Subversive Conversations[*]

Nemanja Antic[†] Archishman Chakraborty[‡] Rick Harbaugh[§]

May 9, 2021

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 dispersed information, transparency, deniability, subversion, cheap talk, conversations.

[*] We thank conference participants at the Australasian Public Choice Conference at QUT (2019), Junior Theory Workshop at U. Bonn (2019), Decentralization Conference at U. Michigan (2019), Stonybrook International Game Theory Conference (2019), UChicago-Northwestern Theory Conference (2019), and the Midwest Theory Conference (2019), and seminar participants at the Delhi School of Economics (2019), Monash University (2019), Northwestern University (2019), Queen Mary College (2019), Toulouse School of Economics (2019), University of Bath (2019), UCLA (2020), Norwegian Business School (2020), and University of Arizona (2021). For helpful comments, we also thank David Austen-Smith, Sandeep Baliga, Gabriel Carroll, Eddie Dekel, Wioletta Dziuda, Georgy Egorov, Jeff Ely, Tim Feddersen, Daniel Garrett, Parikshit Ghosh, Faruk Gul, Jason Hartline, Aaron Kolb, 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 communication may be observed by other parties who have different interests. Competitors discussing a merger must account for antitrust regulators who can subpoena their communications. Scientists evaluating a public health crisis need to be careful lest a skeptical public ignores their policy recommendations. Parents conversing in front of a child should consider the effect of the conversation on the child.

We study this problem of communication under scrutiny. Two players, each with a private signal, must share their dispersed information and decide whether to accept a proposal or not. They must maintain plausible deniability that they have violated a norm or broken a law, avoiding controversy, protests, interventions or penalties that may be imposed on them by regulators, supervisors, or other observers who scrutinize their communication.

Deniability is important when senators try to push through a nominee with shared ideology, or activists organize under state surveillance. If communication could be kept truly private, the players could immediately share all of their information without concern for the consequences. But in reality the chance of exposure always remains. Emails can be hacked, codes can be broken, whistleblowers can go public, and so there is need for caution.[1]

We focus on the possibility of subversion—the players subvert when they do as well as they would if they had complete freedom to exchange information and take decisions while maintaining deniability. Communicating under scrutiny imposes no cost on them. The players have common interests, so when they subvert they take their full information, first-best decision. A fully informed expert with the same interests cannot do better.

For subversion to be possible, the process of communication is important. We show how back-and-forth conversations can be used by two players to take their ideal decision in every state, while maintaining deniability that they were acting in the wider interest. As the conversation progresses, the players share more information based on the context established by previous statements, while withholding information that is best revealed later in the conversation. Over several rounds, this gradual process shares enough information for the players to determine their own preferred action, while also hiding enough information to prevent any objections to the players' determination.[2]

[1] Recent prominent exposures include 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.

[2] Our model of communication between players with identical preferences but different information is different from that of Battaglini (2002) who considers the antipodal problem of multiple experts with different preferences but identical multi-dimensional information and focuses on receiver-optimal rather than sender-optimal outcomes. The related literature is discussed in Section 4.

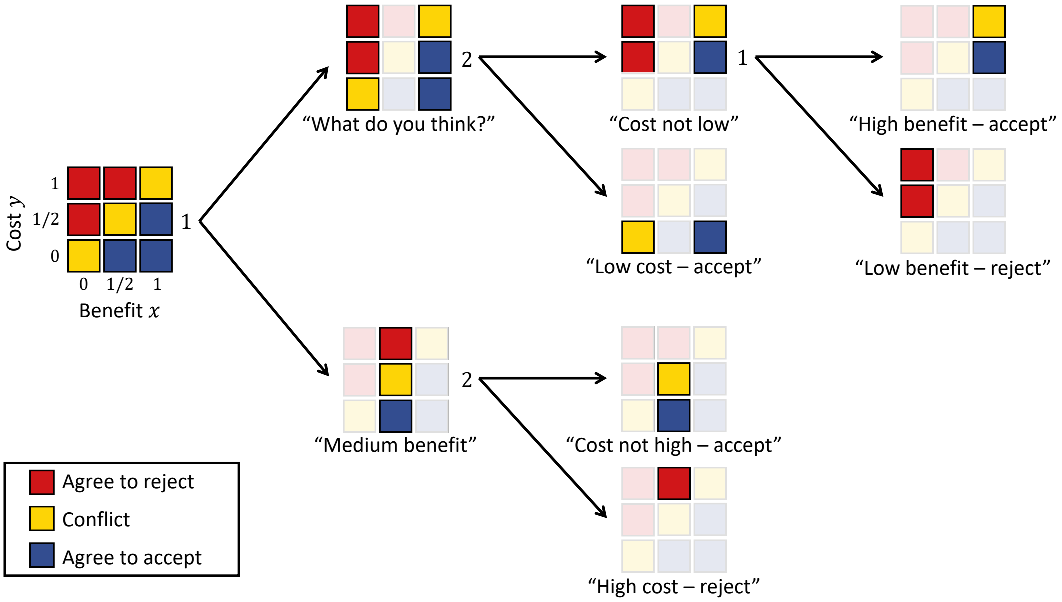

Figure 1: A subversive conversation

To see how a conversation can be subversive, suppose two managers are 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 does the firm. Manager 1 privately knows the project's (economic) benefit $x \in \{0, {}^1\!/\!_2, 1\}$ while manager 2 privately knows the (environmental) cost $y \in \{0, {}^1\!/\!_2, 1\}$. 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 has uniform i.i.d. priors on x and y and only favors a project that has at least an even chance of being good. 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 1 speaks. If the benefit is ${}^1\!/\!_2$ then manager 1 just says so as seen in the first lower branch of the tree in Figure 1. In the second round manager 2 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 ${}^1\!/\!_2$, in which case everyone knows the project is not bad, but only manager 2 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 manager 1 passes the conversation over to the other manager in order to learn whether she has favorable or unfavorable news. In the second round manager 2 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 manager 1 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 $1/2$ or 1 then manager 2 pools this news and tries to gauge what manager 1 now thinks. Having learned that the cost is not 0, in the third round manager 1 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 2 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 contemporaneous monitoring or ex post investigation of their exchanges will not be able to determine that the managers knowingly acted against the public interest.

In this paper, we examine problems like in Figure 1, but with richer type spaces and greater demands on information sharing. We show that subversive conversations exist in many environments, for a variety of preferences and priors, and they need not depend on the details of these environments. The possibility of subversion has implications for organizations when interests 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 transparency by itself may not eliminate bias in hiring.

Instead of different levels of the organization having different preferences, they may have aligned interests but a conflict with the public or an outside authority. Garicano and Rayo (2016) note that concerns about accountability to outsiders may lead executives to forego gathering information from subordinates, resulting in inefficiencies for the organization. We contribute to this research program by showing that even if dispersed information is exchanged in full view, suitable communication protocols can yield plausible deniability for leadership without any efficiency costs for the organization. In the introductory example, for instance, even if a supervisor observes the managers then any ex post investigation will not find she had sufficient grounds to intervene. Organizations have an incentive to design systems and protocols that facilitate subversion. Guiding questions by a supervisor with aligned interests can prevent coordination failures, misunderstandings or untimely revelations, and enforce the right outcomes. Organizational accountability and regulatory compliance may be more difficult to ensure for the public than might otherwise be expected.

Our results also offer insight into the more general role of experts in society. Divergent interests and beliefs between experts and the broader public, and the ability of experts to manipulate public policy, have become a central concern for many issues, from free trade to Brexit to climate change to public health. Even if sunshine laws and other transparency regulations force deliberations out from behind closed doors, we show that the ability of experts to exchange information with each other need not be affected. Experts may be able to push the public in their preferred direction, but the public still benefits from fact-based decision making using expert knowledge. Ignoring experts is worse for the public.

We identify conditions under which subversion is impossible and dispersed information is harmful for the players. For subversion to be possible there must be enough good states to pool with mediocre states, while also separating them from bad states. Since each player can only speak about her own information, there must also be enough diversity in the possible ways these states can be pooled. When the underlying preferences limit the ways players can pool and separate, subversion is impossible and dispersed information is harmful for the players. We link this insight to a well known result in graph theory, Hall's Marriage Theorem (1935).

In some situations, dispersed information can be beneficial for the players. When subversion is not possible even for fully informed experts, because there are not enough good states to pool with mediocre states, it is also impossible with dispersed information. In this case, dispersed information provides a commitment benefit. Players with dispersed information fare better than fully informed experts because they can credibly refrain from using all their information when their communication is scrutinized.

The rest of the paper is organized as follows. In Section 2 we set up our baseline model and show how the committee can be subversive, using an intuitive conversation similar in structure to the introductory example. Next, we generalize the model to a wide class of preferences and priors and construct subversive conversations that work for all specifications of these primitives. Section 3 identifies the costs and benefits of dispersed information. Section 4 reviews the literature, Section 5 concludes, while the Appendix consists of proofs and extensions not contained in the main text.

2 A Model of Subversion

2.1 The baseline model

A committee is composed of two players, 1 and 2, with common interests who can either accept or reject a proposal. Player 1 privately observes signal $x \in [0, 1]$, while player 2 privately observes signal $y \in [0, 1]$. The random variables x and y are uniformly distributed and independent. The players' common payoff from rejecting the proposal is zero, while that from accepting the proposal is $u(x, y) = 1$ if $x + c \geq y$ and -1 otherwise, where $c \in [0, 1]$ is a preference parameter.[3]

The two 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 in an earlier round, and t is odd, player 1 either takes a decision to accept or reject the proposal, or sends a cheap talk message to the other player. Player 2 does the same if t is even. Let M be the set of possible messages with $m_t \in M$ denoting a cheap talk message sent in round t. Abusing notation slightly, let $m_t = A$ denote a decision to accept the proposal and $m_t = R$ a decision to reject it in round t. The game terminates as soon as a player takes a decision.[4]

A pure strategy for player 1 specifies a choice of a message in M or a decision in $\{A, R\}$, in each odd round t, for each possible history of messages m^{t-1} as well as the player's own signal, where m^0 denotes the null history; and similarly for player 2 in even rounds. We refer to a pair of strategies, one for each player, simply as a *conversation*. Our equilibrium notion is perfect Bayesian equilibrium.[5]

As described so far, the problem is straightforward. Player 1 can simply communicate the value of x to player 2 who knows y and then makes the decision. However, the players face a *deniability constraint*. We model the deniability constraint as an uninformed observer with different preferences from the players who will scrutinize their communication ex post. The passive observer could represent a law or social norm, or he could be a regulator, supervisor, or other stakeholder. Because of the preference conflict, the players could be penalized or overruled if they violate the deniability constraint, i.e., if the observer infers that the committee took a decision that is against his interests. While the committee has the authority to take the decision, we assume the committee needs to communicate in a manner that ensures the deniability constraint is met. We now describe the observer's preferences and the deniability constraint facing the committee more formally.

The observer's payoff from the status quo is also zero, while his payoff from the proposal is $v(x, y) = 1$ if $x \geq y$ and -1 otherwise. Since the observer prefers to accept the proposal whenever x is at least as high as y, these variables represent the benefit and cost from the perspective of the observer. For instance, if the observer represents the general public, x and y represent the social benefit and social cost of the proposal.

Let $\mathcal{A} = \{(x, y) \in [0, 1]^2 \mid v(x, y) \geq 0\}$ be the set of states where the observer prefers to accept

[3] We relax these assumptions on priors and preferences in section 2.3. We think of the action taken by the committee as a collective action, although our results extend to some cases of individual decisions taken by each committee member, such as problems of pure coordination.

[4] We set payoffs to zero if they never make a decision. As is standard in cheap talk games, messages have no intrinsic cost or benefit and the message space M is rich enough so that information transmission is constrained only by incentives. Instead of allowing either player to take the decision unilaterally, we could equally employ game forms that let a particular player have decision rights, or assume a decision is taken only after both players vote in favor or ratify it.

[5] In a perfect Bayesian equilibrium, strategies are sequentially rational given beliefs; with beliefs derived via Bayes' Rule if possible, and unrestricted if not.

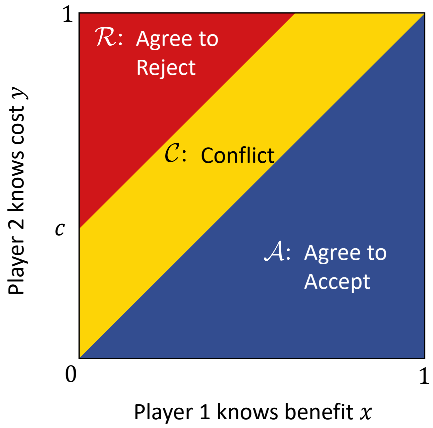
1
$\mathcal{R}$: Agree to Reject
$\mathcal{C}$: Conflict
$\mathcal{A}$: Agree to Accept
Player 2 knows cost y
c
0
1
Player 1 knows benefit x

Figure 2: Conflict and agreement sets

the proposal. Let $\mathcal{R} = \{(x, y) \in [0, 1]^2 \mid u(x, y) < 0\}$ be the set of states where the committee prefers to reject the proposal. The sets $\mathcal{A}$ and $\mathcal{R}$ describe the states where the committee and the observer agree on the optimal decision. The committee also prefers to accept the proposal when $(x, y) \in \mathcal{A}$ while the observer also prefers to reject the proposal when $(x, y) \in \mathcal{R}$. Let $\mathcal{C} = [0, 1]^2 - \mathcal{R} - \mathcal{A}$ denote the zone of conflict between the observer and the committee. For $(x, y) \in \mathcal{C}$, the committee prefers to accept the proposal but the observer does not. The zones of agreement and conflict are depicted in Figure 2.[6]

Our focus is on subversive conversations where the committee subverts the observer's agenda and implements its own ideal decisions, i.e., the committee accepts the proposal if $(x, y) \in \mathcal{A} \cup \mathcal{C}$ and rejects it otherwise. No player has an incentive to deviate from a subversive conversation. After every history of messages m^{t-1} that is followed by a decision in round t, subversion requires updated beliefs to satisfy $\Pr[\mathcal{A} \cup \mathcal{C} \mid m^{t-1}, m_t = A] = 1$ and $\Pr[\mathcal{R} \mid m^{t-1}, m_t = R] = 1$.

The committee must ensure that the uninformed observer never has an incentive to object to the committee decision even after observing the history of messages. This is the deniability constraint. Since the committee never accepts when $(x, y) \in \mathcal{R}$, the deniability constraint following a decision $m_t = A$ can be written as

$$\Pr[\mathcal{A} \mid m^{t-1}, \mathcal{A} \cup \mathcal{C}] \geq \Pr[\mathcal{C} \mid m^{t-1}, \mathcal{A} \cup \mathcal{C}]. \tag{DC}$$

A committee decision to accept the proposal, $m_t = A$, leads the observer to infer that the state belongs to $\mathcal{A} \cup \mathcal{C}$. The deniability constraint says that the observer thinks it is (weakly) more likely that the true state belongs to $\mathcal{A}$ as opposed to $\mathcal{C}$ when the committee accepts the proposal. Under these beliefs, the observer would also prefer to accept the proposal, and so (DC) maintains

[6] In all our figures, the sets $\mathcal{R}$, $\mathcal{A}$ and $\mathcal{C}$ are depicted in red, blue and yellow, respectively. We will call the border between $\mathcal{R}$ and $\mathcal{C}$ the *committee's decision line*, and the border between $\mathcal{C}$ and $\mathcal{A}$ the *observer's decision line*.

deniability that the committee was acting in the observer's interest. A *subversive conversation* is simply a strategy profile that implements the committee's ideal outcome in every state while satisfying (DC).[7]

We impose an admissibility restriction which rules out paradoxical constructions that prevent applying the law of iterated expectations. For a given subversive conversation, let $\mathcal{H}_t^* = \{m^t \mid m_t = A\}$ be the set of all message histories m^t that terminate in round t with a decision $m_t = A$ to accept the proposal. The deniability constraint (DC) must hold for each element of $\mathcal{H}_t^*$. Our admissibility restriction on the conversation says that it must also hold when we integrate over all message histories that belong to any $\mathcal{H} \subset \mathcal{H}_t^*$. It automatically holds in any finite environment. In the continuum, this restriction rules out constructions that create measure-theoretic paradoxes. We conclude this section with some additional remarks on the model.

The conditioning on the history of communication in (DC) reflects the ability of the observer to have ex post access to committee deliberations, e.g., by subpoena or FOIA laws. Our specification of preferences states that both types of error, convicting the innocent and acquitting the guilty, are equally costly for the observer. This makes the deniability constraint equivalent 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 acquitting the committee. This implies the committee will also be acquitted under the more demanding "reasonable doubt" burden of proof used for criminal trials. We assume that the observer understands the meaning of messages in the same way as the players, i.e., encryption is impractical or prohibited.

When (DC) is met, the observer approves of every committee decision ex post. By the law of iterated expectations, the observer has no incentive to object even after observing histories where the committee has not yet made a decision. In other words, a subversive conversation implements the committee's optimal decisions not only under ex post scrutiny but also under contemporaneous scrutiny by the observer. Indeed, we can allow the observer 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 of the conversation, or the message space, or force a vote on the decision, or choose any other ex ante design aspect of the committee meeting. We assume that procedural rules are either given by precedent or history, or that the committee has the authority to design them.

We suppose that the committee encounters its decision problem frequently, such as meeting at regular intervals or making multiple decisions during a single meeting. So it needs to devise a complete contingent plan of communication that will work for every realization of the state.

[7] We use the standard approach of using the joint density to compute conditional probabilities. The constraint (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 $\mathbb{R}^1$, in which case we use the Lebesgue measure on $\mathbb{R}^1$; and similarly for cases where the residual sets are finite.

A conversation is such a plan.[8] We restrict attention to sequential communication, or *polite talk* in the language of Aumann and Hart (2003), since our results do not rely on simultaneous messaging. 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.

2.2 Existence in the baseline model: a gradual conversation

Our first result establishes the existence of subversive conversations in the baseline model and shows that any attempt at subversion must involve a back and forth conversation.

Theorem 1. *In the baseline model, (i) a subversive conversation exists for each* $c \in [0, 1]$*, and (ii) any subversive conversation requires at least three rounds to complete for* $c \in (0, 1)$*.*

Part (ii) of Theorem 1 is easy to see from Figure 2. If $c < 1$, the players must exchange information to determine their ideal decision. But if $c > 0$ and $x < c$, player 1 cannot reveal her signal in round 1 and expect to satisfy (DC) if the proposal is subsequently accepted in round 2. So a decision cannot be taken before round 3. As long as the players need to exchange information and there is any conflict, a back and forth conversation is necessary for subversion.

To demonstrate part (i) of Theorem 1, we construct a subversive conversation similar in structure to the introductory example. Figure 3 depicts this conversation for the case $c = 4/9$. The left most panel depicts the state space at the beginning of the game. Starting there, in each round the player who moves either reveals her signal when it belongs to a specific interval (down arrow to the next bottom panel), or she says "pass" revealing her signal does not belong to that interval (right arrow to the next top panel). When a player reveals her signal, the other player subsequently takes a decision taking into account her own signal. When a player passes, the conversation moves to the next round, where it is the other player's turn to speak. The rightmost pair of panels depict the end phase of the conversation, and we describe what happens there in more detail below.[9]

In round 1, player 1 reveals her signal x if $x \in [4/9, 5/9]$, passing otherwise. Looking at the bottom panel for round 1, notice that middling information in $[4/9, 5/9]$ can be revealed because the deniability constraint (DC) is satisfied if player 2 subsequently accepts the alternative in round 2. In particular, after $x \in [4/9, 5/9]$ is revealed and player 2 proposes acceptance, the

[8] Communication strategies can be modeled as ex ante plans of action that must be interim incentive compatible (Green and Stokey, 2007). 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."

[9] For ease of exposition, we describe the messages sent by each player in terms of the meaning deduced by the other player. In practice, the players may use non-literal language or other conventions such as revealing how surprising their news is (i.e., how far it is from the average), using hesitation as an indication of atypical news, or the time when a player calls for a meeting as an indication of her signal's value, etc. Such conventions could be seen as a form of encryption and so we suppose that the observer can see through them.

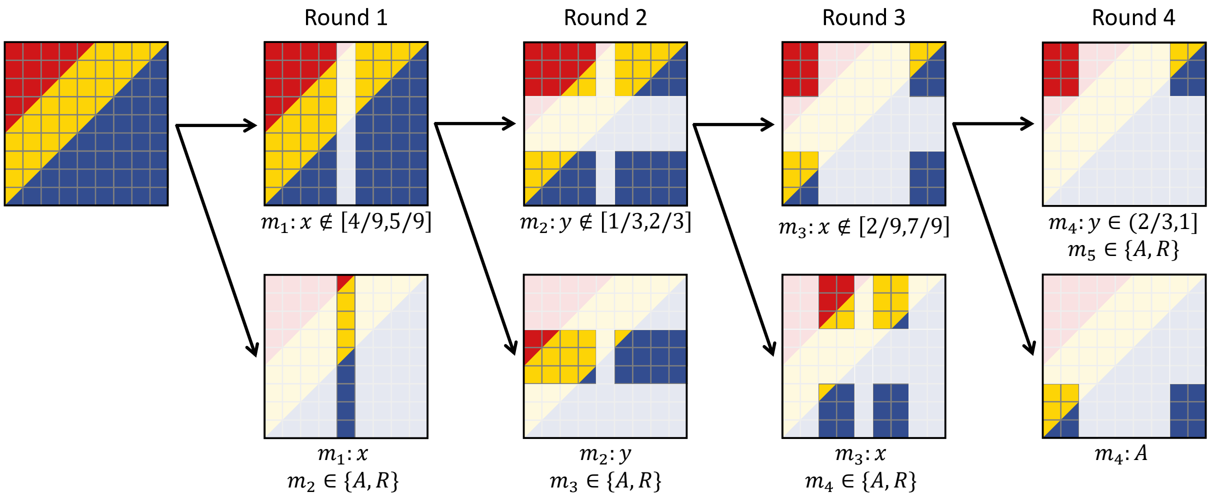

Figure 3: Subversive conversation: gradual protocol for $c = 4/9$.

residual part of $\mathcal{A}$ is the interval $[0, x]$, while the residual part of $\mathcal{C}$ is the interval $[x, x+4/9]$. The former is (weakly) larger than the latter for all $x \geq 4/9$. For $x < 4/9$ the deniability constraint is violated and such x cannot be safely revealed by player 1 in the first round. So player 1 does not reveal such unfavorable information in round 1.

For $x > 5/9$, the deniability constraint will be satisfied, but revealing such favorable information in the first round will jeopardize the possibility of subversion in later parts of the conversation. This is the key point of the construction. If $x \notin [4/9, 5/9]$, player 1 says pass. Everyone learns from this move that $x \notin [4/9, 5/9]$, i.e., that the state (x, y) is either in $[0, 4/9) \times [0, 1]$ or in $(5/9, 1] \times [0, 1]$, as depicted in the top panel of round 1.

Now consider round 2 after player 1 has passed the conversation to player 2. In this round, player 2 perfectly reveals y, provided $y \in [1/3, 2/3]$ (bottom panel); otherwise she says pass (top panel) and we move to the next round. What happens when player 2 reveals y in round 2? Given the history of the conversation, the deniability constraint is satisfied for all possible types that are revealed by player 2 in this round when player 1 subsequently accepts the proposal. For instance, suppose player 2 reveals $y = 2/3$, which is the highest possible cost y that player 2 reveals in this round. If player 1 accepts subsequently, the residual part of $\mathcal{A}$ is $[y, 1] - [4/9, 5/9]$ while the residual part of $\mathcal{C}$ equals $[y - 4/9, y] - [4/9, 5/9]$, and the former is at least as large as the latter for all $y \leq 2/3$. So the deniability constraint is satisfied at this stage when player 2 reveals $y = 2/3$ and player 1 accepts the proposal.

Notice the key role of the information revealed in round 1, namely $x \notin [4/9, 5/9]$, in satisfying the deniability constraint in round 2. If this information was not revealed by player 1's pass in round 1, the deniability constraint would not be satisfied when player 2 reveals $y = 2/3$. In fact, this is true for any $y > 5/9$ that is revealed in round 2. By not revealing types $x > 5/9$ in round 1, player 1 creates slack in the deniability constraint for types $y \in (5/9, 2/3]$ of player 2 where none existed before. The round 1 pass also tightens the deniability constraint for lower types of player 2, $y \in [1/3, 4/9)$, because for those types it removes states that belong to the agreement set $\mathcal{A}$. But these types of player 2 have slack to spare in the deniability constraint.

During the process of information revelation that we are describing, each player creates a context that describes which states remain possible and which do not. This process must be gradual, since otherwise one player may create an unfavorable context for the other player who would no longer be able to reveal her signal or take a decision that she could have otherwise. This was also the key force behind the conversation in introductory example.

A novel feature not present in the introductory example arises here because of a richer type space. The players dynamically create slack in the deniability constraint for each other. By doing this, each player creates a favorable context that allows the other player to provide information that she could not safely reveal before. The process we are describing is indeed a conversation. The players cooperatively create favorable contexts and avoid unfavorable ones, in order to ensure their preferred decisions can be taken once the right context has been created. A subversive conversation is designed for perfect coordination while maintaining deniability.

The remaining panels of Figure 3 continue this line of reasoning. If a decision has not been taken before round 3, player 1 reveals the exact value of x if $x \in [2/9, 7/9]$, passing the conversation back to player 2 otherwise. As shown in the figure, types $x \in [2/9, 4/9)$ can be safely revealed by player 1 in this round, even though they could not be in round 1, because of the favorable context created player 2's pass in round 2. The deniability constraint will be satisfied when the other player subsequently chooses to accept.

When $x \notin [2/9, 7/9]$, player 1 passes again. The remaining possible states are depicted in the two rightmost panels and this is the end phase of the conversation. In round 4, player 2 simply accepts the proposal if $y < 1/3$. She does not need any information from her partner to do this because the residual part of $\mathcal{R}$ is empty for such y. Doing so will also meet the deniability constraint because the residual part of $\mathcal{A}$ is greater than that of $\mathcal{C}$, as can be seen from the figure. If $y > 2/3$, player 2 passes the conversation back to player 1. Subsequently, in round 5, player 1 can reject the proposal if $x < 2/9$ and accept it if $x > 7/9$ while meeting the deniability constraint, as shown in the figure.

In the proof of Theorem 1(i) we use such a construction to establish the existence of a subversive conversation for any $c < 1/2$. In that proof, we extend the existence result also to the case $c \geq 1/2$, although the conversations we employ for this latter case differ from the construction described above, especially in the initial rounds, to take into account the larger conflict between the committee and the observer. We use the insights from this construction throughout the paper—a subversive conversation is a process of creating the right contexts that allow the committee to take its desired decisions while maintaining deniability.

2.3 The generalized model

Consider the following generalization of the baseline model. We assume independent and identical priors for x and y. Let $F(\cdot)$ denote the common cumulative distribution function and let $\mathcal{S} \subseteq \mathbb{R}^2$ denote the support of $F \times F$. The probability measure represented by F can be continuous, discrete, singular, or a mixture of the three.[10]

We generalize the preferences of the committee and the observer. We suppose that these preferences determine the (measurable) agreement and conflict sets $\mathcal{A}$, $\mathcal{R}$ and $\mathcal{C}$ defined earlier. These sets can be arbitrary except for one restriction. We suppose that the observer considers the proposal acceptable if the benefit x is at least as high as the cost y. In other words, $\mathcal{A} \supseteq \mathcal{L}$, where $\mathcal{L} = \{(x, y) \in \mathcal{S} \mid y \leq x\}$ is the lower region of $\mathcal{S}$ on or below the diagonal. Denote by $\mathcal{U}$ the complement of $\mathcal{L}$ in $\mathcal{S}$. Notice that $\mathcal{A}$ is assigned strictly positive probability by the priors.[11]

We can use the cumulative distribution functions to transform the underlying random variables to their quantiles. This is straightforward at all points of continuity of F. At a point of discontinuity, we can determine the exact value of the quantile via a uniform randomization. Since x and y are identically distributed, our assumption on preferences, $\mathcal{A} \supseteq \mathcal{L}$, continues to hold after this transformation. Quantiles are uniformly distributed on $[0, 1]$. So it is without loss of generality to assume that x and y themselves are uniformly distributed, with the underlying state space $\mathcal{S} = [0, 1]^2$. This amounts to reinterpreting the types x and y as their quantiles or, equivalently, that the players converse in terms of quantiles (e.g., player 1 says "x is in the tenth percentile"). An instance of the generalized model is then fully specified by the configuration of preferences. A game $\Gamma = \{\mathcal{R}, \mathcal{C}, \mathcal{A}\}$ is such a configuration.

Suppose a subversive conversation exists for a game Γ. Under what conditions can the same conversation also be subversive in another game $\Gamma' = \{\mathcal{R}', \mathcal{C}', \mathcal{A}'\}$ that potentially has different conflict and agreement sets? To answer this question, notice that in any round t after a history m^{t-1}, the player moving in that round either takes a decision $m_t \in \{A, R\}$ or sends a message $m_t \in M$ to the other player. We may partition her types into *decision types* that take a decision after history m^{t-1}, and *messaging types* that send a message. Of course, one of these two sets may be empty. For a fixed conversation, the partitioning into decision types and messaging types and the messages sent by each messaging type will be unchanged in Γ and Γ', in any round t after any history m^{t-1}.

If the players are to achieve their ideal outcomes, a decision type of the player moving in round t must be allowed to take potentially different decisions in Γ' compared to in Γ, to reflect her possibly different preferences. We allow different decisions to be taken in Γ', by each type

[10] The arguments of this section extend to the case of non-identical priors if the distribution of x dominates that of y in the FOSD sense. They also extend to the case of statistical dependence between x and y under a MLRP condition. See Appendix A.2.

[11] While the example in Figure 1 does not satisfy our assumption on preferences, it is covered by subsequent model variations that we consider. See Section 3.

within the set of decision types, after any history m^{t-1}. However, the partitioning into decision types and messaging types, and the actual message sent by each messaging type, does not differ across Γ and Γ'. If the deniability constraint (DC) is always satisfied in Γ', then the players manage to subvert in the same way that they did in Γ. They use identical messages and the conversation continues or stops under identical conditions. After the players stop exchanging messages, the particular decision taken is tailored to preferences. The same conversation is subversive in both Γ and Γ'. Our first result of this section identifies conditions under which this can happen.

To state the result, we need two more definitions. Call a subversive conversation for Γ a *fine subversion* if every time the committee accepts the proposal, either the realized value of x or the realized value of y has been perfectly revealed. A subversive conversation that is not fine is a *coarse subversion.* In a fine subversion the player who takes the decision necessarily knows the exact state (x, y) but the other player has no informational advantage over the observer and both she and the observer may remain uncertain about one dimension. In a coarse subversion, the player taking the decision may also be uncertain of the exact state but this does not affect her optimal decision.[12]

Lemma 1 (Subset Lemma)**.** *(i) If a fine subversion exists for some game* Γ*, then the same conversation is also a fine subversion for any other game* Γ' *with* $\mathcal{C}' \subseteq \mathcal{C}$*,* $\mathcal{A}' \supseteq \mathcal{A}$ *and* $\mathcal{R}' \supseteq \mathcal{R}$*. (ii) If a coarse subversion exists for some game* Γ*, then the same conversation is also a coarse subversion for any other game* Γ' *with* $\mathcal{C}' \subseteq \mathcal{C}$*,* $\mathcal{A}' \supseteq \mathcal{A}$ *and* $\mathcal{R}' = \mathcal{R}$*.*

Lemma 1 says that a subversive conversation remains so if we change observer or committee preferences in order to reduce, in the sense of subsets, the zone of conflict $\mathcal{C}$ and increase the zone of agreement $\mathcal{A}$. Regardless of whether it is a coarse or fine subversion, the same conversation is subversive in Γ' if $\mathcal{R}' = \mathcal{R}$. For a fine subversion, the case $\mathcal{R}' \supset R$ is also allowed. A fine subversion is especially robust to preference specifications because the player taking the decision does so knowing the exact value of x and y. With full information it is always possible for her to tailor the decision to her exact preferences.

To establish the lemma we need to show the deniability constraint will also be met. This follows from the following observations. In any round t, a decision type of the player who moves in that round sees the same history in Γ and Γ', because the history is generated by the same fixed conversation. Since $\mathcal{A}' \supseteq \mathcal{A}$ and $\mathcal{C}' \subseteq \mathcal{C}$, these same inclusion relationships must be preserved for the residual parts of these sets that one obtains after deleting the states ruled out by the observed history.

In the case of a fine subversion, suppose (without loss of generality) that the exact value of x

[12] The construction depicted in Figure 3 is a coarse subversion since neither x nor y is publicly revealed after round 4. This end phase can be modified to make the conversation a fine subversion that will take one more round to complete.

has been revealed and player 2 has accepted the proposal. Since (DC) obtains in Γ, the observer attaches greater likelihood that the state (x, y) belongs to the residual part of $\mathcal{A}$ compared to the residual part of $\mathcal{C}$. But then the same must be true in Γ', after the same observed history, revealed value of x, and decision to accept. Because the original inclusion relations are preserved for the residual sets, the observer must attach (weakly) higher probability to the residual part of $\mathcal{A}'$ compared to that of $\mathcal{A}$, and a (weakly) lower probability to the residual part of $\mathcal{C}'$ compared to that of $\mathcal{C}$. The deniability constraint (DC) will be met in Γ' since it is met in Γ.

For a coarse subversion, if at any point in Γ a player takes a decision to accept without knowing the other player's type, it must be that the residual agreement set $\mathcal{R}$ is empty conditional on the observed history and the type of the player taking the decision. Otherwise the conversation cannot be subversive in Γ. Since in this case $\mathcal{R}' = \mathcal{R}$, the same must be true in Γ' at the same stage. It follows that the same decision will also be optimal for the players and satisfy (DC) in Γ', once again since $\mathcal{C}' \subseteq \mathcal{C}$ and $\mathcal{A}' \supseteq \mathcal{A}$. This establishes Lemma 1.[13]

Is it possible that the same conversation is subversive for every specification of preferences? We call such a conversation a *universal* subversive conversation since it is subversive for every instance Γ of the generalized model. Our next result is about the existence and length of universal subversive conversations.[14]

Theorem 2. *In the generalized model, (i) there exists a four-round universal subversive conversation, and (ii) any universal subversive conversation requires at least four rounds.*

Part (ii) of Theorem 2 is the analogue of Part (ii) of Theorem 1. To establish part (i), we consider the following conversation:

- Round 1. Player 1 says "x is in $\{a, 1-a\}$" if $x = a$ or $x = 1-a$, for some $a \in [0, 1/2]$.

- Round 2. Player 2 perfectly reveals her type y if $y \in (a, 1-a)$; otherwise she says "y is not in $(a, 1-a)$".

- Round 3. If player 2 has revealed her type y in round 2, player 1 takes a decision in round 3 and the conversation ends. Otherwise, player 1 perfectly reveals her type x in round 3 and the conversation proceeds to round 4.

- Round 4. Player 2 takes a decision.

[13] Notice that Lemma 1 covers cases that go beyond the generalized model described above. In particular x and y do not need to be identically distributed or statistically independent. We use these facts in the appendix where we present a number of additional robustness results.

[14] The length of a conversation is the maximum number of rounds it can take till a decision is made. There is no upper bound on the length of subversive conversations. There exist universal subversions that end almost surely in finite time, even though it is not possible to specify in advance how long the conversation will take.

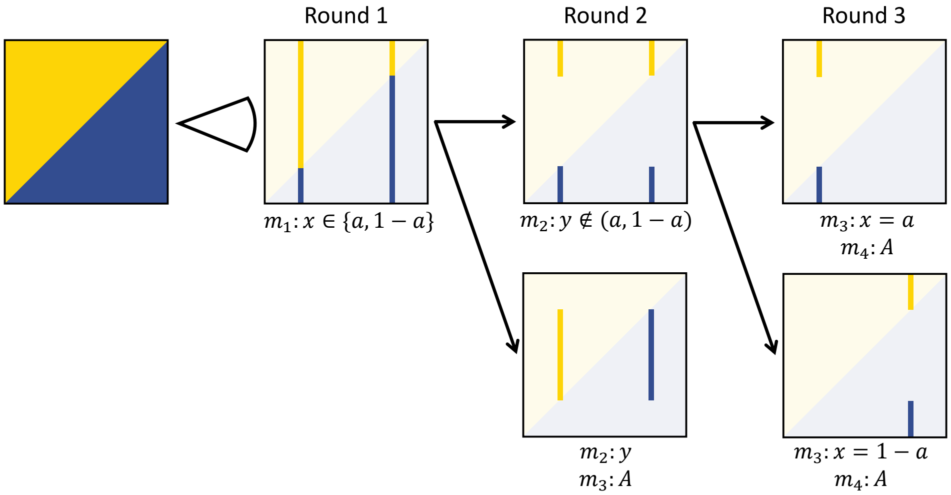

Figure 4: Universal subversive conversation for test case

.

We consider a game $\Gamma = \{\mathcal{R}, \mathcal{C}, \mathcal{A}\}$ in which $\mathcal{R}$ is empty, $\mathcal{A} = \mathcal{L}$ and so $\mathcal{C} = \mathcal{U}$. We show that the conversation described above, and depicted in Figure 4, is a fine subversion for this game. To do so, we need to show why the deniability constraint (DC) will always be satisfied.

Consider first the case where player 1 takes a decision in round 3. Since $\mathcal{R}$ is empty, player 1 will accept the proposal. From the history of the conversation, everyone knows at this stage that $x \in \{a, 1-a\}$ for some $a \leq 1/2$, as well as the exact value of $y \in (a, 1-a)$. We must have $a < 1/2$ since otherwise $(a, 1-a)$ is empty. There are two possible states of the world at this stage, (a, y) and $(1-a, y)$. We must have $(1-a, y) \in \mathcal{A}$ and $(a, y) \in \mathcal{C}$. Since (a, y) and $(1-a, y)$ are equally likely, (DC) is satisfied.

Consider next the case where player 2 takes the decision in round 4. From the history of the conversation, the exact value of $x \in \{a, 1-a\}$ is now publicly known, as is the fact that $y \notin (a, 1-a)$. Once again, since $\mathcal{R}$ is empty, player 2 will accept the proposal. The residual part of $\mathcal{A}$ at this stage is equal to the interval $[0, a]$ regardless of the revealed value of x, while the residual part of $\mathcal{C}$ equals the interval $[1-a, 1]$. Since these two sets are of equal measure, (DC) is met in this case as well.

The game depicted in Figure 4 is a "test case". By Lemma 1, the fine subversion described above for this test case will also be a fine subversion for every other instance of the generalized model. There will be no difference in the messaging by the players. The only difference from the test case that can arise is that a decision type of a player can sometimes reject the proposal if $\mathcal{R}$ is non-empty. Since the observer always agrees with such a decision this will not prevent subversion. We therefore have a universal subversive conversation.

In our problem of subversion, each player holds one piece of information while their optimal decision depends on both pieces of information. We can compare our committee with dispersed information with the benchmark case of a fully informed expert. Such an expert knows both x and y and has the same preferences as the players. Because she has all the information, she knows her optimal decision at the outset and does not need to send messages. She can meet the deniability constraint and subvert if and only if the agreement set $\mathcal{A}$ is at least as large in measure as the conflict set $\mathcal{C}$. This non-negative total slack condition is necessary and sufficient for the fully informed expert to subvert.

By the deniability constraint, non-negative total slack is necessary also for the committee to subvert. But it is not sufficient. The players need to engage in a back and forth conversation to determine their optimal decision while meeting the deniability constraint at the same time. The total slack condition does not guarantee such a conversation exists. Under non-negative total slack, a committee with dispersed information will do at most as well as the fully informed expert. Dispersed information can only be costly. It will impose no cost if and only if the committee is able to subvert. Since Theorem 2 shows the committee can subvert in every instance of the generalized model, we have its first corollary.

Corollary 1. *In the generalized model, the committee with dispersed information does as well as a fully informed expert.*

Corollary 1 shows that in the generalized model, the committee does as well as possible. Because the committee's messages and decisions are both scrutinized by the observer, we may say that the committee communicates in public. Corollary 1 also allows us to compare public communication with the benchmark of secure private communication.

Under secure private communication the committee's decision is scrutinized, 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 that they will coordinate on their optimal decisions. Under this selection, secure private communication is identical to the case of a fully informed expert. So Corollary 1 tells us the committee does as well under public communication as under secure private communication.[15] In Section 3 we consider situations not covered by the generalized model where the committee with dispersed information does strictly worse than the fully informed expert, as well as situations where the committee does better.

Our results so far have assumed that both committee preferences as well as regulator preferences are common knowledge. But our arguments can be extended to cover cases where the conflict between the committee and the observer is uncertain, not only in magnitude but also in sign. The next two corollaries of Theorem 2 address this topic.

[15] Note also that in environments where private communication is not allowed but the committee has somehow engaged in it, the players can hide this fact and maintain the charade of public communication by employing subversive conversations of the kind we have constructed in Section 2.

Consider again the baseline model and suppose that the observer is not certain of the committee's preferences. The observer entertains the possibility that the preference parameter is either c or c' with $c < c'$, as depicted in Figure 5(a). By Theorem 2, this uncertainty will not affect the committee's ability to subvert. If the committee uses a universal conversation, the deniability constraint will be met even when the observer holds the extreme belief that $c = 1$. So it will also be met for any more moderate observer beliefs about committee preferences. The argument does not depend on the assumed linearity of the baseline model and we state it to cover all cases of preference uncertainty in the generalized model.

Corollary 2. *In the generalized model with preference uncertainty, a subversive conversation exists when the committee believes the game is* Γ *whereas the observer believes the game is* Γ'*, for any* Γ *and* Γ'*.*

While Corollary 2 covers all cases where preferences are not exactly known, it assumes that the sign of the conflict between the committee and the observer is common knowledge. The committee can only be biased in one direction relative to the observer. Figure 5(b) shows a different situation where both the sign and the magnitude of the committee's conflict with the observer is state dependent. In the figure, the set $\mathcal{A}$ of states where the observer prefers to accept the proposal is equal to $\mathcal{L}$, as in the baseline model. But the *committee's decision line* that forms the border between the states where the committee prefers one decision or another is given by a (possibly non-linear) continuous function $y = C(x)$. 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(b), the committee's decision line crosses the observer's decision line $y = x$ from above at the point labeled P.[16]

With two possible kinds 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 generalized model which only allows for one kind of conflict and so one kind of deniability constraint. Nevertheless, the following corollary provides conditions under which a subversive conversation exists.

Corollary 3. *In the generalized model with two kinds of conflict and* $\mathcal{A} = \mathcal{L}$*, a subversive conversation exists for any increasing committee decision line* $y = C(x)$*.*

Looking at Figure 5(b), player 1 first discloses whether x lies to the right or left of the x-coordinate of the intersection point P. Player 2 then reveals whether y is above or below the y-coordinate of P. Since $C(x)$ is increasing, there is no conflict between the committee and the

[16] In the situation depicted, the proposal may be "easy" to implement for small values of (x, y) and "hard" to implement for high values of (x, y). The observer does not care about the difficulty of implementation but the committee does, and it is more inclined than the observer to accept an easy project and less inclined than the observer to accept a hard one.

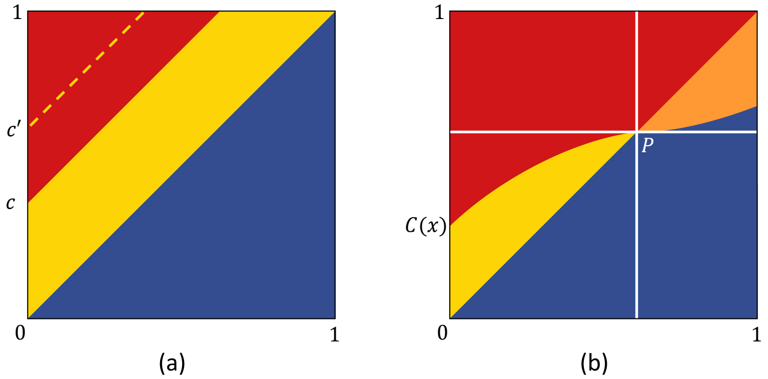

Figure 5: Uncertain conflict: magnitude and sign

observer whenever the state is revealed to be either in the bottom right or the top left element of the resulting partition. The committee can clearly get its way in these states.

On the other hand, if the state is revealed to be in the lower left element of the partition, the situation is covered by Theorem 2, in particular since $\mathcal{A} = \mathcal{L}$ and x and y are identically distributed also conditional on this event. For the same reason, Theorem 2 also covers the case where the state is revealed to be in the upper right element of partition, since this situation is just an "upside down" version of the generalized model with reversed roles for the decisions to "accept" and "reject". 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.

To obtain additional results beyond Theorem 2 and its corollaries, one approach is to construct subversive conversations that always satisfy the deniability constraint with slack. If the slack is bounded away from zero, a subversive conversation for an instance of the generalized model is also subversive for small perturbations of the priors and preferences in the model. In Appendix A.3 we provide examples of subversive conversations based on the gradual protocol that have such slack .

3 Costs and benefits of dispersed information

In this section we consider environments outside the domain of the generalized model where the committee with dispersed information does not do exactly as well as the fully informed expert. In Section 3.1, we provide examples where the committee cannot subvert but the fully informed expert can. Dispersed information is strictly harmful for the players in this case. We also provide necessary conditions for fine subversions that are tighter than the total slack condition and that provide key insight into our constructions of the previous sections. In Section 3.2, we consider the case of negative total slack. We show that in this case the committee with dispersed information can do strictly better than a fully informed expert because of a commitment effect arising out of information being dispersed.

3.1 Dispersed information is costly for the players

Are there situations outside the domain of the generalized model but with non-negative total slack in which the committee with dispersed information cannot subvert? Because non-negative total slack is sufficient for the fully informed expert to subvert, the committee will do strictly worse than the fully informed expert in these cases. We consider this question now. We relate our results to Hall's Marriage Theorem (1935). In order to do so, we restrict attention to cases with a finite number of types for each player and uniform i.i.d. priors. The non-negative total slack condition can then be stated in terms of the cardinality of the sets $\mathcal{C}$ and $\mathcal{A}$. We assume $|\mathcal{A}| \geq |\mathcal{C}|$.

Hall's theorem considers a matching problem between a finite set of "women" and "men". Each woman w has a set $H(w)$ of acceptable men that w is (equally) happy to be matched to. For a man, any woman who finds him acceptable is acceptable. Hall's theorem states that a necessary and sufficient condition to match up the women and men in mutually acceptable pairs is the following: for every subset W of women, the union $\cup_{w \in W} H(w)$ of the sets of their acceptable men must contain at least as many men as there are women in W. Intuitively, this condition requires there be sufficient diversity of preferences. For instance, if two women find one and the same man to be the only acceptable one, it is impossible to match both these women.

To apply Hall's theorem to our setting, notice that subversion involves matching conflict points in $\mathcal{C}$ with agreement points in $\mathcal{A}$. Since the committee's information is dispersed, the committee is further constrained to match conflict and agreement points that lie in the same row or column, in the case of a fine subversion. For each conflict point $w \in \mathcal{C}$, let $H(w)$ be the set of agreement points in $\mathcal{A}$ that lie in the same row or column. In a fine subversion, each w must be matched with an element of $H(w)$. To rule out trivial failures of subversion, we assume that $H(w)$ is non-empty for each $w \in \mathcal{C}$ and that each element of $\mathcal{A}$ belongs to $H(w)$ for some $w \in \mathcal{C}$. Using Hall's theorem, we have the following result.

Proposition 1. *Assume discrete types and uniform i.i.d. priors. A fine subversive conversation exists only if, for every $W \subseteq \mathcal{C}$,*

$$\left| \cup_{w \in W} H(w) \right| \geq \left| W \right|. \tag{HC}$$

Notice first that Hall's condition (HC) implies, but is more demanding than, the non-negative total slack condition, $|\mathcal{A}| \geq |\mathcal{C}|$, identified earlier as a necessary condition. We can think of the problem of subversion as one of "justifying" a conflict point in $\mathcal{C}$ by pooling it with an agreement point in $\mathcal{A}$. Just as Hall's condition can be interpreted in the classic marriage example in terms of diversity of preferences, it can be interpreted in our setting in terms of diversity of possible justifications. Whether or not there is such diversity in any given game depends on both the committee's and the observer's preferences.

To see why (HC) is a necessary condition for the existence of a fine subversion, consider again the fully informed expert who knows both x and y. But suppose now that the fully informed expert is constrained to reveal either the realized value of x or the realized value of y when she takes a decision. Call this benchmark the *constrained fully informed expert.* She is constrained to match each conflict point $w \in \mathcal{C}$ to some agreement point in $H(w) \subseteq \mathcal{A}$ that belongs to the same row or column. The only difference between her and the committee engaged in a fine subversion is that she has all the information required for the decision at the outset and does not need to exchange dispersed information in public.[17]

By Hall's theorem, (HC) is necessary and sufficient for the constrained fully informed expert to subvert. It is necessary because if she can subvert she produces a matching and so (HC) must hold. It is sufficient because if (HC) holds there exists a matching that she can employ to subvert. To see why Proposition 1 follows, notice that whenever the committee has a fine subversive conversation, the constrained fully informed expert can simply mimic the committee and so she can also subvert. So (HC) must hold.

Figure 6(a) provides a simple example where (HC) fails and so a fine subversion is impossible. As can be seen from the shape of $\mathcal{A}$ in the figure, the observer cares only about the cost y and not about the benefit x. Every conflict point in $\mathcal{C}$ can only be matched with a point in $\mathcal{A}$ that lies vertically below it and in the same column. There is insufficient diversity of possible justifications for the conflict points. When W is considered to be the set of conflict points in the rightmost column, (HC) fails. By Proposition 1, there does not exist a fine subversion for the committee.

Indeed, a subversive conversation of any kind, coarse or fine, does not exist in the example of Figure 6(a). Because there is zero total slack, every agreement point must be pooled with a conflict point. Since no agreement point has a conflict point in the same row it can be pooled with, this implies player 2 cannot ever reveal her exact signal. So player 1 has to reveal her exact signal in order for the committee to figure out if the state belongs to $\mathcal{R}$ or not and determine its optimal decision. But whenever player 1 reveals the signal corresponding to the rightmost column, the failure of (HC) in this column implies the deniability constraint cannot be met.[18]

Hall's condition is not sufficient for the existence of a fine subversive conversation. It is not

[17] Glazer and Rubinstein (2004) characterize optimal rules of persuasion for a speaker with knowledge of multiple aspects of a decision problem, communicating with a listener who can obtain evidence on one of the aspects. While the underlying situation is similar, our constrained fully informed expert, and not the listener, volunteers evidence on her own.

[18] The initial imbalance between the conflict and agreement points in the rightmost column will remain even if some player reveals some information prior to player 1 revealing her highest signal. For instance, if player 2 accepts the proposal by pooling the middle row with any one of the bottom two rows before player 1 reveals her signal, (HC) will continue to fail in the right column of the remaining game, preventing a subversion.

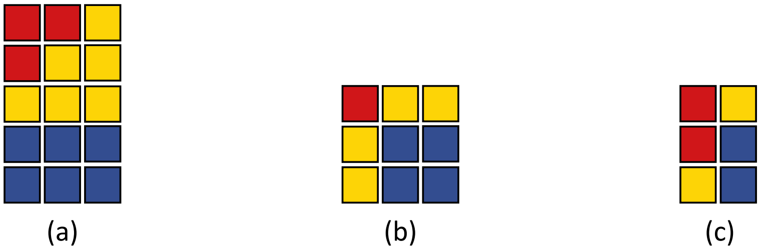

Figure 6: Hall's theorem and subversion

enough that the game satisfies (HC) at the outset. It must continue to do so after every move that is made during a fine subversion. Figure 6(b) illustrates this. Hall's condition is satisfied so the constrained fully informed expert can subvert, but a fine subversion is impossible for the committee. For instance, if player 1 starts the conversation by revealing either the right-most column or its complement, it will be impossible to subvert when the realized state belongs to the two left columns because the number of conflict points exceed the number of agreement points and there is negative slack and so a failure of (HC) in that subset of the state space. This is true regardless of player 1's first move. There is no move for either player that delivers (HC) for each of the possible continuation games.[19]

A fine subversive conversation is an inherently dynamic process that must satisfy (HC) recursively. When it is her turn to move, each player, using her own information, must be able to partition the remaining possible states, each element of which is a continuation game that satisfies (HC). Furthermore, this must be done in a way that enables the other player, using her own information, to do the same thing in the next round. This gradual, branching process of providing increasingly refined contexts must go on till one player perfectly reveals her information and a decision is taken. Such dynamic considerations are absent from the formulation of the matching problem in Hall's theorem—the constrained fully informed expert can take her decision as soon as she learns her type. Dynamic extensions of Hall's condition may yield general insight into the existence of subversive conversations and we leave this question for future research.

An interesting special case where (HC) is sufficient for a fine subversion is when one player (say, player 1) has only two possible types. Figure 6(c) is an example. By Hall's theorem, if (HC) holds a matching exists. What are its properties? Some conflict points in $\mathcal{C}$ may be cross matched with agreement points in $\mathcal{A}$ that belong to the same row of a different column, with the remaining conflict points matched within their own column. Player 2 can first reveal her type for each row corresponding to a cross match. Player 1 then accepts the proposal. Since there are only two columns, this cannot affect any other agreement or conflict point, apart from the two that are pooled. If player 2's type does not belong to the rows corresponding to the cross

[19] A coarse subversion is also impossible in this example, in particular as $\mathcal{R}$ is non-empty and so the players need to exchange some information to determine the optimal decision.

matches, she passes. Since the remaining conflict points are matched within their own columns, these matches can now be made via player 1 revealing her type and player 2 then taking the appropriate decision. In the example depicted in Figure 6(c), the cross match occurs in the lowest row, with the remaining matches occurring within a column.

The reader will recognize Figure 6(c) as the residual state space in the example of Figure 1 after player 1 first passes the conversation over to player 2, thereby revealing she has only two possible types $x = 0$ or $x = 1$. Since (HC) is satisfied after such a move, and trivially satisfied when instead x is revealed to equal 1/2, a fine subversion exists in the example. This argument also sheds light on the construction underlying Theorem 2. If one player can partition her type space, in an admissible way, into 2-column or 2-row pairs that satisfy (HC), then the two players can engage in a fine subversion on each revealed pair. Such an admissible partition exists in the test case depicted in Figure 4, as shown in Section 2.3.[20]

3.2 Dispersed information is beneficial for the players

Suppose we have a game with negative total slack, i.e, $\mathcal{A}$ is strictly smaller in measure than $\mathcal{C}$. Neither the committee not the fully informed expert can subvert since a necessary condition for subversion is violated in either case. What can they do? Consider the discrete type example in Figure 7. Each player has two possible signals and priors are i.i.d. uniform so that we have negative total slack. Since neither the committee nor the fully informed expert can subvert under negative total slack, 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 case of the committee first. There exists an equilibrium where the committee accepts the proposal in some states and rejects it in others without the observer objecting. In this equilibrium, when player 1 has the low signal (corresponding to the left column) she rejects the proposal, and accepts the proposal otherwise. If player 1 does not take a decision in round 1 and deviates by sending a message to the other player, both player 2 and the observer believe that player 1 has the low signal. Since the observer will never approve a subsequent decision to accept the proposal given these beliefs, we can assume without loss that the players always take the decision to reject in the continuation game that follows this deviation. So player 1 has no incentive to deviate by sending a message in round 1. Player 1 also has no incentive to change her decisions. On the path of play, the observer will approve of a decision to accept since he infers the state of the world must be in the right column, and approve of a decision to reject given that he infers the state must be in the left column. Since the observer sometimes approves the proposal, the players get strictly positive expected payoffs.

[20] Strictly speaking, Hall's condition applies to discrete type models, while Section 2.3 allows a continuum of types. The connections we draw here are precise for discrete type models, and provide intuition for continuum of

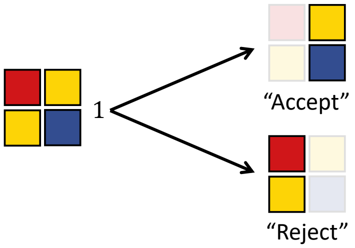

Figure 7: Equilibrium under negative total slack

Next, consider the fully informed expert for the same example. The only decision rule that can be supported in equilibrium involves always rejecting the proposal. For if there is an equilibrium where the proposal is accepted with positive probability, 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 because $\mathcal{C}$ is larger than $\mathcal{A}$. Notice that this point extends to any game with negative total slack. The fully informed expert earns zero expected payoff, yielding our next result.

Proposition 2. *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.*

With negative total slack, the fully informed expert can only reject the proposal and earn zero payoffs. The committee with dispersed information must earn non-negative payoffs in any equilibrium, since otherwise the observer must object after some history that ends with the committee accepting the proposal. Figure 7 is the simplest example of a game with an equilibrium in which the committee with dispersed information earns strictly positive payoffs. Its insight extends to more complex situations with richer type spaces where the players may need to exchange information in order to achieve their desired outcome.

In our cheap talk framework, the fully informed expert cannot commit not to use all her information. Under negative total slack, she ends up with zero payoffs as a consequence. In contrast, the committee can ignore some of its dispersed information. It can refrain from exchanging messages that the observer will observe, and view unfavorably, effectively providing it with commitment not available to the fully informed expert. This difference in effective commitment underlies Proposition 2.

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, 2010). The literature on long cheap talk (Forges, 1990; Amitai, 1996; Aumann and Hart, 2003; Krishna

types versions of the generalized model and beyond that can be made precise via appropriate limits.

and Morgan, 2004; Chen, Goltsman, Horner, and Pavlov, 2017) asks how cheap talk can expand the set of equilibrium outcomes. We ask a complementary question. Fixing the optimal joint decision rule for a common interest committee, we ask how alternative communication protocols that implement this fixed decision rule differ in the amount of information revealed publicly. Our focus on subversion identifies situations where the problem of communication under scrutiny can be solved costlessly, without distortions in decisions. Our results require two-sided private information and rely on unrestricted sequential communication.[21] The general question of identifying all equilibrium outcomes of communication under scrutiny, with or without common interest, remains open.

Krishna and Morgan (2001) and Battaglini (2002) consider communication by two experts with different preferences but the same information. Their focus is on the receiver's best equilibrium in which the experts perfectly reveal the state. This structure and focus is the opposite of our case. Our focus is on the optimal equilibrium from the perspective of two experts with different information but the same preferences who must also hide information from an observer with different preferences. Back and forth conversations achieve these goals, an insight new to the literature. 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 from the perspective of an uninformed principal. The idea of consensus in their paper is related to the notion of deniability used in this paper. Extended communication has no role in their setting since communication is not scrutinized by the principal.

Since players are communicating to each other and the observer, this paper is related to the literature on cheap talk with multiple audiences starting with Farrell and Gibbons (1989). And since each player has relevant information for the decision, it is related to the literature on communication to receivers with private information (e.g., Watson, 1996). When a player sends 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.

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. Our decision problem is identical in the case of state independent sender preferences but our questions are different. We focus on sender optimal outcomes, when information about the different aspects is dispersed among multiple senders, and the receiver cannot independently obtain evidence. To precisely identify the effect of dispersed information, we consider a benchmark of a fully informed sender who

[21] This differentiates our model also from that of 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.

knows all decision-relevant aspects and who must voluntarily reveal one or the other aspect on her own while persuading the receiver. We show that Hall's Marriage Theorem (1935) provides the conditions under which such a sender achieves her optimal outcome.

Since the players attain their ideal outcomes in a subversive conversation, it is optimal under both cheap talk and commitment (Kamenica and Gentzkow, 2011; Lipnowski and Ravid, 2020). Our subversive conversations also extend to verifiable message games where message spaces depend on types. The pooling of conflict points with agreement points that is a key feature of our subversive conversations is reminiscent of strategic argumentation by a single expert communicating with an uninformed receiver (Dziuda, 2011). Since in our model each player has private information, such pooling and separation must not only persuade the receiver but also share enough information to determine the committee's ideal decisions. Cheap talk games can sometimes lead to pooling of disjoint types in equilibrium (e.g., Golosov, Skreta, Tsyvinski and Wilson, 2014). In our model, such pooling conceals unfavorable information till more favorable information has been revealed.

A subversive conversation ensures that the observer agrees with the players, so decision making power is effectively held by the players, even when formal authority lies with the observer. Authority and delegation have been studied in the literature on organizations (see, e.g., Aghion and Tirole, 1997; Dessein, 2002; Alonso, Dessein, and Matouschek, 2008). We show that the ability to manage the process of communication may be enough for the players to gain effective authority. How the observer would like to constrain this process remains an open question.

Within the broad political economy literature on communication and process, Gradwohl and Feddersen (2018) ask similar questions as this paper (see also Wolinsky, 2002; Feddersen and Gradwohl, 2020). They 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. Exploiting the richer type spaces of our model, we identify conditions when the committee can subvert even under transparency, regardless of the level of conflict. In these cases, transparency has no effect on outcomes.

5 Conclusion

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 the amount of information they reveal publicly. We show how a back and forth conversation can create a sequence of contexts that allow more information to be revealed, and so 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.

Like any equilibrium, subversion needs customs and conventions that ensure the process will play out as intended. A custom of staying till the stipulated end of the meeting will reduce any incentive to make untimely revelations because of impatience. Acquiring information in the order needed for decision making will also be useful for the same reason. There are situations when subversion is impossible. And even when subversion is possible, the observer may want to design rules and systems that prevent it. The problem of communication under scrutiny has many open questions. In this paper we provide some initial answers.

A Appendix

A.1 Omitted proofs

Proof of Theorem 1: To prove part (i), we treat the cases $c < 1/2$ and $c \geq 1/2$ separately.

Case 1 ($c < 1/2$): For arbitrary $z \in [0, 1]$, let the message in round $t = 1, 2, ...,$ be $m_t = z$ if $z \in \left[z_t^L, z_t^H\right]$ and $m_t =$ "pass" otherwise, where $z = x$ if t is odd and $z = y$ if t is even. So, for odd t, player 1 reveals x perfectly when $x \in \left[z_t^L, z_t^H\right]$ and by "passing" reveals that x does not belong to $\left[z_t^L, z_t^H\right]$. If player 1 reveals x in round t, $m_t = x$, then player 2 takes the committee's ex-post optimal decision by sending message $m_{t+1} = A$ if $(x, y) \in \mathcal{A} \cup \mathcal{C}$, and R otherwise. If instead, player 1 passes in round t, then the conversation moves to the next round where it is player 2's turn to speak. Even rounds where player 2 speaks and $z = y$, follow mutatis mutandis.

To define z_t^L and z_t^H, first let $T = \lceil \log_2((2-2c)/(1-2c)) \rceil$. Since $c > 0$, we must have $T \geq 2$. Set $z_0^L = 0$, and $z_t^H = 1 - z_{T-1}^L$ for all t. For $t = 1, ..., T-2$, let $z_t^L = 1/2 - (2^t - 1)(1 - 2c)/2$. For $t = T-1$ let $z_{T-1}^L = \max\{1/2 - \left(2^{T-1} - 1\right)(1 - 2c)/2, z_{T-2}^L/2\}$. Moreover, for $t = T$, let $z_T^L = z_{T-1}^L$. We are now left with 4 squares like in the top picture in Round 3 of figure 3 (round T insures they are squares instead of rectangles as shown in the figure). The game can be finished in much the same way. The player who speaks in round $T + 1$ coarsely reveals if the states are high or low and the other player can then take a decision n at time $T + 2$.

It remains to check that the deniability constraint is satisfied. First consider $t < T - 1$ and suppose without loss of generality that x has been revealed in round t (i.e., t is odd). Then given history m^{t-1} of t passes followed by the revealed value of x in round t and subsequent decision $m_{t+1} = A$, for all $x \in [z_t^L, z_t^H]$, we have that $\Pr\left[x \geq y | m^t, m_{t+1} = A\right] = \Pr\left[x \geq y | x, y \notin \left(z_{t-1}^L, z_{t-1}^H\right), x, x - y + c \geq 0\right] = x/(x + c - \left(z_{t-1}^H - z_{t-1}^L\right))$. This probability is weakly greater than $z_t^L/(z_t^L + c - \left(z_{t-1}^H - z_{t-1}^L\right)) = 1/2$. An identical argument applies for the case where y is revealed and player 1 takes the decision. This establishes (DC) for all $t < T - 1$. A similar argument also applies for $t \geq T - 1$. This completes the proof of Case 1.

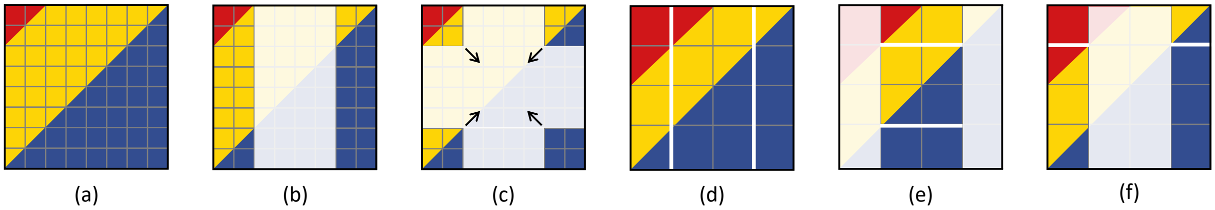

Figure 8: Subversive conversation with high conflict, $c \geq 1/2$

Case 2 ($c \geq 1/2$): Figure 8(a) depicts initial state space for $c > 1/2$. At $t = 1$, player 1 sends $m_1 = A$ if $x \in [1-c, c]$, without revealing the exact value of x, as shown in panel (b). Conditioning on $m_1 = A$, $\mathcal{C}$ is equal in measure $\mathcal{A}$, and so (DC) holds. If $x \notin [1-c, c]$ then player 1 passes at $t = 1$. Player 2 then sends $m_2 = A$ if $y \in [1-c, c]$, or $m_2 =$ "pass" otherwise. As in $t = 1$, this satisfies (DC) and implements the committee's ideal decision. Panel (c) depicts the residual state space after both players pass in the first two rounds, removing the states $[1-c, c] \times [0, 1]$ and $[0, 1] \times [1-c, c]$. This residual space is strategically identical to the state space where we paste the four remaining squares together, i.e., the case where $c = 1/2$.

To complete the proof we show a subversive conversation for $c = 1/2$. Player 1 sends $m_1 = [1/4, 3/4]$ if $x \in [1/4, 3/4]$ or $m_1 = \neg[1/4, 3/4]$ otherwise, as depicted in panel (d). If $x \in [1/4, 3/4]$ the conversation moves to panel (e). Here, if $y \in [1/4, 3/4]$, player 2 sends $m_2 = A$. Otherwise player 2 passes and player 1 perfectly reveals his signal $x \in [1/4, 3/4]$ at $t = 3$. Subsequently, player 2 sends $m_4 = A$ if $(x, y) \in \mathcal{A} \cup \mathcal{C}$, and $m_4 = R$ otherwise. Conditioning on: the value of $x \in [1/4, 3/4]$; $y \notin [1/4, 3/4]$; and $(x, y) \in \mathcal{A} \cup \mathcal{C}$, the measure of the agreement zone is always at least the measure of the zone of conflict, and so (DC) holds.

Panel (f) shows the state space after player 1 revealed $x \notin [1/4, 3/4]$ at $t = 1$. In this case, player 2 sends $m_2 = y$ if $y \leq 3/4$ or else says "pass". When y is revealed and player 1 proposes acceptance, (DC) is satisfied as can be seen from the figure. If $y > 3/4$ and player 2 passed, then player 1 makes the committee's optimal decision. This completes the proof of part (i).

To prove part (ii), assume by way of contradiction that $c < 1$ and that there exists a two round subversive conversation. Note that player 1 cannot perfectly reveal $x \in [0, c)$ and satisfy the deniability constraint (DC). However, the player 2 cannot take the first-best decision in round 2 if she does not know the exact value of $x \in [0, \min\{c, 1-c\})$, which is a positive measure set when $0 < c < 1$, and thus we have a contradiction. ■

Proof of Theorem 2(ii): Assume by way of contradiction that there exists a three-round subversive conversation that works for all $c \in (0, 1)$. While we focus on pure strategies in the paper, we allow for mixed strategies in this impossibility proof. If player 1 is to make her ideal decision in round 3 he must know whether x bigger than $1 - c$ for all $c \in (0, 1)$; this is only possible if player 2 perfectly reveals $y = 1$ in round 2, since a universal conversation cannot depend on c except when making a decision.

Consider the following set of histories, H: player 1 in round 1 sends some message m_1 which does not perfectly reveal the x-dimension of the state and instead pools, player 2 in round 2 then sends message $y = 1$ and player 1 sends $m_3 = A$.

More formally, let $x' = \min\{c, 1/2\}$ and observe that player 1 cannot perfectly reveal $x \in [0, x')$ in round 1 and satisfy the deniability constraint (DC); thus these x values must be pooled in round 1. Let $\sigma(x, m^{t-1})$ denote the strategy of player 1 whose type is x, following history m^{t-1}. Consider now the set of histories H induced by types $x \in [0, x')$ and $y = 1$, i.e., $M_1 := \{m_1 \in \text{supp}(\sigma(x, m^0)) : x \in [0, x')\}$, followed by player 2 revealing $y = 1$ in round 2 and player 1 saying "Accept" in round 3. That is to say let $H := \left\{m^t : m_1 \in M_1, m_2 = 1, m_3 = A\right\}$. The deniability constraint (DC) integrated over this set H must be satisfied by an admissible conversation $\int_{m^t \in H} \left\{\Pr[\mathcal{A}|m^t] - \Pr[\mathcal{C}|m^t]\right\} \; d\nu\left(m^t\right) \geq 0$. Note that the integral here is with respect to the measure ν which is induced by the strategies and prior (could be a mixed strategy). However, given $y = 1$, only the singleton point $(1, 1) \in \mathcal{A}$ and thus $\int_{m^t \in H} \Pr[\mathcal{A}|m^t \in H] = 0$. For any $c > 1/2$, we have that $\int_{m^t \in H} \Pr[\mathcal{C}|m^t \in H] \geq \Pr[x \in [1 - c, 1/2)] = c - 1/2 > 0$ and thus the deniability constraint fails, which is a contradiction. ■

A.2 Subversion with non-iid priors

In this section we extend the construction of Theorem 2 to the case of non-iid priors for x and y. Consider first the case where x and y are distributed independently but not identically. Let F_x be the CDF of x and F_y the CDF of y. Notice that Lemma 1 applies and so we can restrict attention to the test case depicted in Figure 4, where $\mathcal{R}$ is empty, $\mathcal{A} = \mathcal{L}$ and so $\mathcal{C} = \mathcal{U}$. If F_x first order stochastically dominates F_y, then the transformation to quantiles employed in Section 2.3, yields a new game with $\mathcal{A} \supseteq \mathcal{L}$, $\mathcal{C} \subseteq \mathcal{U}$ and $\mathcal{R}$ empty. Applying the subset lemma again, it suffices to consider the case $\mathcal{A} = \mathcal{L}$ and $\mathcal{C} = \mathcal{U}$. Since quantiles are uniformly distributed, we can now employ the same conversation depicted in Figure 4, using the quantiles instead of the random variables x and y at every stage. The same conversation will be subversive for every configuration of preferences of the generalized model. We have the following result.

Proposition 3. *If x and y are independently distributed with CDFs F_x and F_y, where F_x first order stochastically dominates F_y, then the four round conversation of Figure 4 is subversive for all preference configurations of the generalized model.*

Consider next the case where x and y may not be statistically independent. We assume 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. Suppose for the prior $g(x, y)$, the conversation depicted in Figure 4 is subversive for the test case with $\mathcal{A} = \mathcal{L}$ and $\mathcal{C} = \mathcal{U}$. The following result identifies conditions for the same conversation to be subversive when the prior is instead given by another joint density $h(x, y)$ and associated conditionals $h(x|y)$ and $h(y|x)$.

Proposition 4. *If the conversation of Figure 4 is subversive when $\mathcal{A} = \mathcal{L}$ and $\mathcal{C} = \mathcal{U}$, when the prior of x and y is given by the joint density $g(x, y)$, it is also subversive for every configuration of preferences of the generalized model when the prior is given by the joint density $h(x, y)$ that satisfies (a) $\frac{h(x|y)}{g(x|y)}$ is non-decreasing in x, for all $y \in [0, 1]$, and (b) $\frac{g(y|x)}{h(y|x)}$ is non-decreasing in y, for all $x \in [0, 1]$.*

The conditions above say that h assigns higher weight to higher values of x and lower values of y, relative to g, in the sense of likelihood ratios. Since x is a benefit and y is a cost, this raises total slack.[22] We show below that under the assumed conditions, the same conversation will be subversive in the test case where $\mathcal{A} = \mathcal{L}$ and $\mathcal{C} = \mathcal{U}$, when the prior is h. Lemma 1 will then allow us to conclude that the same conversation is subversive for every configuration of preferences of the generalized model.

Suppose player 1 has revealed $x \in \{a, 1 - a\}$ for some $a \leq 1/2$ in round 1. Consider first the case where $y \in (a, 1 - a)$ and so player 2 has revealed y in round 2. We must have $a < 1/2$ since otherwise $(a, 1 - a)$ is empty. Note that $(1 - a, y) \in \mathcal{A}$ and $(a, y) \in \mathcal{C}$. Since (DC) is satisfied when the prior is g we must have $\frac{g(1-a,y)}{g(a,y)} \geq 1$. To show that (DC) will be satisfied when instead the prior is h, it suffices to show $\frac{h(1-a,y)}{h(a,y)} \geq \frac{g(1-a,y)}{g(a,y)}$. Employing Bayes Rule and rearranging, this is equivalent to showing $\frac{h(1-a|y)}{g(1-a|y)} \geq \frac{h(a|y)}{g(a|y)}$. Since $1 - a > a$, this is implied by the first assumed condition on h.

Consider next the case where player 2 has revealed $y \notin (a, 1-a)$ in round 2, following which player 1 reveals the exact value of $x \in \{a, 1 - a\}$ in round 3. We consider the case $x = a$, since the case $x = 1 - a$ is entirely analogous. Since (DC) is satisfied when the prior is g and player 2 accepts, we must have $\frac{G(a|x=a)}{1-G(1-a|x=a)} \geq 1$, where $G(y|x)$ is the CDF of y given x under g. To show that (DC) will also be satisfied when instead h is the prior, it suffices to show$\frac{H(a|x=a)}{1-H(1-a|x=a)} \geq \frac{G(a|x=a)}{1-G(1-a|x=a)}$, where $H(y|x)$ is the CDF of y given x under h. But this follows if $G(y|x = a)$ first order stochastically dominates $H(y|x = a)$, which is implied by the second assumed condition on h. This establishes Proposition 4.

A.3 Subversion with slack

In this section we amend the gradual protocol in Theorem 1 to construct subversive conversations that satisfy (DC) with slack. Since the slack is uniformly bounded away from zero along all possible path of play, the same conversation will be subversive for perturbations of preferences and priors in arbitrary directions. Proposition 5 covers the case $c < 1/2$ in the baseline model.

Proposition 5. *For any $c < \frac{1}{2}$, there exists a subversive conversation such that (DC) has at least $\varepsilon = (1/2 - c)/2$ slack.*

[22] The monotonicity of the likelihood ratios is not necessary for the result. The condition $h(x, y) \geq g(x, y)$ iff $x \geq y$ is an alternative sufficient condition on h that dispenses with monotonicity.

Proof. We will consider three separate cases, for $c \leq 1/4$, $c \in (1/4, 3/10]$ and $c > 3/10$. We provide full details of the first and a more terse proof of the other cases.

Case 1 ($c \leq 1/4$). Let $\varepsilon = \frac{1}{4} - \frac{1}{2}c$. At $t = 1$ player 1 perfectly reveals $x \in \left[x_1^L, x_1^H\right]$ else says "pass", where the cutoffs are defined as $x_1^L = c + \varepsilon$ and $x_1^H = 1 - c - \varepsilon$. Perfectly revealing x_1^L is where (DC) is most binding, but there is a c measure of conflict and $c + \varepsilon$ measure of agreement.

At $t = 2$ perfectly player 2 perfectly reveals $y \in \left[y_2^L, y_2^H\right]$ else passes, where the cutoffs are defined as $y_2^L = \max\left\{0, y_2^H - x_1^H + x_1^L\right\}$ and $y_2^H = x_1^L + c$. Since $x_1^H = 1 - c - \varepsilon$, there is an extra ε worth of slack in (DC) for any value that player 2 perfectly reveals.

At $t = 3$, player 1 says either $x \in \left[x_3^P, x_1^L\right]$, pooling all those states, or perfectly reveals $x \in (x_1^H, x_3^H]$, or says "pass". Two of the cutoffs, x_1^L and x_1^H, are defined above and $x_3^P = \max\left\{y_2^L - c, 0\right\} = \max\left\{c - \varepsilon, 0\right\}$, $x_3^H = 1 - c\mathbf{1}_{\left\{x_3^P > 0\right\}}$. To check there is sufficient slack in (DC), observe that $x_1^H - x_1^L \geq y_2^H - y_2^L$, so perfectly revealing x_1^H (and anything to the right of it) entails at least an ε of slack in (DC). Consider the pooling revelation $\left[x_3^P, x_1^L\right]$. To see that this will satisfy (DC) with slack, consider first the case where $x_3^P = 0$, (i.e., $y_2^L \leq c$ or $c \leq \frac{1}{6}$). In this case, $\left[x_3^P, x_1^L\right] = [0, c + \varepsilon]$ and player 2 will decide to accept in the case of $y < y_2^L \leq c$. This has at most a (Lebesgue 2-dimensional) measure $\frac{1}{2}c$ of conflict zone (yellow) and at least a (Lebesgue 2-dimensional) measure $\frac{1}{2}c + \varepsilon$ of agreement zone (red), so there is at least ε slack in that final decision by player 2. In the case where $x_3^P > 0$, we have that $x_3^P - c = \frac{1}{2}c - \frac{1}{4} \geq \varepsilon$ (for $c \leq \frac{1}{4}$) and so there is at least an ε of "extra" agreement zone.

In the case where $x_3^P > 0$, (i.e., $\frac{1}{6} < c \leq \frac{1}{4}$), the game continues with player 2 revealing perfectly $y \in [0, y_2^L)$ and says "pass" if the states are $(y_2^H, 1]$. Player 1 then takes the appropriate decision. Since $x_3^H = 1 - x_3^P - \varepsilon$ in this instance, (DC) is satisfied when $y \in [0, y_2^L)$ is perfectly revealed. Let's now check the remaining states $(y_2^H, 1] \times (x_3^H, 1]$, i.e., the top-right corner. In particular, we have that $x_3^H = 1 - c$, so that player 1 can simply take a decision. At x_3^H there is an additional $1 - y_2^H - c = 1 - 2c - \varepsilon - c = 1 - 3c - \varepsilon \geq \varepsilon$ (for $c \leq \frac{1}{4}$) of slack in (DC). As such, the measure of the agreement zone is at least ε larger than the measure of the disagreement zone on that final decision by player 1.

Case 2 ($c \in (1/4, 3/10]$). At $t = 1$ player 1 perfectly reveals $x \in \left[x_1^L, x_1^H\right]$ else says "pass", where $x_1^L = c + \varepsilon$ and $x_1^H = \frac{1}{2}$. At $t = 2$ perfectly player 2 perfectly reveals $y \in \left[y_2^L, y_2^H\right]$, where $y_2^L = x_1^L$ and $y_2^H = x_1^L + c$.

At $t = 3$, player 1 says either $x \in \left[x_3^P, x_1^L\right]$, pooling all those states, or perfectly reveals $x \in (x_1^H, x_3^H]$, or says "pass". The right cutoff of the pooling region, x_1^L is defined above, and $x_3^P = \varepsilon$. Consider the pooling revelation $\left[x_3^P, x_1^L\right] = [\varepsilon, c + \varepsilon]$. The second player will accept if $y < y_2^L$. To check there is sufficient slack in (DC) following this decision, note that we have a measure of $\frac{1}{2}c^2$ of conflict, but that the overall region has measure $c(c + \varepsilon)$, so that on average there is $c\varepsilon$ extra slack and so (DC) is satisfied. Consider next the perfect revelation of $x \in (x_1^H, x_3^H] = (1/2, 1 - 2\varepsilon]$. The (DC) is most binding at $x = y_2^H = 1 - 3\varepsilon$, since the y

cut has not removed any conflict region and has removed only agreement zone. Following this revelation, player 2 will accept if $y \leq y_2^H + c$. Following this decision, there is a measure c of conflict and a measure $y_2^H - c = 2c + \varepsilon - c = c + \varepsilon$ of agreement, which indeed has an extra ε slack in (DC).

If player 1 passes, in the next period player 2 perfectly reveals y if $y \in [0, y_2^L)$ or passes if the states are $(y_2^H, 1]$. Player 1 then takes the appropriate decision. Since $x_3^H = 1 - 2\varepsilon$ and $x_3^P = \varepsilon$, there is an ε of slack in (DC) following a revelation of any $y \in [0, y_2^L)$. To check (DC) on the remaining states (where player 1 will accept, i.e., $(y_2^H, 1] \times (x_3^H, 1])$, note that $x_3^H = 1 - 2\varepsilon$ and $y_2^H = 1 - 3\varepsilon$, so there is always an extra ε of agreement zone for low y.

Case 3 ($c \in (3/10, 1/2)$). For $t \in [1, t_e] \cap \mathbb{N}$ let $z_t^L = \frac{1}{2}\left(1 + \varepsilon + (-1)^t \varepsilon - 2^t \varepsilon\right)$ and $z_t^H = z_t^L + \left(2^t - 1\right)\varepsilon$, where $t_e = \left\lfloor \log_2\left(\frac{1}{2\varepsilon} - 1\right)\right\rfloor$. The first few $\left(z_t^L, z_t^H\right)$ pairs are $(c + \varepsilon, 1/2)$, $(c + \varepsilon, 1/2 + 2\varepsilon)$, $(1/2 - 4\varepsilon, 1/2 + 3\varepsilon)$, $(1/2 - 7\varepsilon, 1/2 + 8\varepsilon)$, $(1/2 - 16\varepsilon, 1/2 + 15\varepsilon)$. At each t, the player to speak perfectly reveals $\left[z_t^L, z_t^H\right]$ or says pass, where $z = x, y$ depending on whether the period is odd or even. It is easy to check that this indeed satisfies (DC) with ε slack.

If t_e is even, player 1 at time $t_e + 1$ will perfectly reveal remaining x values between $x_{t_e+1}^L = x_{t_e-1}^L - \gamma$ and $x_{t_e+1}^H = x_{t_e-1}^H + \gamma$, where $\gamma = \min\left\{1 - y_{t_e}^H + \varepsilon - x_{t_e-1}^H, x_{t_e-1}^L - \frac{1}{2}\left(y_{t_e}^L + \varepsilon\right)\right\}$. Following this player 2 perfectly reveals states in $[0, y_{t_e}^L)$ or says pass. In both cases, player 1 finishes the game by taking a decision.

If t_e is odd, player 2 at time $t_e + 1$ perfectly reveals y between $y_{t_e+1}^L = y_{t_e-1}^L - \delta$ and $y_{t_e+1}^H = y_{t_e-1}^H + \delta$, where $\delta = \left(x_{t_e}^L + c - y_{t_e-1}^H + \varepsilon\right)/2$. At time $t_e + 2$ player 1 perfectly reveals x in $x_{t_e+2}^L = x_{t_e}^L - \gamma$ and $x_{t_e+}^H = x_{t_e}^H + \gamma$, where $\gamma = 1 - y_{t_e}^H + \varepsilon - x_{t_e-1}^H$. Following this player 2 perfectly reveals states in $[0, y_{t_e}^L)$ or she passes. In either case, player 1 finishes the game by taking her ideal decision. ■

Proposition 5 shows that subversive conversations are robust to model perturbations, when the conflict between the committee and the observer is small enough. To obtain the same robustness result for larger conflicts, one needs to amend the model. This is easiest to see for the case $c = 1$ which has zero total slack. One approach to adding slack is to assume $\mathcal{A}$ contains the set $\{(x, y) \in \mathcal{S} \mid y \leq x + \epsilon\}$ for some $\epsilon > 0$. One can then use constructions similar to those employed for Theorem 2 that satisfy (DC) with slack along every possible path of play. Details are available upon request.

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. 2010. "Persuasion by Cheap Talk," *American Economic Review*, 100(5): 2361–2382.

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

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

[9] Crawford, Vincent P. and Joel Sobel, 1982. "Strategic Information Transmission," *Econometrica*, 1431–1451.

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

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

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

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

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

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

[16] Glazer, Jacob and Ariel Rubinstein. 2004. "On Optimal Rules of Persuasion," *Econometrica*, 72(6): 1715–36.

[17] Golosov, Mikhail, Vasiliki Skreta, Aleh Tsyvinski, and Andrea Wilson. 2014. "Dynamic strategic information transmission." *Journal of Economic Theory*, 151:304–341.

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

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

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

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

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

[23] 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.

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

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

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

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

[28] Silberman, Alan H and Leah R. Bruno. 2007. "Sunk By Your Own Torpedoes! How Emails and Memos Can Lead to Antitrust and Other Litigation Issues," presentation, Dentons.com.

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

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