, constrain the message space, force a vote on the decision, or choose any other ex ante design aspect of the committee meeting. Procedural rules are either given by precedent or history, or the committee has the freedom to design them.[11] We use the device of an observer to provide one justification for the deniability constraint but the constraint could itself be a primitive. It could represent the players' own desire to avoid scandal or outrage and protect themselves against accusations of violating a social norm or acting against the public interest.

[11] The problem where the observer designs the committee meeting is interesting but relatively well understood. One can invoke the Revelation Principle and use direct mechanisms to characterize the decision rules that can be implemented (and associated payoffs). Our problem is more novel. The decision rule is fixed to be the committee's first best decision rule. The focus is on information revealed publicly by different indirect mechanisms that all implement this rule, and identifying the ones that maintain deniability.

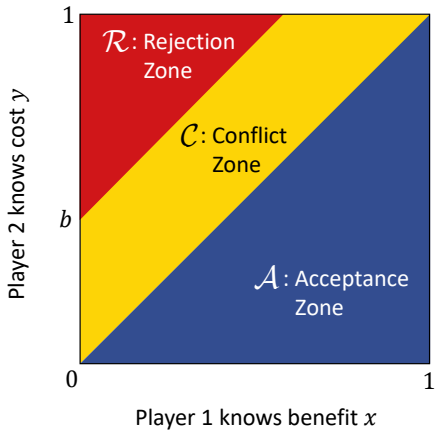

Figure 2: The state space for the baseline model.

2.2 Baseline model: uniform priors, constant conflict

We start by considering a simple baseline example of the abstract environment introduced above. Suppose G is the uniform distribution on $\mathcal{S} = [0,1]^2$. Let $\mathcal{A} = \{(x,y) \in \mathcal{S} \mid y \leq x\}$ and $\mathcal{R} = \{(x,y) \in \mathcal{S} \mid y > x + b\}$, where $b \in [0,1]$ is a preference parameter. We can think of the random variables x and y as the observer's benefit and cost. The observer prefers to accept the proposal if and only if $x \geq y$. Since $\mathcal{C} = \{(x,y) \in \mathcal{S} \mid x < y \leq x + b\}$, the committee has a bias in favor of accepting the proposal. The parameter b captures the size of this bias. The zones of acceptance, rejection and conflict are depicted in Figure 2.

Our first result establishes the existence of subversive conversations for this baseline model. It also shows that any attempt at subversion must involve a back and forth exchange, at least in some states (x,y). We say that a conversation requires t rounds (or is t rounds long) if it takes t rounds to take decisions with probability 1.

Proposition 1 *In the baseline model with uniform priors and constant conflict, (i) for each $b \in [0,1]$, there exists a subversive conversation, (ii) a subversive conversation requires four rounds for $b \in (0,1)$ sufficiently high and three rounds for any $b \in (0,1)$.*

That all subversive conversations need at least 3 rounds is easy to see from Figure 2. If $b \in (0,1)$, the players must exchange information to determine their ideal decision. However if $x < b$, X cannot reveal her type in round 1 and satisfy (DC) if the proposal is subsequently accepted in round 2. Player X must send some other message in round 1 without revealing her type. So the conversation must last at least one more round in order for the committee to determine its ideal decision and a decision cannot be taken before round 3. As long the players do not know their optimal decision ex ante ($b < 1$) and there is some conflict between the players and the observer ($b > 0$), back and

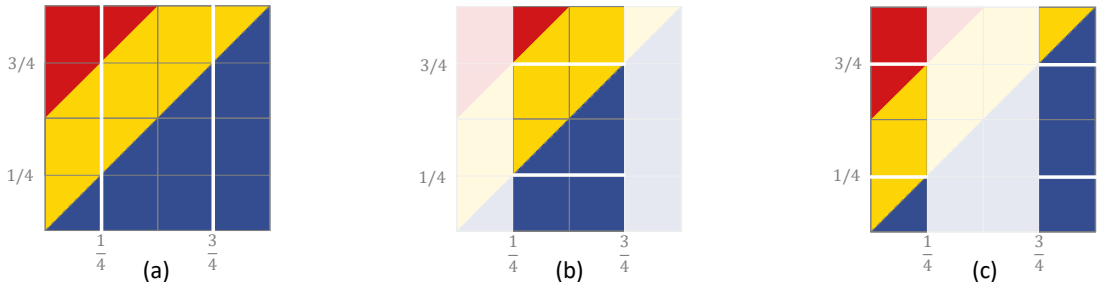

Figure 3: A subversive conversation for the baseline model.

forth communication is necessary for subversion. In the proof of Proposition 1, we show that when $b \in (0,1)$ is large enough a subversive conversation requires four rounds.

For the proof of part (i), we construct a subversive conversation, and an attendant action plan, for each $b \in [0,1]$. Figure 3(a) depicts the case $b = 1/2$. In round 1, X reveals either that x belongs to $[\frac{1-b}{2}, \frac{1+b}{2}] = [1/4, 3/4]$, or she states that $x \notin [1/4, 3/4]$. She does not reveal the exact value of x. In the first case we move to panel (b) of the figure while in the second case we move to panel (c). Each panel only depicts the states that are known to be possible given X's first message.

Consider panel (b). If $y \in [1/4, 3/4]$, Y knows the committee's optimal decision and she accepts the proposal immediately. Otherwise, she requests more information from her partner, following which X reveals the exact value of x and Y subsequently takes the committee's optimal decision. When Y accepts the proposal without learning x, everyone infers (x, y) belongs to $[1/4, 3/4] \times [1/4, 3/4]$. As can be seen from the figure, the sets $\mathcal{A}$ and $\mathcal{C}$ have equal measure conditional on this event, and so the deniability constraint (DC) is satisfied in this case. The deniability constraint is also satisfied when Y asks for the exact value of x and then accepts the proposal. Conditional on this event, the residual part of $\mathcal{A}$ is $\{x\} \times [0, 1/4)$ whereas the residual part of $\mathcal{C}$ is a subset of $\{x\} \times (3/4, 1]$ and the latter is (weakly) smaller in measure, as can be seen from the figure.

Consider now panel (c) where everyone knows $x \notin [1/4, 3/4]$. In this case, Y reveals the exact value of y whenever $y \in [1/4, 3/4]$. When $y \notin [1/4, 3/4]$, Y either reveals $y > 3/4$ or reveals $y < 1/4$, without revealing her exact type. In all cases, her partner has enough information to determine the committee optimal decision. The deniability constraint is always satisfied since the residual part of the agreement set $\mathcal{A}$ is at least as large as the residual part of the conflict set $\mathcal{C}$, as can be seen from the figure.

The key feature of the conversation depicted in Figure 3 is gradualism. Neither player reveals all that she knows immediately but instead waits for the right moment. They gradually partition the state space into increasingly smaller subsets. Each such subset is a context that determines how future revelations will be evaluated by the players and the observer. When a decision can be taken, and what that decision should be, are determined by the contexts created during the

preceding conversation. Since the right context can create slack in the deniability constraint when none existed initially, a well-designed conversation is the key ingredient for subversion.

To see the role of a proper context, note that X cannot reveal any $x \in [1/4, 1/2)$ in round 1 because the deniability constraint will not be met if Y accepts the proposal afterwards. But such a type can be safely revealed later in the conversation. The context created by X's first message "$x \in [1/4, 3/4]$" and Y's subsequent statement "$y \notin [1/4, 3/4]$" allows X to reveal any $x \in [1/4, 1/2)$ in round 3; while if $y \in [1/4, 3/4]$ no player needs to reveal her exact signal and Y can take the committee's optimal decision in round 2. In the gradual process of creating suitable contexts, each player must initially conceal unfavorable news so as not to run afoul of the deniability constraint. But each player must also conceal sufficiently good news in order to create a favorable context just in case her partner has bad news. A subversive conversation ensures either that information required for a decision can be safely revealed once the proper context has been created, or that it is not necessary to reveal it.[12]

We have described the conversation above in terms of the intended (and inferred) meaning of statements, i.e., in terms of subsets of the state space. But it is not hard to construct natural language versions of it. The following is one example. In round 1, when $x \in [1/4, 3/4]$, X says something like, "My news is intermediate, I think you should take the decision," to which Y responds either with, "We should accept," (when $y \in [1/4, 3/4]$) or with, "Tell me what you know!" (when $y \notin [1/4, 3/4]$). In the latter case, X reveals her exact signal and Y takes a decision. In contrast, when $x \notin [1/4, 3/4]$, X can start off by requesting Y's information. Player Y then reveals her exact type when $y \in [1/4, 3/4]$ and otherwise reveals only whether y is very low ($y < 1/4$) or quite high ($y > 3/4$). In all cases, X has sufficient information to take the committee optimal decision. In the proof of Proposition 1, we show that similar conversations yield a subversion for each $b \geq 1/3$, with decisions taken within four rounds. For $b < 1/3$, the protocol has a similar structure but the smaller conflict allows it to be slightly simpler and decisions can be taken within three rounds.

In our problem of subversion, each player holds one piece of information while their optimal decision depends on both pieces of information. The players have to exchange this dispersed information to determine their optimal decision. We can compare our committee with a benchmark case of a fully informed expert. Such an expert knows both x and y and has the same preferences as the players. So she can subvert whenever the committee can subvert because she can mimic the committee's protocol. But since the fully informed expert has all the information, she knows her optimal decision at the outset. She can simply take a decision without engaging in a conversation. This decision will meet deniability if and only if there is *non-negative total slack,* i.e., the agreement set $\mathcal{A}$ is at least as large in measure as the conflict set $\mathcal{C}$. The non-negative total slack condition is

[12] In the rest of the paper, we will identify these contexts variously, by the history of messages m^t that precede it, by the residual state space $\mathcal{S}(m^t)$ that remains possible after this history, as well as by the continuation game defined by this residual state space.

also necessary for the committee to subvert. But it is not sufficient. The players need a back and forth conversation to determine their optimal decision, while maintaining slack in the deniability constraint at all times. Under non-negative total slack, the committee can do at best as well as the fully informed expert. Proposition 1 shows that in the baseline model the committee does as well as possible. In what follows, we will extend this result to domains beyond the baseline model.

Since the committee's messages and decisions are both observed, we may say that the committee communicates in public. Proposition 1 allows us to compare public communication with secure private communication. Under secure private communication the committee's decision is observed but the observer does not have access to the messages exchanged by committee members. Since the players communicate securely in the knowledge that the observer will not observe their messages, it seems reasonable to suppose they will coordinate on their optimal decisions. Under this equilibrium selection, secure private communication is outcome equivalent to the case of a fully informed expert. Since the committee does as well as the fully informed expert whenever it subverts, Proposition 1 shows that public communication is as good as secure private communication, a result we will also extend beyond the baseline model in what follows. The ability to subvert in public has additional implications. It says that the committee would do well to engage in a subversive conversation, even when communication takes place in private, in order to protect itself against leaks or hacks. Conversely, in environments where private communication is not allowed but the players somehow engage in it, they can hide this fact and maintain the charade of public communication by employing a subversive conversation.

Subversive conversations require coordination on the amount and nature of information that is to be revealed at each stage. Acquiring information as and when needed for the conversation can avoid excessive or untimely disclosures. This can benefit activists seeking protection from state surveillance, or organizations avoiding regulatory oversight. Guiding questions from aligned interests during public hearings can also perform the same role. If the players only have the information they need, or can only answer the questions they are asked, the possibility of coordination failure is reduced.

2.3 Biased committee: the lower-triangular model

We now generalize the baseline model by relaxing the uniform priors assumption and allowing for the general committee preferences defined in Section 2.1. Suppose x and y have identical cumulative distribution functions, $F(\cdot)$, so that $G = F \times F$ and $\mathcal{S} \subseteq \mathbb{R}^2$. The probability measure represented by G can have atoms.[13]

With respect to the observer, the deniability constraint (DC) remains unchanged. But whereas

[13] In the Appendix, we extend our arguments of this section (and the next) to allow for statistical dependence between x and y, under a MLRP condition.

in the baseline model we assumed an unbiased observer, or $\mathcal{A} = \mathcal{L} \equiv \{(x, y) \in \mathcal{S} \mid y \leq x\}$, we now allow $\mathcal{A} \supseteq \mathcal{L}$ in order to handle atoms in the distribution of types more easily. As before, we suppose that $\mathcal{A} \subseteq \mathcal{R}^c$, so that $\mathcal{C} = \{\mathcal{A} \cup \mathcal{R}\}^c$ is the zone of conflict where the committee prefers to accept the proposal and the observer prefers to reject it. Thus, the committee is biased in favor of the proposal like in the baseline model but the conflict between the committee and the observer is no longer a constant parameter. The committee's preferences can be arbitrary, provided the sets $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$ are measurable. Since $\mathcal{L}$ is the lower region of $\mathcal{S}$ on or below the diagonal $y = x$, we refer to this model as the lower-triangular model.

Notice first 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), we can reinterpret quantiles as the true underlying types. At points of continuity of F this is a one-to-one transformation. For an atom, many new types (quantiles) may map into the same original (atomic) type, but this is immaterial since all these quantile types have the same payoffs as the atom. The transformed state space is $\mathcal{S} = [0, 1]^2$ and we can re-interpret the three sets $\mathcal{A}$, $\mathcal{C}$ and $\mathcal{R}$ as subsets of this new space. For the rest of the paper, we will employ this quantile transformation, i.e., assume without loss of generality that priors are iid uniform.[14]

With priors understood to be uniform, a particular instance of a model is fully specified by the configuration of preferences that determine the (measurable) sets $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$. We will refer to the triple $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$ as a *game* in the rest of the paper. Note that in the lower-triangular model, since x and y are identically distributed, the transformation to quantiles preserves the assumptions on preferences made in the original type space. In particular, $\mathcal{R}^c \supseteq \mathcal{A} \supseteq \mathcal{L}$ continues to hold in the space of quantiles.

Proposition 2 *Consider the lower-triangular model. There exists a conversation that is subversive for every game* $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$.

Proposition 2 states there exists a conversation σ which, when paired with a suitable action plan α, yields a subversive protocol (σ, α) in every instance $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$ of the diagonal model. The conversation itself does not depend on the game $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$. The timing of (non-null) decisions as specified by the action plan also does not depend on the game. But since the committee accepts the proposal when $(x, y) \in \mathcal{A} \cup \mathcal{C}$ and rejects it when $(x, y) \in \mathcal{R}$, and these sets vary from game to game, a specific decision prescribed by the action plan after a given history may depend on the game. The committee's preferences may change from one decision problem to another. Proposition

[14] Note that a mixed strategy for an atomic type can be represented as a collection of pure strategies for the quantiles that make up that atomic type. Note also that this transformation results in a continuum of types for each player with priors that admit a uniform density. We will use this density to compute posterior probabilities when Bayes' Rule cannot be applied.

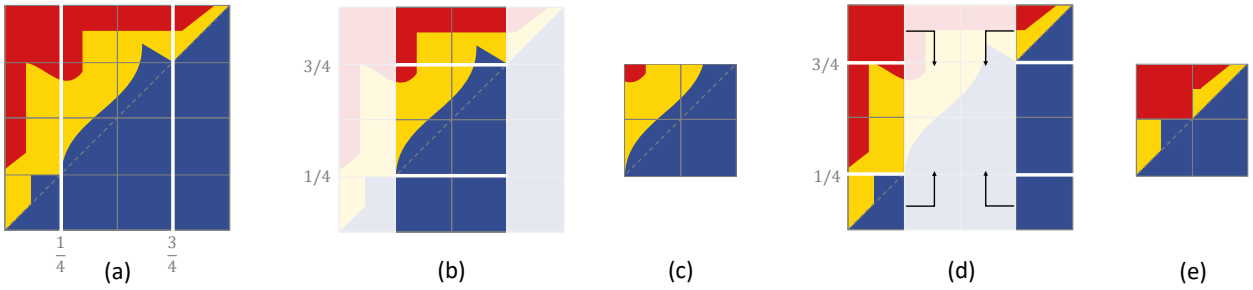

Figure 4: A recursive construction.

2 shows that the committee can use the same conversation in every problem, adapting only its specific decisions to the particularities of the situation.[15]

We prove Proposition 2 using a process that is similar to the one depicted in Figure 3, the only difference being that it is used recursively to cover the wider class of allowed preferences. Figure 4 depicts the argument. In round 1 of the conversation, X either says "$x \in [1/4, 3/4]$" or says "$x \notin [1/4, 3/4]$", regardless of the exact game $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$. She does not reveal her type. For the first message, we move to panels (b) and (c) of the figure while for the complementary message we move to panels (d) and (e). Each panel only depicts the states that are known to be possible given the history of messages.

Consider panel (b) first where it is common knowledge that $x \in [1/4, 3/4]$. When $y \notin [1/4, 3/4]$, Y reveals this fact and requests X's exact type. Subsequently, X reveals x and Y takes the committee's optimal decision. As can be seen from the figure, the deniability constraint is met for each possible revealed value of x, given the observed history of messages—the residual part of $\mathcal{A}$ is always at least as large as that of $\mathcal{C}$. On the other hand, when $y \in [1/4, 3/4]$, Y only reveals this fact. The residual state space at this point equals $[1/4, 3/4] \times [1/4, 3/4]$ and it is depicted in panel (c). If this residual state space is rescaled to equal the unit box, it becomes an instance of the lower-triangular model. This is one case of recursion in the protocol—the players can continue with the conversation in the same fashion in this residual state space (perhaps with one player announcing the rescaling).

Consider next panel (d) where it is common knowledge that $x \notin [1/4, 3/4]$. When $y \in [1/4, 3/4]$, Y reveals her exact type and X subsequently takes the optimal decision while meeting the deniability constraint. Otherwise, the residual state space equals $[1/4, 3/4]^c \times [1/4, 3/4]^c$. This residual state space can also be thought of as an instance of the diagonal model, once the four elements are "pasted together" and appropriately rescaled, as can be seen from panel (e) of the figure. This is

[15] While we formally allow the players to send a message *and* take a decision in every round, we adopt the convention that after any history m^t, if some types of a player take a non-null decision $d_{r+1} \in \{A, R\}$ that ends the game, then the accompanying message is the same regardless of the decision, history or game. This is equivalent to assuming the conversation terminates before any non-null decision.

the other recursive step in the protocol and the conversation continues as described above in the residual state space. Since each recursion takes place in a strictly smaller subset of the original state space (by a factor of 1/4), with decisions taken in the remaining part, decisions are taken with probability one in finite time. Anytime a decision is taken, the particular decision depends on the game$\{\mathcal{A},\mathcal{C},\mathcal{R}\}$. But the conversation that takes place before the decision does not depend on the game.[16]

The same conversation is subversive in every instance of the lower-triangular model because it has the following property—any time a decision is taken the player who takes the decision knows both x and y. The other player always reveals her exact type during the conversation that precedes the decision. Call a conversation *fine* if it always results in the player taking the decision being fully informed. Call subversive protocol (σ,α) a *fine subversion* if σ is a fine subversive conversation.[17]

Subset property. A fine subversive conversation for $\{\mathcal{A},\mathcal{C},\mathcal{R}\}$ is also a fine subversive conversation for any $\{\mathcal{A}',\mathcal{C}',\mathcal{R}'\}$ 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 with arbitrary sets. In particular, since $\mathcal{C}'\subseteq\mathcal{C}$ and $\mathcal{A}'\supseteq\mathcal{A}$, these inclusion relations will be preserved for the conflict and agreement sets that remain at some residual state space $\mathcal{S}\left(m^t\right)$ reached after history m^t. The set $\mathcal{S}\left(m^t\right)$ does not depend on the game given the fixed conversation that is employed in both games. So we must have $\mathcal{C}'\cap\mathcal{S}\left(m^t\right)\subseteq\mathcal{C}\cap\mathcal{S}\left(m^t\right)$ and $\mathcal{A}'\cap\mathcal{S}\left(m^t\right)\supseteq\mathcal{A}\cap\mathcal{S}\left(m^t\right)$. Since (DC) obtains in the original game $\{\mathcal{A},\mathcal{C},\mathcal{R}\}$, it must also be met in the new game $\{\mathcal{A}',\mathcal{C}',\mathcal{R}'\}$ that has a smaller (residual) conflict set and larger (residual) sets where the players and observer agree on the decision. Since the player taking the decision is always fully informed she can fine tune the decision to suit her preferences, just like she would be if there was no deniability constraint. It is the conversation that takes place before the decision that allows her to be fully informed and still meet deniability.

The subset property obtains beyond the lower-triangular model, and we will generalize it in Section 2.5. To use it in the lower-triangular model, notice that the conversation depicted in Figure 4 results in a fine subversion in the "worst case" scenario where $\mathcal{R}$ is empty, $\mathcal{A}=\mathcal{L}$ and $\mathcal{C}=\mathcal{L}^c$. The players want to accept the proposal regardless of the state. The conversation described in Figure 4 allows them to do that while meeting deniability. By the subset property, the same conversation is also subversive in every other instance of the lower-triangular model. Because the conversation is fine and the player taking the decision is always fully informed, only the action plan needs to be amended to take into account the specific game.

[16] In the proof of Proposition 2 we extend this argument to cover a set of models that contains the lower-diagonal model as a special case.

[17] We call a subversive conversation that is not fine a coarse subversive conversation, and the associated protocol a coarse subversion. An example is the conversation depicted in Figure 3 where the player taking the decision is not fully informed after some histories.

Fine subversions allow us to handle uncertainty about preferences. Consider, for example, the baseline model where the conflict is given by a constant b, but suppose that the observer is not certain of the value of b. If the committee uses the fine subversive conversation described in Figure 4 regardless of its bias b, and adjusts its action plan to suit the particular value of b, then it will be able to subvert regardless of observer beliefs about b. The observer will not be able to update his beliefs about committee preferences from the fixed conversation. As for the decisions, since the deniability constraint is met for every possible value of b, it will be met regardless of observer beliefs about these values. The argument extends to arbitrary kinds of uncertainty about committee preferences, including higher order uncertainty. It also covers (higher order) uncertainty about the observer's preferences, subject to staying within the ambit of the lower-triangular model, i.e., subject to it being common knowledge that $\mathcal{A} \supseteq \mathcal{L}$ in the space of quantiles. Call a conversation that is subversive regardless of beliefs about preferences *belief-independent*.

Corollary 1 *A belief-independent subversive conversation exists for the lower-triangular model.*

Belief-independence is an important robustness property because every aspect of preferences may not be common knowledge in reality. A belief-independent subversive conversation can avoid suspicion and second guessing the committee's actual motives because the conversation does not depend on these motives. But while Corollary 1 covers arbitrary kinds of uncertainty about preferences, it assumes the direction of conflict is common knowledge—since $\mathcal{R}^c \supseteq \mathcal{A}$, it is common knowledge that the committee is partial towards accepting the proposal, relative to the observer. It can never be the case that the committee prefers to reject the proposal while the observer prefers to accept it. We conclude this section by presenting another corollary of Proposition 2, where the direction of conflict between the committee and the observer is ex ante unknown. Depending on the state, the committee may prefer to reject the proposal when the observer prefers to accept it, and vice versa.

Let $\mathcal{C}_0$ be the set of states where the committee prefers to reject the proposal while observer prefers to accept it that we now allow to be non-empty. Let $\mathcal{C}_1$ be the zone of conflict where the committee prefers to accept but the observer prefers to reject and, as before, let $\mathcal{A}$ and $\mathcal{R}$ be the acceptance and rejections sets where everyone agrees on the optimal decision. When the state is in $\mathcal{A} \cup \mathcal{C}_1$ the committee prefers to accept the proposal. 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*. Similarly, the observer prefers to accept the proposal when the state is in $\mathcal{A} \cup \mathcal{C}_0$ and we may call the border between this set and its complement the *observer's decision line*.

With uncertain direction of conflict, a subversive conversation has to take into account two kinds of deniability constraints, one where the committee accepts the proposal and another where the committee rejects it. This situation is not directly covered by the lower-triangular model which

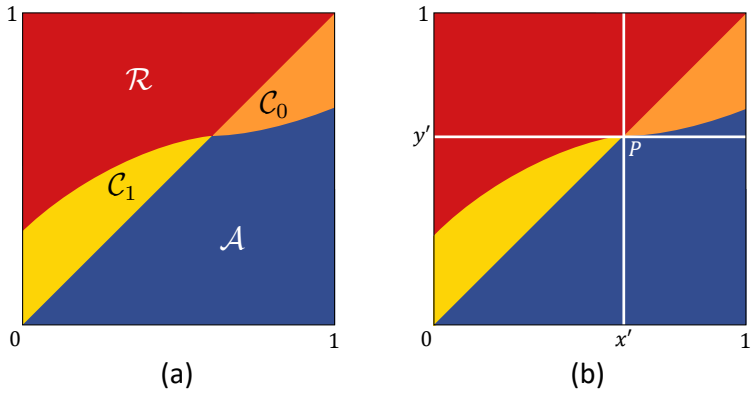

Figure 5: Uncertain direction of conflict.

only allows for one kind of conflict and so one kind of deniability constraint. Nevertheless, using our results for the lower-triangular model, the following corollary provides sufficient conditions under which a subversive conversation exists even in this case.

Corollary 2 *With uncertain direction of conflict and $\mathcal{A} \cup \mathcal{C}_0 = \mathcal{L}$, a subversive conversation exists for any increasing committee decision line $y = C(x)$.*

Corollary 2 is illustrated by Figure 5. 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 5, the committee's decision line crosses the observer's decision line $y = x$ at the point labeled P, with coordinates (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 the lower-triangular model (once appropiately rescaled) and so subversion is possible in this subspace via Proposition 2. On the other hand, if X reveals $x > x'$, then Y can accept the proposal if $y < y'$. Otherwise, when $y \geq y'$, the residual state space is just an "upside down" version of the lower-triangular model with reversed roles for the decisions to "accept" and "reject". By Proposition 2 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, some player will learn the direction of conflict at some stage during the conversation. For instance, just before taking a decision, the player taking the decision 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 also be inferred from the decision itself, the player can reveal it herself and still meet deniability. In any subversive conversation the

direction of conflict will be revealed ultimately. 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 any additional strategic issues for the problem of subversion. For this reason, in the rest of the paper we focus on the case where the direction of conflict is common knowledge at the outset.

2.4 Biased observer: the upper-triangular model

In the model of Section 2.2, the parameter $b > 0$ represents the committee's bias in favor of accepting the proposal. In this section we consider the mirror image case where the players would like to accept the proposal if and only if the benefit x exceeds the cost y. But, the observer, or the public he represents, is biased against the proposal, with the parameter $c > 0$ representing the size of the conflict with the players.[18]

Suppose $\mathcal{S} = [0,1]^2$. Let $\mathcal{R} = \mathcal{U} \equiv \{(x,y) \in \mathcal{S} \mid y \geq x\}$, where $\mathcal{U}$ is the upper triangular area on or above the diagonal $y = x$. Let $\mathcal{A} = \{(x,y) \in \mathcal{S} \mid y < x - c\}$. The parameter $c > 0$ represents the conflict between the observer and the committee and $\mathcal{C} = \{(x,y) \in \mathcal{S} \mid x - c \leq y < x\}$ is the zone of conflict. Assume priors are uniform iid so that an instance of the model is fully specified by the constant c. As noted before, the restriction to uniform priors is without loss of generality since we transform types to their quantiles. For the moment we restrict attention to a constant conflict c but relax this assumption later in the section. Since a necessary condition for subversion is non-negative total slack, $\Pr[\mathcal{A}] \geq \Pr[\mathcal{C}]$, and $\Pr[\mathcal{R}] = 1/2$, we must have $\Pr[\mathcal{A}] \geq 1/4$ in order for subversion to be possible. This translates to the condition $c \leq c^* \equiv 1 - 1/\sqrt{2}$, which is also required for a fully informed expert to be able to subvert.

Figure 6(a) depicts the state space for $c = 1/4$ as well as the initial stages of a subversive conversation for this case. In round 1 of this conversation, X either says "$x \in [1/4, 1/2] \cup [3/4, 1]$" (denoted by a message m_1 in panel (a)) or she says "$x \notin [1/4, 1/2] \cup [3/4, 1]$" (denoted by the message m_1'). Regardless of X's message, Y sends one of two similar messages in round 2—either she says "$y \in [0, 1/4] \cup [1/2, 3/4]$" (message m_2) or "$y \notin [0, 1/4] \cup [1/2, 3/4]$" (message m_2'). Panel (b) of Figure 6 depicts the four possible (re-pasted) residual state spaces that are left after each of the two messages send by each player in the first two rounds.

Consider first the residual state space after the messages (m_1, m_2). It is an instance of the lower-triangular model. By our previous results the committee can subvert in this case. Similarly, the residual state space after the pair of messages (m_1', m_2') is also covered by our results for the lower-triangular model. The bottom right quadrant is an instance of the lower-triangular model.

[18] In cheap talk models with a one dimensional state space (e.g., Crawford and Sobel, 1982), only the relative bias of the sender vis-a-vis the receiver matters. This is not true in our setting with a two dimensional state space where the shapes of the acceptance and conflict sets are important.

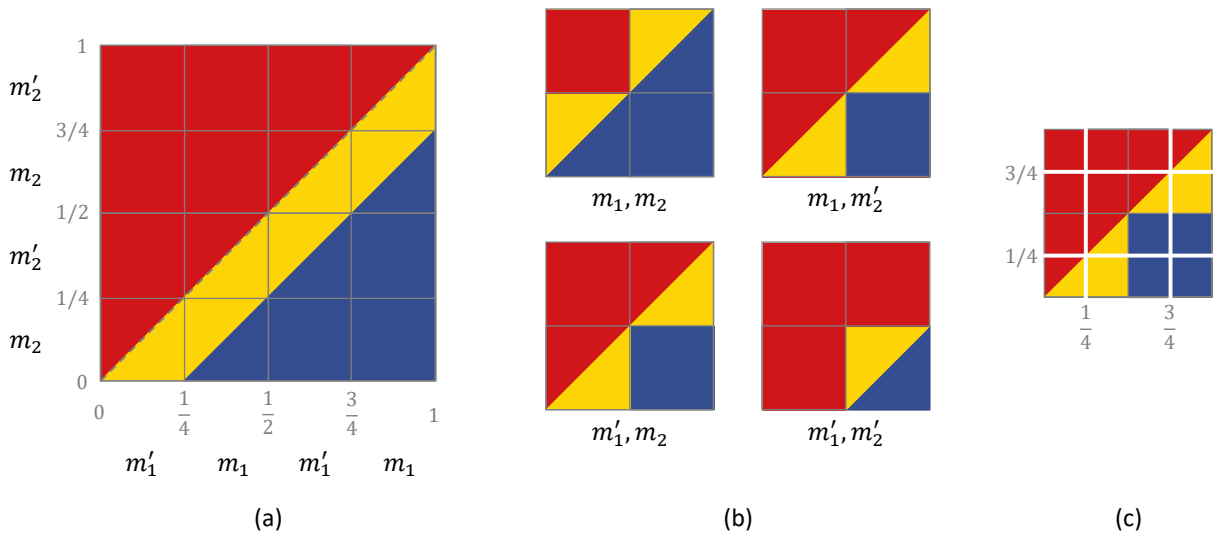

Figure 6: A subversive conversation for $c = 1/4$.

The two players can coordinate to this part of the state space in the first two rounds and then use a subversive conversation for the lower-triangular model (if the state lies in that quadrant), rejecting the proposal otherwise.

The remaining two residual state spaces depicted in panel (b), corresponding to the messages (m_1, m_2') and (m_1', m_2), are identical to each other. Panel (c) depicts how the players can subvert in this case, using a recursive construction similar to the one depicted in Figure 4. Player X either says "$x \in [1/4, 3/4]$" or "$x \notin [1/4, 3/4]$" and Y responds with the similar messages "$y \in [1/4, 3/4]$" or "$y \notin [1/4, 3/4]$". When the resulting state space is either $[1/4, 3/4] \times [1/4, 3/4]$ or $[1/4, 3/4]^c \times [1/4, 3/4]^c$, the problem is identical to the one the players started from (with appropriate pasting and rescaling), and so the players can restart the conversation and continue onwards. Otherwise, when the residual state space is $[1/4, 3/4] \times [1/4, 3/4]^c$, Y can reveal her signal and X can take the optimal decision while meeting deniability. Similarly, when the residual state space is $[1/4, 3/4]^c \times [1/4, 3/4]$, X can reveal her signal and Y can take the decision. Since each recursion takes place in a strictly smaller subset of the starting state space (by a factor of $1/4$), with decisions taken in the remaining part, the conversation ends with probability one in finite time and it is subversive.

Proposition 3 below shows that a fine subversive conversation exists in this model for the case $c = c^*$. Since we transform types to their quantiles, we can use the subset property to generalize the model described above, as we did for the lower-triangular model. To this end, assume (i) arbitrary iid priors for x and y (including the possibility of atoms) that satisfy $\Pr[\mathcal{C} \mid x] < c^*$ for all x, and (ii) $\mathcal{R} \supseteq \mathcal{U}$. Condition (i) amounts to reinterpreting c^* as a uniform upper bound on the probability of conflict. Condition (ii) allows us to call this model the upper-triangular model—the assumptions on

preferences in the original type space, $\mathcal{R} \supseteq \mathcal{U}$ and $\mathcal{C} \subseteq \{(x, y) \in \mathcal{S} \mid x - c^* \leq y < x\}$, are preserved in the quantile space. A game is summarized by the triple $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$ in this space. By the subset property, the fine subversion for the case $c = c^*$ and uniform iid priors is a fine subversion in every game. So it is a belief-independent subversion subject to it being common-knowledge that we are within the ambit of the upper-triangular model.

Proposition 3 *Consider the upper-triangular model. There exists a conversation that is subversive for every game* $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$.

Corollary 3 *A belief-independent subversive conversation exists for the upper-triangular model.*

Just like in the lower-triangular model, the committee with dispersed information does as well as the fully informed expert in the upper-triangular model even under uncertainty about preferences. Public communication is as good as secure private communication. We focus on the lower- and upper-triangular models because they capture natural economic situations. The constructive, geometric approach we employ to establish our existence results has an additional benefit. It shows that each residual state space depicted in our figures is itself a model of subversion when interpreted as the whole game. So subversion is possible in a wider class of situations than captured by these two models. In the next section we provide some invariance properties of subversive conversations that allow us to further extend our existence results.

2.5 Invariance properties

Propositions 2 and 3 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 is a method for extending existence established for one game to other games. For instance, the transformation to quantiles that we employ throughout the paper is an invariance property because regardless of the actual priors the conversation can always be in terms of quantiles. This is an example of a more general *relabeling property* that we formally define in this section and provide other examples below.

Robustness is a second reason why invariance properties are important. For instance, the subset property of fine subversions was used to establish the existence of belief-independent subversions in the lower- and upper-triangular models. This is an invariance property because a fine subversive conversation for some game is also a fine subversive conversation for any other game where the acceptance and rejection sets are larger (supersets) and the conflict set is smaller (a subset). The subset property is a special case of a more general *decision-measurability* property that we also formally define in this section. We consider the relabeling property first and the decision-measurability property next, proceeding in each case from examples to general definitions. The examples we will use are depicted in Figure 7.

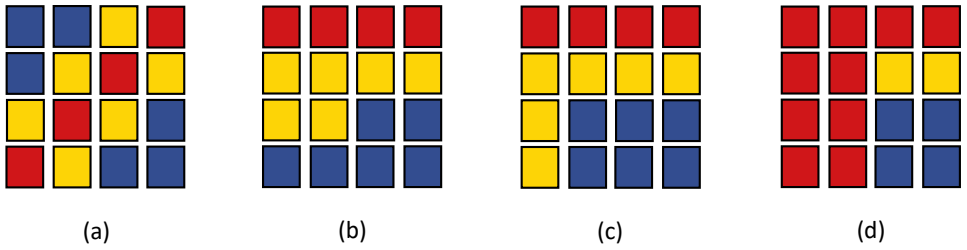

Figure 7: Invarince properties: some examples.

Panels (a) and (b) of Figure 8 depict the two games in panels (a) and (b) of Figure 7. 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 the proposal when the two signals match exactly, and prefer to accept it if the signals are sufficiently dissimilar. For instance, a defendant could be on trial and the proposal up for discussion is whether to convict him or acquit him. When the two signals corroborate each other, the defendant's alibi checks out and everyone prefers to acquit. But the players are more biased in the direction of conviction relative to the observer since they would like to convict unless the signals match up perfectly. We will construct a subversive conversation for the game in panel (a) that, after an appropriate relabeling of types, is also subversive for the game in panel (b). Even though the two games capture different economic situations, they share a subversive conversation.

Start from the game in Figure 8(a). In round 1 of the conversation, X either says the state corresponds to the two left columns as shown in panel (i) or to the two right columns as shown in panel (i$'$). Consider panel (i) first. If we permute two rows corresponding to two different types of Y as depicted in the figure, we obtain the situation depicted in panel (ii). Since a permutation is just a relabeling, it does not change any essential aspect of the continuation game under consideration at this point. The players can converse and take decisions as they would otherwise after taking into account the relabeling. Panel (iii) depicts how the conversation proceeds in the relabeled continuation game. Player Y reveals whether she has the lowest type or not. In the former case, X takes a decision. In the latter case, X reveals her exact type and Y takes a decision. For each of the three continuation games created by these messages, panel (iv) depicts how we can permute rows and columns within each game. Panel (v) pastes together the three continuation games of panel (iv). In panel (vi) we have performed another row permutation in the resulting game, as depicted in the figure. Panels (ii$'$) through (vi$'$) depict an analogous sequence of moves starting from the continuation game depicted in panel (i$'$). If we now paste together the games depicted in panels (vi) and (vi$'$) and perform the column permutation depicted in the figure, we obtain the

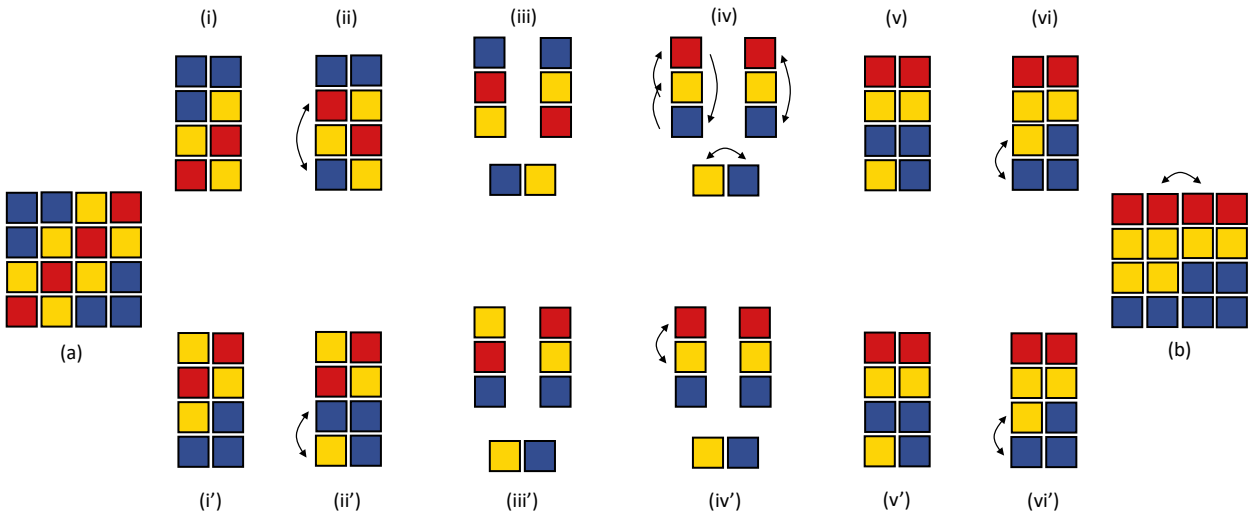

Figure 8: An illustration of the relabelling property.

game depicted in panel (b) of Figure 8 (and of Figure 7).

The same conversation is subversive for these two games because we can go from one game to the other by relabeling the residual type spaces at every history generated by the fixed conversation. In particular, our results on the existence of subversive conversations extend to games where the acceptance, rejection and conflict sets are not monotonic, convex or even connected. Games that differ in these properties may have the same subversive conversation. We now define the relabeling property formally, taking into account that types may lie in a continuum.

Fix a subversive conversation σ for a game $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$. Let $\mathcal{S}_i\left(m^t\right)$ denote the residual type space of player $i = X, Y$, following history m^t, so that $\mathcal{S}\left(m^t\right) = \mathcal{S}_X\left(m^t\right) \times \mathcal{S}_Y\left(m^t\right)$ is the residual state space generated by σ after history m^t. Given m^t, an *admissible relabeling* of $\mathcal{S}_i\left(m^t\right)$ is a measure-preserving bijection $\rho_i(\cdot|m^t) : \mathcal{S}_i\left(m^t\right) \to \mathcal{T}_i\left(m^t\right) \subseteq \mathbb{R}$ whose inverse is also measure-preserving.[19] The set $\mathcal{T}_i\left(m^t\right)$ must have the same cardinality as $\mathcal{S}_i\left(m^t\right)$. If $\mathcal{T}_i\left(m^t\right) = \mathcal{S}_i\left(m^t\right)$ and $\rho_i(\cdot|m^t)$ is the identity map, player i's residual type space is not relabeled after m^t.

If $\mathcal{T}_i\left(m^t\right) = \mathcal{S}_i\left(m^t\right)$ but $\rho_i(\cdot|m^t)$ is not the identity map, the relabeling delivers a *permutation* of $\mathcal{S}_i\left(m^t\right)$. Figure 8 is a discrete type example, with permutations performed at multiple histories generated by the fixed conversation. Another example with a continuum of types can be found in the upper right quadrant of Figure 5(b). This is an "upside-down" instance of the lower-triangular model, i.e., all "rows" have been permuted at the history where this continuation game is reached.

If $\mathcal{T}_i\left(m^t\right) \neq \mathcal{S}_i\left(m^t\right)$, we call $\rho_i(\cdot|m^t)$ 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 the transformation to quantiles performed previously. The rescaling described in the context of Figure 4 and used to establish

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

Proposition 2 is another example, and we will see more in the next section.

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

We turn now to the second invariance property of subversive conversations, decision-measurability. Consider again the game of Figure 7(b). Apart from the fine subversion depicted in Figure 8, it also admits the following coarse subversion. Player Y moves first and she rejects the proposal when her type corresponds to the top row; otherwise she turns the conversation over to X who takes the committee-optimal decision in round 2. Let m_1 denote Y's message that the state is not in the top row and let $\mathcal{S}(m_1, A)$ denote the residual state space over which X accepts the proposal after message m_1. Notice that if we alter $\mathcal{S}(m_1, A)$ arbitrarily while still allowing X to take a committee-optimal decision in round 2, we have the same subversive conversation in the altered game. Only the action plan may change. An example is Figure 7(c) where we have shuffled the acceptance and conflict points in a measure-preserving manner not allowed by the relabeling property (in particular, the shuffling cannot be generated by permutations of rows or columns). Player X can still take the optimal decision of accepting the proposal after this shuffle and meet deniability.

Figure 7(d) provides an example where $\mathcal{S}(m_1, A)$ is altered in a way that is not measure-preserving. In particular, we have increased the rejection set and reduced the acceptance and conflict sets. This perturbation also preserves the property that X can take the committee-optimal decision in round 2. She can reject the proposal when the state is in the left two columns, and accept it otherwise, while meeting deniability. While the action plan may change across the three games depicted in panels (b) through (d), the same conversation is subversive in all of them. As shown above, the game in panel (b) also shares a fine subversive conversation with the game in panel (a) of Figure 7. But this conversation cannot be used in panel (c). So the multiplicity of subversive conversations in any given game is a robustness feature. Each such conversation generates its own set of games where subversion is possible in the same way.

To define the decision-measurability property formally, fix a subversive conversation σ for a game $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$. Let $\mathcal{S}\left(m^t, d_{t+1}\right)$ be the residual state space after a history of messages m^t generated by σ, followed by a decision $d_{t+1} \in \{A, R\}$.

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

The decision-measurability property allows us to modify the residual state space $\mathcal{S}\left(m^t, d_{t+1}\right)$ in a way that permits the player taking the decision to still be able to determine the committee-

optimal decision (condition (ii)) and to meet the deniability constraint (DC) if she takes a decision to accept the proposal (condition (i)). 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 (ii) 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 (i). Indeed, the decision-measurability property applies both to fine and coarse subversive conversations.

These invariance properties show that our existence results of the previous sections are more general than they might appear. Each subversive conversation for a given game provides an implicit characterization of a larger set of games where subversion is possible in the same way. In addition to these two properties, subversive conversations only require specifying the committee's ordinal ranking of decisions in every state. So they are also invariant to the cardinal specification of the committee's payoffs. Our focus on subversion is important in obtaining these properties. They will not obtain for non-subversive equilibrium conversations for which incentive constraints matter.

3 Failures of subversion

In this section we identify conditions under which subversion is impossible. Recall that a necessary condition for the committee to subvert is non-negative total slack, $\Pr[\mathcal{A}] \geq \Pr[\mathcal{C}]$. But this is not sufficient. Since each player has one piece of information, they need to exchange information in order to determine their optimal decision in finite time with probability one. The players must maintain non-negative slack in all continuation games that arise during this process. The three-type example in Figure 9(a) shows a simple case where this is impossible. If X partitions the columns into two or more elements with her first move, at least one element of the partition will have negative total slack. So (DC) cannot be satisfied in the continuation game corresponding to that element. The same is true if Y makes the first move. Since $\mathcal{R}$ is non-empty, some information must be exchanged to determine the optimal decision and so neither player can take a decision at the outset. We conclude that subversion is impossible if one restricts attention to pure strategies, and now extend the argument to cover mixed strategies.

Transform types to quantiles so that x and y are continuous and uniformly distributed on $[0,1]^2$. Mixed strategies in the original finite type game of panel (a) correspond to pure strategies in the transformed games with a continuum of quantile types, depicted in panels (b) and (c). Suppose a subversive protocol exists in this transformed game. Note first that a decision cannot be taken with positive probability in round 1. For 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, we must have negative total slack in the continuation game corresponding to complementary part where X does not take a decision in round

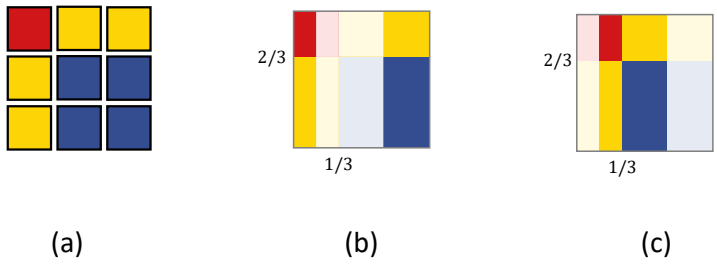

Figure 9: Non-existence of a subversive conversation.

1. Subversion is 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 with her first move, each of which has zero total slack.[20] An example is depicted in panels (b) and (c) of Figure 9. Notice that each of these elements can be rescaled back to the original game using the invariance properties described in Section 2.5. But then a decision cannot be taken with positive probability in round 2 either, for each such element, exactly because it could not be taken in round 1 for the original game. Continuing the argument, it follows that a decision can never be taken with positive probability. Since a subversive protocol must take a decision in finite time with probability one, this yields the desired contradiction.

Recall that non-negative slack is necessary and sufficient for the fully informed expert to subvert. The same is true if the committee could engage in secure private communication. Figure 9 is the simplest example of a situation where public communication with dispersed information imposes a cost on the committee. We can use it to construct larger examples, with strictly positive total slack, where the committee may be able to take decisions in some states and for some time. But ultimately they will run up against a situation where decisions cannot be taken with positive probability while preserving non-negative slack, just like in Figure 9.

Consider next any game with negative total slack, i.e., $\Pr[\mathcal{A}] < \Pr[\mathcal{C}]$. Neither the committee with dispersed information nor the fully informed expert can subvert since a necessary condition for subversion is violated in either case. What can they do? As before, we focus on the payoff-dominant equilibrium for both the committee and the fully informed expert. Since neither the committee nor the fully informed expert can subvert, in principle they may accept the proposal even when the state is in $\mathcal{R}$ and reject it when the state is in $\mathcal{A}$. They need to ensure that the observer has no incentive to object to any decision that is made in equilibrium.

Consider the fully informed expert first. The only decision rule that can be supported in equilibrium involves always rejecting the proposal. For if there is an equilibrium where the proposal

[20] We allow an uncountable number of elements, subject to the admissibility restriction on subversive protocols imposed in Section 2.1.

is accepted with positive probability, then the fully informed expert will accept whenever the state is in $\mathcal{A} \cup \mathcal{C}$. Inferring this, the observer must object to such a decision, since $\mathcal{C}$ is larger in measure than $\mathcal{A}$. So the fully informed expert must earn zero expected payoffs. For the committee, there is always an equilibrium with zero payoffs in which the proposal is always rejected because the observer believes $\mathcal{A}$ is strictly smaller than $\mathcal{C}$ regardless of history, and so will always object to a decision to accept. In the payoff-dominant equilibrium the committee will do weakly better than this babbling equilibrium.

Proposition 4 *In any game with negative total slack, the committee with dispersed information does at least as well as the fully informed expert, and in some cases strictly better.*

In our cheap talk framework, the fully informed expert cannot commit not to use all her information. In contrast, the committee can compromise by credibly ignoring some of its dispersed information. Each player can conceal some information from her partner in an incentive compatible manner, or take decisions that cannot be reversed without objections from the observer, effectively providing the committee with commitment not available to the fully informed expert. This difference in effective commitment underlies Proposition 4. While in any game with non-negative total slack the fully informed expert does at least as well as the committee with dispersed information, and in some cases strictly better, this ordering is reversed under negative total slack. Notice that the observer also weakly prefers information to be dispersed, and strictly so when the committee decision is informative.

To see how the committee can do strictly better than the fully informed expert, consider the example in Figure 10 with iid uniform priors and negative slack so that the fully informed expert earns zero payoffs. Unlike the case of subversion where we only needed to specify the committee's ordinal ranking of actions, we now need to describe their cardinal payoffs and take into account incentive constraints. Suppose the players' common payoff from accepting the proposal is equal to 1 if $(x, y) \in \mathcal{A} \cup \mathcal{C}$, and equal to -1 otherwise. Figure 10 depicts the committee optimal equilibrium conversation. In this conversation, X accepts the proposal in round 1 when her type $x = 2$ and otherwise passes the conversation over to Y. Subsequently in round 2, Y accepts the proposal when $y \in \{0, 1\}$ and otherwise passes the conversation back to X. In round 3, X accepts the proposal when $x = 3$ and rejects it otherwise.

In order to check incentive compatibility, notice that type $x = 1$ is the only type of player X and type $y = 2$ is the only type of player Y who do not get their first best payoffs and so these are the only cases to check. When X takes a decision in round 3, type $x = 1$ rejects the proposal. She does not have an incentive to mimic $x = 3$ and accept the proposal at this stage because Y pools her types $y = 2$ and $y = 3$ and so X infers the state is equally likely to be in $\mathcal{R}$ or in $\mathcal{A} \cup \mathcal{C}$. For the same reason, $x = 1$ also does not have an incentive to pretend to be $x = 2$ and take a unilateral decision in round 1. With respect to Y, since X pools the types $x = 0$ and $x = 1$ throughout the

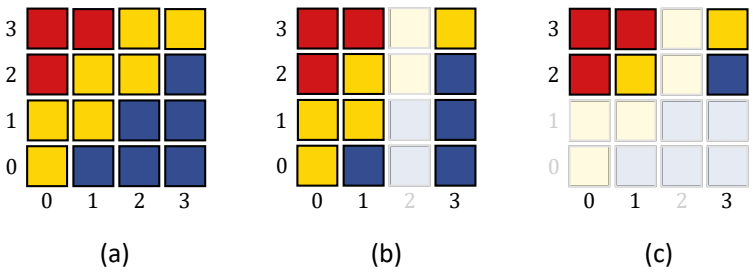

Figure 10: A conversation under negative total slack.

conversation, type $y = 2$ does not have any incentive to mimic types $y < 2$ and accept the proposal unilaterally in round 2.[21]

From the perspective of the committee the only inefficiency occurs in state $(x, y) = (1, 2)$. So they have an incentive to discern if this is the state and accept the proposal in that case. We suppose that if either player tries to send an off-the-path message (or a neologism) at any stage during the conversation, the observer will infer that discerning this state must be the objective of the deviation. So he will object to any subsequent decision to accept the proposal, ruling out such deviations. The last point illustrates why the committee does better than the fully informed expert. The fully informed expert can discern the exact state without sending a public message that the observer will also see. Anticipating this, the observer will never approve of a decision by the fully informed expert to accept the project.[22]

Proposition 4 depends on our assumption of no commitment. If it was common knowledge that the fully informed expert could commit to ignore some of her information, she could simply mimic the committee's conversation and do as well as the players (if not better). Similarly, if the players could commit to their strategies, in general they would do at least as well as without commitment. Given our specification of committee payoffs, the committee earns the optimal commitment payoffs under cheap talk in the example of Figure 10. More generally, commitment can only help the players when subversion is impossible, and sometimes strictly so (e.g., for the example in Figure 9). Whether or not the players have commitment, the key point is that conversations are often necessary for the players to do as well as possible. It is the deniability constraint, not incentive constraints, that makes conversations necessary.

[21] This decision rule is committee-optimal because it implements the committee's first best decision in all but one state $(x, y) = (1, 2) \in \mathcal{A} \cup \mathcal{C}$. This is the minimal amount of compromise necessary to restore non-negative slack in the set of states where the committee accepts the proposal. It is not difficult to show that there is a unique incentive compatible committee-optimal decision rule in this example, and a back and forth conversation is necessary to implement it.

[22] For the same reason, the observer should be skeptical of a decision to accept the proposal when the players engage in secure private communication.

4 Related Literature

This paper belongs to the literature on cheap talk (Crawford and Sobel, 1982; Green and Stokey, 2007) with multi-dimensional information (Battaglini, 2002; Chakraborty and Harbaugh, 2007, 2010) and multiple audiences (Farrell and Gibbons, 1989), some of whom have may private information (Watson, 1996). Our model of communication between players with identical preferences but different information is antipodal to that of Battaglini (2002) who considers multiple experts with different preferences but identical multi-dimensional information and focuses on receiver-optimal rather than sender-optimal outcomes.

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 our requirement is for an outside observer to consent. Since communication within the committee is not scrutinized by any outsider in their setting, back and forth conversations have no role to play.

The literature on long cheap talk (Forges, 1990; Amitai, 1996; Aumann and Hart, 2003; Krishna and Morgan, 2004; Golosov, Skreta, Tsyvinski and Wilson, 2014; Chen, Goltsman, Horner, and Pavlov, 2017) asks how extended cheap talk can alter the set of equilibrium outcomes. Matthews and Postlewaite (1995), in particular, provide examples similar to our model where two privately informed players engage in sequential communication in front of an observer who takes decisions. They show that multi-round sequential communication leads to different outcomes compared to single round or no communication, although the presence of the observer distorts decisions away from what is ideal for the players. We extend this line of research and identify a large class of situations where the problem of communication under scrutiny can be solved costlessly. In a subversive conversation, the players persuade the observer to implement their own ideal decision rule. The general question of identifying all equilibrium outcomes of communication under scrutiny, whether or not subversion is possible, remains open.

Krishna and Morgan (2001) consider communication by two experts with different preferences but the same information. They ask how a process of debate and rebuttal between two experts can result in the receiver's best equilibrium in which the experts perfectly reveal the state.[23] In contrast, the two experts of our paper have different information but the same preferences. We ask if the experts can achieve their optimal outcome, while maintaining deniability, simply by structuring their discussion suitably. Sequential communication in the form of back and forth conversations may be enough to achieve these goals, an insight new to the literature. When a player sends

[23] See also Meyer-ter-Vehn, Smith, and Bognar (2017) who model communication in the form of repeated voting by two players in a debate setting with costly delay.

a message, no additional information is inferred by the other player that cannot be inferred by the observer, so messages are not encrypted (Shannon, 1949). Encryption is unnecessary because, unlike in the cryptography literature which assumes a "malevolent" third party, there is enough commonality of interests between the players and observer in our model.

Since the players attain their ideal outcomes in a subversive conversation, subversion is possible under cheap talk if and only if it is possible under commitment (Kamenica and Gentzkow, 2011; Lipnowski and Ravid, 2020). Even when subversion is possible, the ability to design communication protocols ex ante is likely to be useful for organizations, since this can avoid coordination failures. When subversion is not possible, the ability to commit to protocols will often be valuable for the players. Irrespective of whether the players have commitment or not, conversations are essential for them to do as well as possible in many situations.

Glazer and Rubinstein (2004) consider optimal rules of persuasion from the perspective of a single listener facing a single speaker informed about multiple aspects of a decision problem when the listener can obtain a limited amount of evidence. We focus on sender optimal outcomes, when information about the different aspects is dispersed among multiple senders, and the receiver cannot independently obtain evidence. Matching conflict points with agreement points is a key feature of our subversive conversations. This is reminiscent of strategic argumentation by a single expert communicating with an uninformed receiver (Dziuda, 2011). Since in our model each expert has private information, such pooling (and separation) must be designed not only to persuade the receiver but also to share enough information publicly and coordinate on the experts' ideal decision. Hall's Marriage Theorem (1935) provides a necessary condition for a fine subversion in our setting, describing the different ways conflict and agreement points can be matched by the players. But Hall's theorem does not yield a sufficient condition for fine subversions (nor a necessary condition for coarse ones), since information is dispersed among multiple players.

A subversive conversation ensures that the observer agrees with the players, so decision making power is effectively held by the players, even when legal authority lies with the observer. This echoes the distinction between formal and real authority drawn by Aghion and Tirole (1997). In our setting, no player has all the information necessary for the decision and they have to aggregate dispersed information in public. Nevertheless, the ability to manage the process of communication may give them effective authority. Because information is dispersed among multiple players, the circumstances under which delegating formal authority is optimal for the organization (Dessein, 2002; Alonso, Dessein, and Matouschek, 2008) remains an open question.

Within the large political economy literature, this paper is most directly related to studies of committee deliberations. Gradwohl and Feddersen (2018) (see also Wolinsky, 2002; Feddersen and Gradwohl, 2020) study the effect of transparency in a cheap talk model with multiple senders who have common interests and correlated binary signals. 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 the committee and the fully informed expert is related to this transparency versus opacity distinction. In our different environment, we identify conditions when the committee can subvert even under transparency, as well as situations where transparency can help communication because of its effect on incentives.

5 Conclusion

This paper analyzes a common situation in information transmission between players with similar interests—their communications may be overheard by players 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 make 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 number of interesting open questions emerge from this paper. As mentioned earlier, the general problem of communication under scrutiny remains open, although our results extend in a number of directions beyond the model considered in this paper. The deniability constraint can be modified to cover the case where the observer cares about the value of the decision, e.g., when his payoff from accepting the proposal equals $x - y$, with that from rejecting it normalized to zero. It is easy to see that the same conversation depicted in Figure 3 for the baseline model remains subversive if the observer's payoffs take this form. But the general problem remains open. Even when the observer cares about magnitudes, meeting the deniability constraint may preclude more detailed inquiries when investigations are costly. Our results are also robust to allowing a less demanding evidentiary standard for the observer, although a full treatment of such cases also remains open.

We have assumed the committee faces a binary decision problem and that there is no conflict of interest within the committee. Subversion is impossible in continuous action problems whenever the observer can perfectly infer the state from the observed messages and decisions. But subversion does not require binary decisions. Our results extend to a number of discrete choice problems such as a committee choosing between hiring one of multiple candidates, including the possibility of hiring no one. Conflicts of interest within the committee create incentive constraints even when there is no observer. A natural extension is to ask whether an incentive-efficient decision rule of the communication game without scrutiny can be implemented in the game with scrutiny, via an

outcome-equivalent conversation. Our results are robust in this direction but a full treatment may yield new insights, in particular because incentive-efficient decision rules may not be unique.

Finally, we have assumed that the players have complete freedom over the design of communication. The mirror-image case is where the observer is the designer. While unrestricted mechanism design by the observer is unlikely to yield results that are new to the literature, more restricted forms may be interesting. Even when the observer is the designer, our results shed light on the case where the players retain the freedom to engage in unrestricted and public side-communication. What kind of restrictions on this freedom would the public like to impose remains an open question. Comparing the two problems, one where the players are free to design the rules of communication and the other where they are not, may yield novel insight on the nature of authority within an organization when information is dispersed, as well as new understanding of the scope of regulatory oversight.

6 Appendix A: Proofs

Proof of Proposition 1.

We prove part (i) here via breaking up the proof into two cases, $b > 1/3$ and $b \leq 1/3$. Assume first $b > 1/3$, and consider the following protocol:

- In round 1, X either says "$x \in [(1-b)/2, (1+b)/2]$" or she says "$x \notin [(1-b)/2, (1+b)/2]$".
- If $x \in [(1-b)/2, (1+b)/2]$, then in round 2:
  - Y accepts the proposal for all $y \in [(1-b)/2, (1+b)/2]$. Doing so is optimal for the committee since we must have $y \leq (1+b)/2 = (1-b)/2 + b \leq x + b$ and so $(x, y) \in \mathcal{R}^c$ whenever $(x, y) \in [(1-b)/2, (1+b)/2]^2$. This will satisfy (DC) because, conditional on this history that reveals only $(x, y) \in [(1-b)/2, (1+b)/2]^2$, the residual acceptance set, $\{(x, y) \in [(1-b)/2, (1+b)/2]^2 \mid y \leq x\}$, is equal in measure to the residual conflict set, $\{(x, y) \in [(1-b)/2, (1+b)/2]^2 \mid x < y \leq x + b\}$.
  - Y passes the conversation back to X when $y \notin [(1-b)/2, (1+b)/2]$. Subsequently, X perfectly reveals x and Y then takes the committee-optimal decision. Whenever Y accepts, the residual acceptance set equals $\{x\} \times [0, (1-b)/2)$ while the residual conflict set is a subset of $\{x\} \times ((1+b)/2, 1]$. Both these sets have zero measure in $\mathbb{R}^2$. Throughout the paper, in all such cases, we will use the joint density to compute the induced measure on $\mathbb{R}^1$ and associated conditional probabilities. In this instance that is the uniform distribution on the relevant subset of $\mathbb{R}^1$. It is easy to see see that a decision to accept will meet (DC), since the residual acceptance set has measure $(1-b)/2$ (in $\mathbb{R}^1$) and the residual conflict set is a subset of a set of measure $(1-b)/2$ (in $\mathbb{R}^1$).

- If $x \notin [(1-b)/2, (1+b)/2]$, then in round 2:

  - Y perfectly reveals y when $y \in [(1-b)/2, (1+b)/2]$. Subsequently X takes the committee-optimal decision. A decision to accept will meet deniability since conditional on this history, the residual acceptance set equals $((1-b)/2, 1] \times \{y\}$ which is at least as large (in the induced measure on $\mathbb{R}^1$) as the residual conflict set which is a subset of $[0, (1-b)/2) \times \{y\}$.
  - Otherwise, Y either says "$y > (1+b)/2$" or "$y < (1-b)/2$".
    - When $y > (1+b)/2$, X rejects the proposal if $x < (1-b)/2$ because she deduces $(x, y) \in \mathcal{R}$ from the fact that $x + b < (1-b)/2 + b = (1+b)/2 < y$. On the other hand, X accepts the proposal when $x > (1+b)/2$ since she infers $(x, y) \in \mathcal{R}^c$ from the fact that $x + b > 1 \geq y$ using $b \geq 1/3$. A decision to accept will satisfy deniability since the residual acceptance set, $\{(x, y) \in \{((1+b)/2, 1] \times ((1+b)/2, 1] \mid y \leq x\}$, has a conditional probability equal to that of the residual conflict set $\{((1+b)/2, 1] \times ((1+b)/2, 1] \mid x < y\}$, given this history.
    - When $y < (1-b)/2$, X accepts the proposal since she deduces $(x, y) \in \mathcal{R}^c$ from the fact that $x + b \geq b \geq (1-b)/2 > y$, using $b \geq 1/3$. This will satisfy (DC) since the residual conflict set, $\{[0, (1-b)/2) \times [0, (1-b)/2) \mid y > x\}$ is strictly smaller in measure than its complement, the residual acceptance set given the history.

This completes the construction for the case $b > 1/3$. For $b \leq 1/3$, we use the following conversation:

- In round 1, X perfectly reveals x when $x \in [b, 1-b]$ and Y then takes the committee optimal decision. Whenever Y accepts, the residual acceptance set equals $\{x\} \times [0, x]$ while the residual conflict set is a subset of $\{x\} \times (x, x+b]$ and so (DC) is met since $b \leq x$.

- When $x \notin [b, 1-b]$, X simply reveals this fact in round 1. Subsequently, Y perfectly reveals y when $y \in [b, 1-b]$ and X then takes the committee-optimal decision. Otherwise Y either says "$y > 1-b$" following which X rejects if $x < b$ and X accepts if $x \geq 1-b$, or Y says "$y < b$" following which X accepts the proposal in round 3. In all cases, (DC) holds whenever the proposal is accepted, for reasons identical to that described above for the case $b > 1/3$, and so we omit the details.

The fact that a subversive conversation will take three rounds to complete for all $b \in (0, 1)$ follows from the discussion in the text and here we prove the remaining part of the claim in (ii). Fix $b > 1/3$ and assume by way of contradiction that there exists a subversive protocol (σ, α) such that decisions take place by round 3 with probability 1. Let $\mathcal{A}(m^t)$, $\mathcal{C}(m^t)$ and $\mathcal{R}(m^t)$ be, respectively, the residual

acceptance, conflict and rejection sets after history m^t and let $\mathcal{S}(m^t) = \mathcal{A}(m^t) \cup \mathcal{C}(m^t) \cup \mathcal{R}(m^t)$ be the residual state space.

Consider first the case $b \geq 1/2$ and $(x, y) \in (0, 1-b) \times (b, 1)$. Observe that X cannot perfectly reveal any $x \in (0, 1-b)$ in round 1 and satisfy (DC) and so for X to take the players' ideal decision in round 3, it must be that Y perfectly reveals each $y \in (b, 1)$ in round 2 in order for player X to be able to take the ideal decision in round 3. Let $H = \{\sigma_X(x, y; m_0), m_2 \mid x \in (0, 1-b)\} \subset H_A^2(\sigma, \alpha)$ be the set of all two round message histories that are generated when $x \in (0, 1-b)$ and a decision is taken in round 3. Our admissibility restriction implies that (DC) must be satisfied when we integrate over histories in H.

Suppose Y reveals $y = b + \varepsilon$ in round 2, for $\varepsilon > 0$ and small. Since all $x \in (0, 1-b)$ generate a history $m^2 \in H$, $\mathcal{C}(m^2)$ must have measure at least $1 - b - \varepsilon$ at such y. Since (DC) is met if X then accepts, $\mathcal{A}(m^2)$ must have at least the same measure. Since the measure of $\mathcal{R}(m^2)$ is ε, the measure of $\mathcal{S}(m^2)$ must then be at least $2(1 - b - \varepsilon) + \varepsilon$. But the measure of $\mathcal{S}(m^2)$ after X's first message cannot depend on y since X has no information about y, the measure of $\mathcal{S}(m^2)$ must be at least $2(1 - b - \varepsilon) + \varepsilon$ for all $y \in (b, 1)$.

Consider now the case where Y reveals $y = 1 - \varepsilon$ in round 2. For this history, the $\mathcal{A}(m^2)$ has measure at most ε while $\mathcal{R}(m^2)$ has measure at most $1 - b - \varepsilon$. Since the measure of $\mathcal{S}(m^2)$ is at least $2(1 - b - \varepsilon) + \varepsilon$ for all $y \in (b, 1)$, this means that for $m^2 \in H$ corresponding to $y = 1 - \varepsilon$, $\mathcal{C}(m^2)$ has measure at least $1 - b - \varepsilon$. But then (DC) cannot be met for ε small enough. The proof for $b \in (1/3, 1/2)$ follows along similar lines and we omit the details.■

Proof of Proposition 2. We prove the result for a generalization of the lower-triangular model. Suppose $\mathcal{A} \supseteq \widetilde{\mathcal{L}}$ for some set $\widetilde{\mathcal{L}}$ that satisfies:

- Monotonicity: if $(x, y) \in \widetilde{\mathcal{L}}$, then $(x', y') \in \widetilde{\mathcal{L}}$ whenever $x' \geq x$ and $y' \leq y$.
- Symmetry: if $(x, y) \notin \widetilde{\mathcal{L}}$, then $(1 - x, 1 - y) \in \widetilde{\mathcal{L}}$.

The set $\mathcal{L}$ defined in Section 2.3 is an example of such a set $\widetilde{\mathcal{L}}$.

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.

The conversation is the same for each game $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$. Fix $z \in (0, 1/2)$. The conversation proceeds as follows in the first two rounds:

- In round 1, X either says m_1: "$x \in [z, 1-z]$" or she says m_1': "$x \notin [z, 1-z]$".
- In round 2, Y either says m_2:"$y \in [z, 1-z]$" or she says m_2': "$y \notin [z, 1-z]$".
- After history m_1, m_2': the residual state space is $[z, 1-z] \times [z, 1-z]^c$; player X perfectly reveals x in round 3 provided the ordered pair $(z, z) \in \widetilde{\mathcal{L}}$.

- 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, player Y perfectly reveals y in round 4 provided the ordered pair $(1-z, 1-z) \in \widetilde{\mathcal{L}}$.

- Otherwise, the residual state space inherits both the monotonicity and symmetry property of the acceptance set. In each case the conversation restarts in a rescaled state space as described below.

  - 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 2.5. We then have an instance of the generalized lower-triangular model and so the conversation can proceed as described above.
  - After history m_1', m_2': the residual state space is $[z, 1-z]^c \times [z, 1-z]^c$ and using a similar bijection as in the previous case, we generate another instance of the generalized lower-triangular model in which the conversation proceeds as described above.
  - The same argument also applies after history m_1, m_2' when $(z, z) \notin \widetilde{\mathcal{L}}$, and after history m_1', m_2 when $(1-z, 1-z) \notin \widetilde{\mathcal{L}}$, and the conversation is restarted in these cases as well.

By symmetry, either $(z, z) \in \widetilde{\mathcal{L}}$ or $(1-z, 1-z) \in \widetilde{\mathcal{L}}$, so a decision will be taken with probability at least $2z(1-2z) =: q > 0$ in each recursion of the process described above. So, a decision will be taken in finite time with probability 1.

It remains to show that a decision to accept will meet (DC). Suppose that $(z, z) \in \widetilde{\mathcal{L}}$ and consider the history m_1, m_2', after which player X perfectly reveals x. If Y accepts the proposal, by the monotonicity property all (x, y) with $x \geq z$ and $y \leq z$ belong to $\widetilde{\mathcal{L}} \subseteq \mathcal{A}$. So, conditional on m_1, m_2', x, the residual part of $\mathcal{A}$ is $\{x\} \times [0, 1-z)$ 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 $\mathbb{R}^1$). Similarly, after history m_1', m_2 when the ordered pair $(1-z, 1-z) \in \widetilde{\mathcal{L}}$, identical arguments establish (DC) will be met after an acceptance.

This establishes the result. Since the conversation is fully specified by the properties of $\widetilde{\mathcal{L}}$ and the choice of z, it does not depend on the particular game $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$. Only the action plan (i.e., the particular decision) depends on the specific game $\{\mathcal{A}, \mathcal{C}, \mathcal{R}\}$.■

Proof of Proposition 3.

We will provide a fine subversion for the case $c = c^*$ and there is zero total slack. The construction is outlined in Figure 11(a). In round 1, player X breaks down into zones 1a, 1b and 1c, as labeled in the figure, which together exhaust the whole space.

We describe below what happens in each of the three cases, up to the point where one of the players reveals her type perfectly. Following this the other player takes a decision. So the action plan α is defined as follows: player i takes a non-null decision, $\alpha_i \neq N$, if and only if the other player

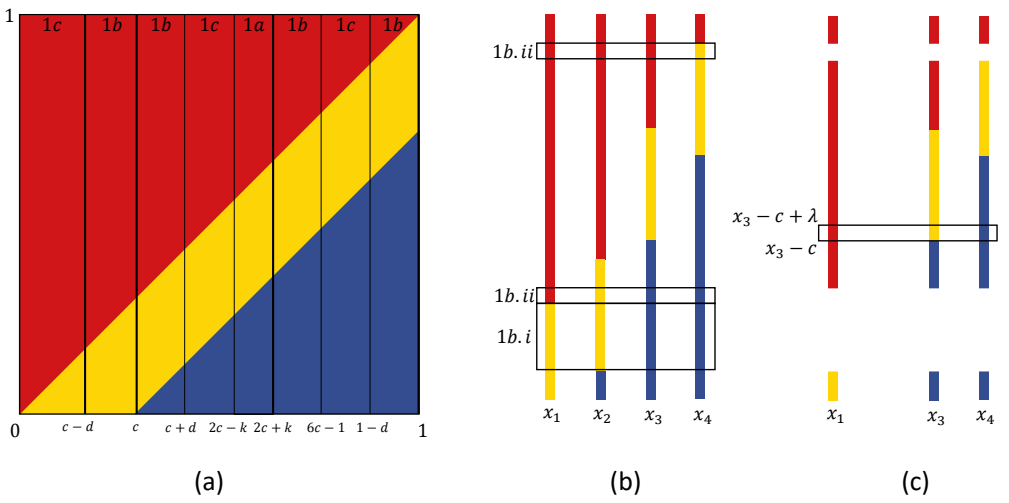

Figure 11: Subversion when $c = c^*$: cases 1a and 1b.

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. The cutoffs and the sequence of messages that we describe below have been chosen so that (DC) will always hold, as can also be verified using the accompanying figures.

Case 1a. In round 1, if $x \in (2c - k, 2c + k)$, player X pools two states $\{2c - z, 2c + z\}$ for $z \in [0, k)$, where $k = 7c - 2 \geq 0$. This region is labeled 1a in Figure 11(a). Following this Y reports the exact value of y if $y \in [c - z, c]$, otherwise she passes and X then perfectly reveals the value of $x \in \{2c - z, 2c + z\}$. See lemma 1 at the end of the proof for details.

Case 1b. Consider $x \in [c - d, c) \cup [c, c + d) \cup [6c - 1 - d, 6c - 1) \cup (1 - d, 1]$, where $d = 1 - 3c$ is the width of each interval. In this case, X pools four types $\{x_1(z), x_2(z), x_3(z), x_4(z)\} := \{c - d + z, c + z, 6c - 1 - d + z, 1 - z\}$ for each $z \in [0, d)$ in round 1. In Figure 11(a) these four types are denoted by thick black lines when $z = 0$, and as z increases the first three types in the pool move right and while the fourth moves left.

Following this, Y will either (i) perfectly reveal y if $y \in [z, c - d + z]$, or (ii) pool all types in $(c - d + z, c) \cup (1 - d, 1 - z)$, or (iii) pass. Cases (i) and (ii) are labeled in Figure 11(b), with case (iii) being the complementary set. In case (ii), X will reveal whether or not $x = x_4(z)$. If $x \neq x_4(z)$, Y perfectly reveals y. Recall that the conversation ends whenever a player perfectly reveals her type, with the other player taking the ideal decision in the next round.

In case (iii) X reveals whether or not $x = x_2(z)$ in round 3. Panel (c) of Figure 11 depicts what happens if $x \neq x_2(z)$. Player Y then perfectly reveals $y \in (x_3(z) - c, x_3(z) - c + \lambda)$, where $\lambda = 2 - 6c - 2z$ (see the rectangle in panel (c)), or she passes. If she passes, X reveals whether or not $x = x_3(z)$. If $x \neq x_3(z)$, Y perfect reveals $y \in [0, z)$, or she passes following which X perfectly reveals $x \in \{x_1(z), x_4(z)\}$.

Case 1c. In this case the conversation has a recursive structure. The three remaining rectan-

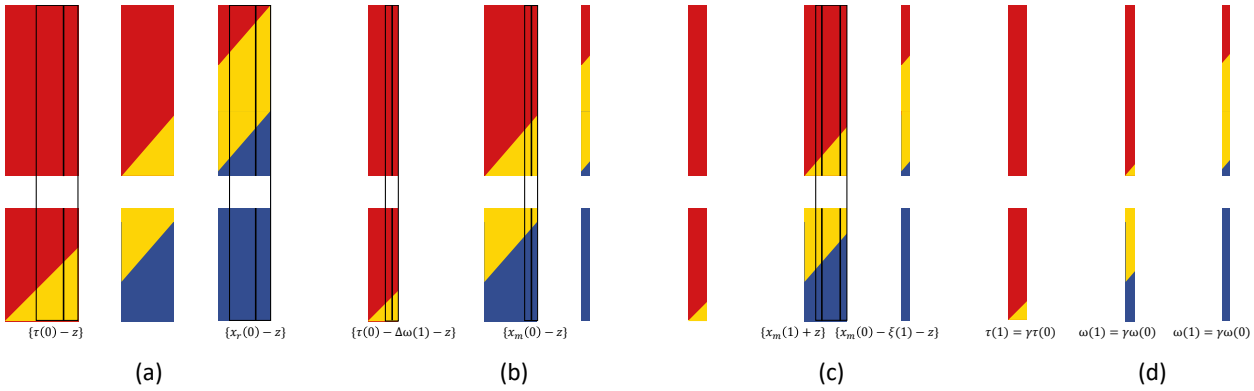

Figure 12: Subversion when $c = c^*$: case 1c.

gular regions of the state space are labeled with $1c$ at the top of Figure 11(a). We are left with a triangular conflict set on the left-most piece of the residual state space and the two other pieces have the same width. These are shown in Figure 12(a). The recursive structure will be such that these properties of the three pieces will be maintained, with each of their widths is scaled by a factor of $\gamma := 3 - 2\sqrt{2}$ in each step of the recursion. In each step of the recursion, decisions are taken with probability at least $1 - \gamma > 0$. One step of the recursion is two rounds of communication during which each player will reveal some information.

Let $\tau(n)$ be the width and height of the triangular conflict area of the left-most piece and let $\omega(n)$ be the common width of the other two pieces in step $n \geq 0$ of the recursion. Let the x-coordinate of the right edge of the right-most piece be $x_r(n)$ and x-coordinate of the right edge of the middle piece be $x_m(n)$. At the beginning of this process we have: $\tau(0) = 4c - 1$, $x_r(0) = 3c$, $x_m(0) = 2 - 5c$, $\omega(0) = 1 - 3c$. For $n \geq 1$, define:

$$\begin{aligned}
\tau(n) &= \tau(n-1)\gamma, \quad \Delta\tau(n) = \tau(n) - \tau(n-1), \\
\omega(n) &= \omega(n-1)\gamma, \quad \Delta\omega(n) = \omega(n) - \omega(n-1), \\
\xi(n) &= \Delta\tau(n) - \Delta\omega(n), \\
x_r(n) &= x_r(n-1) - \Delta\omega(n), \\
x_m(n) &= x_m(n-1) - \Delta\omega(n).
\end{aligned}$$

Step $n \geq 1$ of the recursion consists of two rounds $2n$ and $2n+1$. Player Y perfectly reveals the value of y in round $2n$ if y is in the interval $[x_m(n-1) - c, x_m(n-1) - c + d\gamma^n)$ or $(x_r(n-1) + c, 1]$, in which case the conversation ends and a decision is taken. Figure 12(a) highlights the states Y reveals in round 2, while the remaining states are shown as slightly faded. A decision to accept will meet (DC) because the two right pieces have a common width $\omega(n-1)$. In round $2n+1$, X pools types

(i) $\{\tau(n-1)-z, x_r(n-1)-z\}$ for $z \in (0, \Delta\omega(n))$;

(ii) $\{\tau(n-1)-\Delta\omega(n)-z, x_m(n-1)-z\}$ for $z \in [0, \xi(n))$;

(iii) $\{x_m(n)+z, x_m(n-1)-\xi(n)-z\}$ for $z \in [0, \frac{1}{2}(x_m(n-1)-\xi(n)-x_m(n))]$.

Case (i) reduces the width of the right-most piece of the residual state to $\omega(n)$, by eliminating states from the right and pooling them with some states from the left-most piece, as depicted in Figure 12(b) for the case of $n=1$. Case (ii) reduces the width of the left-most piece to its final width at this step of the recursion $\tau(n)$. Panel (c) shows X revealing the states in case (ii). Finally, case (iii) reduces the middle piece to width $\omega(n)$, as shown in panel (d) of the figure. This completes step n of the recursion. Figure 12(e) shows the remaining states after step 1 of the recursion. In all three cases, the proof of will be completed by using lemma 1 that we present below.■

Consider a model where one player has only two possible types, while the other player's type space is arbitrary. For instance, let $x \in \{l, r\}$, $l < r$, and let $\mathcal{S}_Y \subseteq \mathbb{R}$. Assume that $\mathcal{A}$ and $\mathcal{R}$ are both 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 1 *Let* $\Pr[\mathcal{C}] \leq \Pr[\mathcal{A}]$ *and assume monotonic* $\mathcal{A}$ *and* $\mathcal{R}$*. 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.*

($\Leftarrow$) Assume the right stick has negative total slack, $\Pr[(\{r\} \times \mathcal{S}_Y) \cap \mathcal{C}] > \Pr[(\{r\} \times \mathcal{S}_Y) \cap \mathcal{A}]$, but a subversive conversation 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 $Z \subset (\{r\} \times \mathcal{S}_Y) \cap \mathcal{C}$. But by monotonicity of $\mathcal{A}$ the acceptance region 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.

($\Rightarrow$) Assume the right stick has non-negative total slack. If the left stick also has non-negative total slack, then X can just reveal $x \in \{l, r\}$ 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 region just above the acceptance region 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 and meet (DC).■

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

7 Appendix B. Additional results

Quick conversations. Proposition 2 establishes the existence of a conversation that is subversive for every instance of the (generalized) lower-triangular model, using a recursive construction that utilizes a symmetry property of $\mathcal{L}$. When this property is satisfied, it can be used to show that X can partition $\mathcal{S}_X = [0,1]$ into elements $\{a, 1-a\}$, $a \in [0, 1/2]$, so that each element satisfies the conditions of lemma 1. After this partition in round 1 by player X, the conversation then proceeds as in the lemma, so decisions are taken by round four with probability one.

We present the recursive construction in the main text because it provides more insight into the role of gradualism for subversive conversations. We also provide a recursive construction in the proof of Proposition 3 to establish the existence of a belief-independent subversion for the upper-triangular model. This construction can be amended to ensure decisions will be taken within eight rounds in all cases.

Correlated types. Throughout the paper we assume x and y are statistically independent. This is the case of interest because it implies neither player has any advantage over the observer when it comes to decoding a message sent by the other player. Nevertheless, in some situations the two signals may be correlated and we briefly consider such cases here.

Suppose x and y admit a strictly positive joint density $g(x,y)$ and denote by $g(x|y)$ and $g(y|x)$ the conditional densities derived from this joint density. Consider another joint density $h(x,y)$ that satisfies (a) $\frac{g(y|x)}{h(y|x)}$ is non-decreasing in y, for all x, and (b) $\frac{h(x|y)}{g(x|y)}$ is non-decreasing in x, for all y. We show that if there exists a fine conversation that is subversive for every instance of the lower-triangular model when priors are given by g, then the same is true when priors are given by h. The argument is as follows.

Since the conversation is fine, decision-making is fully informed, and so either the value of x or the value of y must be publicly revealed for every history m^t that precedes a decision. Suppose x is publicly known. Assume, for the moment, $\mathcal{A} = \mathcal{L}$ and $\mathcal{C} = \mathcal{L}^c$. Then the residual acceptance set given m^t (a subset of $\mathcal{S}_Y$) must be an union of intervals that is contained in $[0,x]$ while the residual conflict set is an union of intervals that is contained in $(x,1]$. Since (DC) is met under priors g, by the monotone likelihood ratio property (a), it must also be met for the priors h. The same is true when instead y is publicly known before a decision, using property (b). Using the subset property of fine subversions, the same conversation must be subversive for every other instance of the lower-triangular model.

The identical result obtains also for the upper-triangular model, using the same argument. Relative to g, the density h makes higher values of the benefit x, and lower values of the cost y, more likely, in the sense of likelihood ratios. We leave a fuller treatment of non-iid priors for future research.

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] Amitai, Mor. 1996. "Cheap-Talk with Incomplete Information on Both Sides," working paper.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[21] Hall, P., 1935. "On Representatives of Subsets," *Journal of the London Mathematical Society*, 10: 26–30.

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

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

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

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

[26] 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.

[27] Meyer-ter-Vehn, Moritz, Lones Smith, and Katalin Bognar. 2017. "A Conversational War of Attrition," *Review of Economic Studies*, 85 (3): 1897–1935.

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

[29] Shannon, Claude. 1949. "Communication Theory of Secrecy Systems," *Bell System Technical Journal*, 28(4): 656–715.

[30] 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.

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

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