Subversive Conversations[*]

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

October 27, 2020

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 optimal 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 experts, cheap talk, subversion, deniability, 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, 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), and University of Bath (2019). For helpful comments, we also thank David Austen-Smith, Sandeep Baliga, Eddie Dekel, Wioletta Dziuda, Georgy Egorov, Jeff Ely, Daniel Garrett, Parikshit Ghosh, Faruk Gul, Jason Hartline, Marilyn Pease, Nicola Persico, Doron Ravid, Ludovic Renou, Patrick Rey, Alvaro Sandroni, Joel Sobel, Lars Stole, Jean Tirole and Bilge Yilmaz.

[†] 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

The purpose of minutes is not to record events, it is to protect people."
— Sir Humphrey Appleby, *Yes Prime Minister, Official Secrets*, BBC, 1987.

We study communication under scrutiny. To decide on the best plan of action, people need to share information. But at the same time they may need to conceal information from other parties who have different interests and may access their communication. Scientists have to exchange knowledge to understand an issue, but revealing too much to a skeptical public might undermine their policy recommendations. A committee must share information to determine whether to hire a candidate, but may fear controversy if its deliberations are exposed to the public. This need to both share and conceal can arise whenever information is decentralized, such as firms proposing a merger to regulators, activists organizing under state surveillance, military leaders recommending an intervention to political leaders, or legislators deliberating on a law before an election.

If conversations can be kept truly private, people with the same interests can share all their information with each other while concealing it from other parties with different interests. But in reality the chance of public exposure always remains—emails can be hacked, codes can be broken, memos can be subpoenaed, whistleblowers can go public.[1] Even for traditionally private attorney-client communication, legal counsel are now advised to "assume every word will somehow get published".[2] Without confidence in the secrecy of their communications, how can people share enough information to coordinate on the right decision?

We show how back-and-forth communication – a conversation – between two players can reconcile these needs to *coordinate* and *conceal* even if the entire conversation is (made) public.[3] As the conversation progresses, the players share more information based on the context established by previous statements, while withholding information that the other side does not need to know or that is best revealed later in the conversation. Initially the players don't reveal very good or very bad information but rather pool it together while waiting for news from the other side. Over several rounds, this process shares enough information for the players to determine

[1] Recent exposures include the leak of climategate emails by a server breech, the subpoena of private documents in the VW dieselgate scandal and in the Purdue Pharma opiod case, the leak of DNC emails attained by password phishing, preemptive self-reporting in antitrust cases, and even the release of inherited emails from a deceased expert in the Census citizenship and Congressional redistricting controversies. As Pep Guardiola of Manchester City Football Club put it after a hacking scandal, "Today there are no secrets any more... Everyone knows it." Our focus is on how careful conversations can conceal enough information even if the conversation becomes public.

[2] Quote from 2017 presentation "Sunk By Your Own Torpedoes! How Emails and Memos Can Lead to Antitrust and Other Litigation Issues" by Alan H. Silberman and Leah R. Bruno of Dentons, the world's largest law firm. Available at Dentons.com.

[3] The analysis applies equally if player communications might be leaked or if they are required to be public, such as sunshine laws for government proceedings, FOIA rules for government communications, and accounting rules for internal workflows.

their own preferred action, while also hiding enough information to prevent any objections to the players' determination.[4]

We analyze this problem in a cheap talk setting where two players ("experts") with private information communicate in order to reach a decision while trying to ensure that a third party (an "observer") with different interests does not intervene or otherwise penalize them. The observer, who could represent a supervisor, a government official, or the public more generally, listens and fully comprehends the experts' communication strategy. Under reasonable conditions we show the existence of equilibria that are not just persuasive, in that on average the experts gain relative to no communication, but that are also "subversive", i.e., they achieve the experts' optimal outcome in all states, including states where the observer prefers a different action. The experts subvert the observer's agenda and implement their own. Hence the experts can do just as well through a public conversation as through completely secure private communication.

To see how a conversation can be subversive, suppose two managers (the experts) are evaluating whether to accept or reject a new mining operation. The project has both environmental costs and economic benefits, but 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 they would like to undertake the project as long as 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 sides support good projects and oppose bad projects, but preferences for mediocre projects are in conflict.[5]

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 upper branch of the tree in Figure 1. In the second round manager 2 then recommends rejection if the cost is 1, in which case everyone opposes the project; or recommends acceptance 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

[4] Our problem of communication by players with identical preferences but different information who face scrutiny by a third party is novel in the cheap talk literature started by Crawford and Sobel (1982). For instance, Battaglini (2001) considers the antipodal problem of multiple experts with different preferences but identical information, and focuses on the receiver-optimal rather than sender-optimal outcome. The related literature is discussed in more detail in Section 5.

[5] The game is not zero-sum so there is no issue of deception or strategic feints (Crawford, 2003). Cooperation by the players is not by itself damaging to the observer as in Baccara and Bar-Isaac (2008).

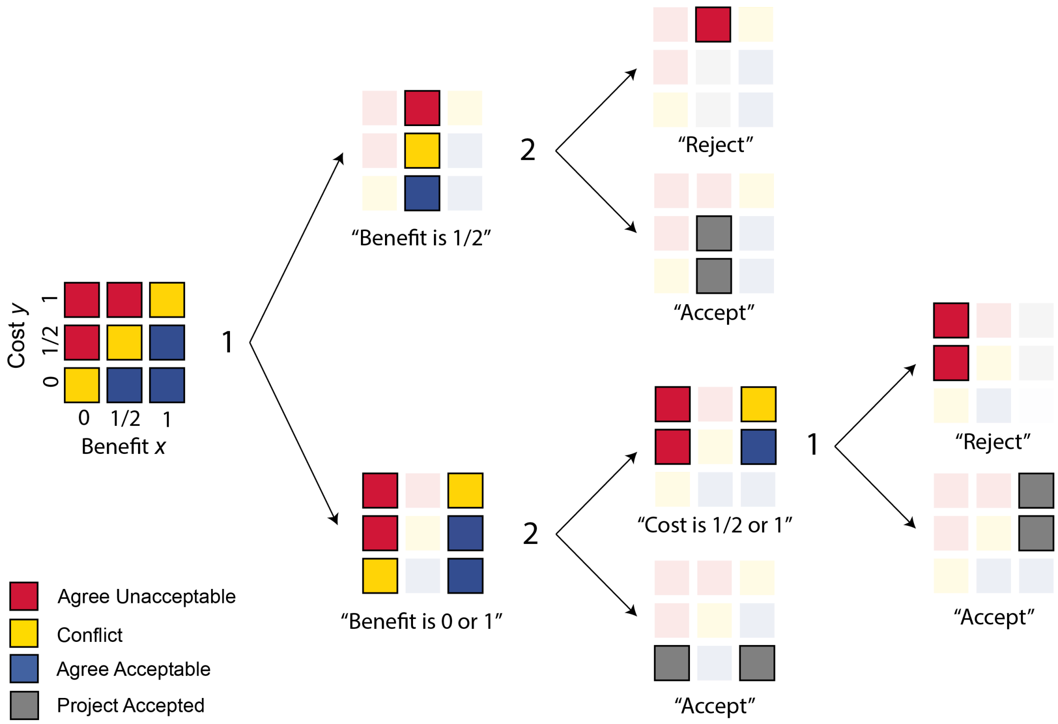


Figure 1: Conversation tree

public that they have no reason to oppose it.

If the benefit is either 0 or 1 then in the first round manager 1 hides this potentially good or bad news as seen in the first lower branch of the tree. Such pooling effectively passes the conversation back to the other manager for further discussion while keeping the public from intervening.[6] In the second round manager 2 then recommends acceptance 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 does not object. If instead the cost is 1/2 or 1 then manager 2 pools this news since revealing high cost will ensure objection by the public. Having learned that the cost is not 0, in round 3 manager 1 now recommends rejection if the benefit is 0 and acceptance 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, and only manager 2 knows which, so the public does not object to the project.

By the end of the conversation the managers have pooled all states where the project is mediocre or good and they want to accept the project, while identifying all states where the project is bad and 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. Despite

[6] In our cheap talk framework messages only have meaning in equilibrium so it is equivalent if the manager says something less literal to pass the conversation back like "What do you think?" or "Not sure yet".

the conflict of interest and the fact that the public has access to their communication and fully understands what is going on, the managers are able to use this communication protocol to subvert the public and get their ideal outcome in every state of the world.[7] As long as they follow the subversive communication protocol in the above example, 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.

We examine problems like in Figure 1 but with richer information structures and more exacting coordination demands. We show that subversive conversations exist in surprisingly many cases, for a wide variety of preferences and priors, independently of the precise details of the environment and even in the presence of uncertainty about these details. When subversive conversations exist, the fact that the experts hold decentralized information and converse in public imposes no cost on them. They do equally as well as they would if they could guarantee secure and private information exchange and had full control over their decisions.

We start with a baseline model that is similar to the above example except there is a uniform distribution of states in the unit box and the biased experts have linear preferences. In Theorem 1 we show constructively that, for any degree of conflict between the experts and the observer, there exists a subversive conversation. The players reveal more information in each round in an intuitive manner that progressively eases the constraints on further information disclosure. Because of the need to delay disclosure of decisive information by one expert until more information has been revealed by the other expert, we show that this back and forth aspect of communication is necessary in that, for any bias, a subversive conversation must take at least three rounds. In Theorem 2 we then show constructively the existence of a universal subversive conversation that does not depend on the exact degree of conflict. The construction effectively collapses the information exchange from the more gradual conversation into a four-round conversation. We also show that this is the fastest possible universal conversation.

In the universal conversation of our baseline model, the exact same conversation remains subversive for any degree of conflict. Looking at this issue more generally, in Theorem 3 we show that if a subversive conversation exists for a game with one set of conflict states and agreement states, then the same conversation remains subversive in any game where the conflict and agreement states are respective subsets and supersets of the original game. Applied to our baseline model, we find that this result implies robustness of the subversive conversation to nonlinear preferences, to uncertainty over these preferences including a lack of common knowledge, and to non-uniform priors. We also use the theorem to show that subversion is robust to allowing the direction of the expert bias to be state-dependent so that the experts are sometimes biased

[7] This conversation, and the symmetric one where manager 2 speaks first, are the only subversive conversations for this example. As in any communication game with pooling, better types might want to separate out from the equilibrium by revealing their type. But unlike in most communication games, in our environment the managers do as well as they possibly can by following the equilibrium strategy of pooling, so there is no potential gain from such separation.

toward the status quo and sometimes toward the proposal.

In variations of the baseline model, we use the same approach to show that subversion is robust to situations where the observer leans toward the status quo, either ex ante because of unfavorable priors or ex post because of bias in his preferences, and to the observer having expectation preferences rather than the probability preferences we focus on. We also discuss how the model extends to situations where the observer puts different weights on the loss from a mediocre proposal versus the gain from a good proposal, and to situations with conflicts of interest between the experts.

Finally, we investigate the general problem of pooling mediocre and good states while separating bad states that is necessary for subversion. We ask when the experts cannot subvert. One kind of failure of subversion occurs when the experts are unable to subvert even if they could confer in private. We show that in this case public communication generally benefits the experts. A second kind of failure of subversion occurs when the experts can still get their ideal outcome if they can privately communicate, but the need to both exchange information with each other and conceal information from outsiders makes subversion under public communication impossible. We show that this type of failure of subversion is related to Hall's marriage theorem (1935). We use Hall's theorem to provide examples where subversion is not possible because the experts and the observer do not agree on the extent to which the different pieces of information held by the two experts are relevant for the decision.

The robust existence of subversive conversations has implications for communication design within 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 to eliminate bias in hiring it may be necessary to go beyond transparency and consider how restrictions on the form and flow of information can minimize the ability of the committee to coordinate on a biased decision.

Instead of different levels of the organization having different preferences, they may share the same preferences but have a conflict with the public or an outside authority. In this case communication design can be used instead by an organization to maintain "plausible deniability" for leadership even if private deliberations become public. Executives may want to implement subversive communication protocols in order to avoid learning compromising information.[8] For instance, in the introductory example, if the managers follow the protocol and report their recommendation to the executive then any ex post investigation of their deliberations will not find that the executive had sufficient grounds to intervene and stop the project.

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

Our results also offer insight into the more general role of experts in society. Divergent interests and beliefs between experts and the broader public have become a central concern for many policy issues, from free trade to Brexit to climate change. This paper shows that the ability of experts to exchange information with each other, and their ability to successfully persuade the public, need not be affected by this divergence though the form of information revelation becomes less straightforward. Even if sunshine laws and other transparency regulations force deliberations out from behind closed doors, careful experts can still subvert the public interest, e.g., persuade voters to go to war, elect a candidate, confirm a judge, etc.[9] Conversely, to the extent that experts have different policy preferences than the public due to greater knowledge and understanding, our results show that fact-based decision making using expert knowledge is less affected by public interference than might be expected.

This paper also offers new insight into the role of secure versus insecure communication in commercial, governmental, and national security contexts. If the players could encrypt their messages they could communicate their exact information to each other privately and coordinate on their optimal action.[10] In our model encryption is unnecessary for subversion since the concealment property of conversations allows the players to successfully coordinate even if a third party is listening in on the conversation in full awareness of the intended meaning of messages. Hence monitoring the unencrypted communication of suspects in antitrust or other illicit activities need not be sufficient to prevent successful collusion by careful conspirators. And the spread of the surveillance state need not always prevent successful subversion by the citizens. When a player pools together different states with a message in the model, there is no additional information inferred by the other player that cannot be inferred by the third party, so messages are not encrypted (Shannon, 1949). Instead, such pooling allows the other player to better use their own information to make a decision or to reveal more information. Limited concealment through a conversation is sufficient because, unlike in the cryptography literature, there is some commonality of interests between the players and the third party.

The rest of the paper is organized as follows. In Section 2 we set up our baseline model and show that decentralized experts can subvert whenever a centralized expert can. We also extend the existence result to alternative specifications of preferences, priors and other details of the environment. In Section 3, we provide extensions and further robustness properties of our baseline results. Section 4 considers cases where subversion is impossible. Section 5 reviews the literature, Section 6 contains our concluding remarks, while Section 7 contains all proofs that are not contained in the main text.

[9] Most models of transparency have focused on whether receivers know the sender's bias (e.g., Morgan and Stocken, 2003; Chakraborty and Harbaugh, 2010). We assume the experts are known to be biased, and consider whether their communications are transparent (public) or not.

[10] In practice even "end-to-end" encryption fails if communication devices themselves can be corrupted (Padlipsky, Snow, and Karger, 1978). Moreover, use of encoded communications is often prohibited by public policy or by internal corporate policy.

2 A Model of Subversion

2.1 Players, Preferences, and Information

A committee is composed of two players, 1 and 2. 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 have common interests and decide to either accept or reject a proposal. Their payoff from the status quo 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. Absent other constraints and because of common interests, the problem is straightforward—the players can reveal their signals to each other and choose the action they both like.

However there is a constraint on the committee. A third party, referred to as the observer, has different preferences from the committee and is privy to the communication. 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 is -1 otherwise. The observer is uninformed and does not know the realization of x or y. But he listens to the communication between committee members and fully understands the intended meaning of everything he hears. The task before the committee is to exchange enough information in order to implement the committee's own desired decision rule without provoking objections or intervention by the observer at any time. It does not matter for our results if the committee takes the decision and the observer only has access to committee deliberations ex post, or if the committee deliberates in the presence of the observer who takes the actual decision.

We interpret x and y as the social benefit and social cost of the proposal relative to the status quo. The observer prefers the proposal whenever the social benefit x is at least as high as the social cost y. Relative to the observer, the committee is biased in favor of the proposal. The parameter c represents the conflict between the committee and the observer. Thus, when $x + c \geq y$ but $x < y$ the observer prefers the status quo while the committee prefers the proposal. Otherwise, the observer and the committee agree on the optimal decision. Given the specification of observer preferences, the uninformed observer is willing to accept the proposal if it is more likely than not that the social benefit of the proposal exceeds the social cost, conditional on the information the observer infers from the communication between the players. We define the communication game between the committee members and related notions more formally now.

Let $\mathcal{R} = \{(x,y) \in [0,1]^2 | u(x,y) < 0\}$ be the set of states where the observer and the committee both agree the proposal should be rejected; and let $\mathcal{B} = \{(x,y) \in [0,1]^2 | v(x,y) > 0\}$ be the set of states where they both agree the proposal should be accepted. Let $\mathcal{Y} = [0,1]^2 - \mathcal{R} - \mathcal{B}$ denote the zone of conflict where the committee prefers to accept the proposal but the observer does not. The assumption $c > 0$ ensures that $\mathcal{Y}$ is non-empty and the players need to conceal information from the observer. The assumption $c < 1$ ensures that $\mathcal{R}$ is non-empty and the

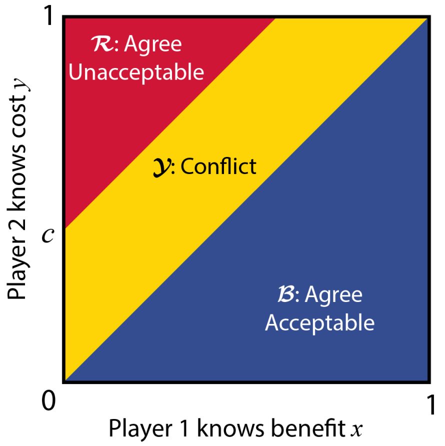

1
$\mathcal{R}$: Agree Unacceptable
$\mathcal{Y}$: Conflict
$\mathcal{B}$: Agree Acceptable
Player 2 knows cost y
c
0
1
Player 1 knows benefit x

Figure 2: Conflict and agreement sets

players need to exchange information in order to coordinate on their optimal decision.[11]

The communication between the players takes the form of *polite talk* in the sense of Aumann and Hart (2003). There are sequential rounds of cheap talk, indexed by $t = 1, 2, ...$, where player 1 (player 2) sends a message in odd (even) rounds t, while the other player listens. A (pure) strategy for each player specifies a message m_t for each possible history of messages m^{t-1} and the expert's own signal. Let M be the set of possible messages.[12]

We refer to a pair of strategies, one for each player, simply as a *conversation.* A conversation terminates in round t with a recommendation by a player to accept or reject the proposal. Denote by $m_t = A$ a recommendation to accept the proposal and by $m_t = R$ a recommendation to reject it and suppose the committee always implements the recommended decision. The equilibrium notion is perfect Bayesian equilibrium.[13]

The observer's preferred decision rule is the proposal be accepted if $(x, y) \in \mathcal{B}$ and rejected otherwise. But our focus is on *subversive conversations*, in which the committee subverts the observer's agenda and implements its own ideal decisions in all states (x, y), without the observer objecting. A subversive conversation (if it exists) leads to accepting the proposal if $(x, y) \in \mathcal{B} \cup \mathcal{Y}$ and rejecting it otherwise. It is automatically an equilibrium conversation because no player has an incentive to deviate from a conversation that implements her ideal decision rule. For this reason, we refer to an equilibrium subversive conversation simply as a subversive conversation.

[11] In all our figures, the sets $\mathcal{R}$, $\mathcal{B}$ and $\mathcal{Y}$ are depicted in red, blue and yellow, respectively. We will call the line $y = x + c$ that forms the boundary between $\mathcal{R}$ and $\mathcal{Y}$ the committee's decision line, and the line $y = x$ that forms the boundary between $\mathcal{Y}$ and $\mathcal{B}$ the observer's decision line.

[12] As is standard in cheap talk games, messages have no intrinsic cost or benefit. We assume that the message space M is rich enough so that information transmission is constrained only by incentives and not by the availability of messages.

[13] We can (and will) ignore off the path of play messages because any time a cheap talk equilibrium exists where some messages are not used on the path of play, an outcome equivalent equilibrium exists where all messages are used on the path of play (and vice versa).

There is no need for commitment from the players to implement a subversive conversation.[14]

Subversion requires that the experts share enough information with each other in order to be able to implement their optimal decision rule. In particular, the *coordination constraint* requires that after observed history of messages m^{t-1},

$$\begin{aligned} &\Pr[\mathcal{B} \cup \mathcal{Y} | m^{t-1}, m_t = A] = 1, \\ &\Pr[\mathcal{R} | m^{t-1}, m_t = R] = 1, \end{aligned} \tag{CC}$$

where $\Pr[Z|m^{t-1}, m_t]$ denotes the probability that the realized state (x, y) is in the set Z, given a history m^t and a recommendation to accept or reject, $m_t \in \{A, R\}$. Notice from the border between $\mathcal{R}$ and $\mathcal{Y}$ in Figure 2, that the cutoff value of y, where the committee is indifferent between accepting and rejecting the proposal, may depend on x. When it does, at least one committee member needs to publicly reveal the exact value of her signal in order to satisfy the coordination constraint (CC).

Subversion also requires that the conversation obscures enough information that the observer does not object to a committee decision for acceptance, even when in fact $(x, y) \notin \mathcal{B}$. In particular, the *deniability constraint* requires that in round t with a recommendation for acceptance $m_t = A$ after history m^{t-1} the observer also prefers accepting the proposal to the status quo, $\Pr[\mathcal{B}|m^{t-1}, m_t = A] \geq 1/2$. Since in a subversive conversation the experts never propose acceptance when $(x, y) \in \mathcal{R}$, this condition is equivalent to requiring that

$$\Pr[\mathcal{B}|m^{t-1}, \mathcal{B} \cup \mathcal{Y}] \geq \Pr[\mathcal{Y}|m^{t-1}, \mathcal{B} \cup \mathcal{Y}]. \tag{DC}$$

The deniability constraint (DC) says that the observer thinks it (weakly) more likely that the true state belongs to $\mathcal{B}$ as opposed to $\mathcal{Y}$ when the committee recommends accepting the proposal. It is equivalent to saying that the measure of the residual part of $\mathcal{B}$ is at least as large as the measure of the residual part of $\mathcal{Y}$, where by measure of a residual set we simply mean the (Lebesgue) measure of a set obtained after deleting the states that have been ruled out by the observed history of messages.[15]

Since the committee is biased in the direction of the proposal relative to the observer, (DC) covers the only case where intervention by the observer is possible. If the committee proposes rejecting the proposal in a subversive conversation, the observer has no incentive to intervene. Since the observer approves of every committee decision if (DC) holds, the observer should have no incentive to intervene before the committee comes to a decision.

The conditioning on m^{t-1} in (DC) reflects the ability of the observer to observe committee

[14] It is also immediate that if a subversive conversation exists in our cheap talk framework, it extends to a framework with verifiable information.

[15] The residual sets $\mathcal{B}$ and $\mathcal{Y}$, after some states have been deleted, may in fact be subsets of $\mathbb{R}^1$, and as such they would have measure zero when the state space is considered to be $\mathbb{R}^2$. In these cases by measure of the residual set we mean the Lebesgue measure on $\mathbb{R}^1$.

communications or have access to them, e.g., by subpoena or FOIA laws. Indeed, we call (DC) the deniability constraint to emphasize that the committee has to conceal information from the observer while deliberating in public, without recourse to literal concealment via secret or private information exchange. Because of our focus on subversion, an optimal communication strategy that satisfies (DC) remains optimal if some or all of the history is observed only with some probability, e.g., a leak or hack of private communications.[16]

To conclude the formal description of our baseline model we impose an admissibility restriction on subversive conversations that rules out paradoxical constructions associated with the continuum. For a given subversive conversation, let $\mathcal{H}_t^* = \{m^t : m_t = A\}$ be the set of all message histories that terminate in round t with a recommendation to accept the proposal when the players use this conversation. The deniability constraint (DC) must hold for each $m_t \in \mathcal{H}_t^*$. Our admissibility restriction on the conversation says that it must also hold when we integrate over all message histories that belong to any $\mathcal{H} \subset \mathcal{H}_t^*$.[17] We turn now to constructive proofs of the existence of subversive conversations that are admissible in this sense.

2.2 Existence

Figure 3 depicts a subversive conversation for the case $c = 4/9$. Panel (a) depicts round 1 of the conversation in which player 1 reveals the exact value of x when x is in $[4/9, 5/9]$. Player 2 then recommends acceptance if $y \leq x + c$ and rejection otherwise. Middling information in $[4/9, 5/9]$ can be revealed because the deniability constraint (DC) is satisfied when player 2 proposes accepting the alternative after such information is revealed. In particular, the residual part of $\mathcal{B}$ after $x \in [4/9, 5/9]$ is revealed and player 2 proposes acceptance is the interval $[0, x]$ while the residual part of $\mathcal{Y}$ is the interval $[x, x + 4/9]$, and 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. On other hand, although the deniability constraint is satisfied for $x > 5/9$, revealing such x in the first round will jeopardize the possibility of subversion in later parts of the conversation and in other states of the world. Indeed, if $x \notin [4/9, 5/9]$, we suppose that player 1 says "pass", i.e., she turns the conversation over to player 2 and we enter round 2. The observer learns from this move that the set of possible states (x, y) is either in $[0, 4/9) \times [0, 1]$ or in $(5/9, 1] \times [0, 1]$. The pooling created by player 1's pass preserves a sufficiently high chance

[16] The deniability constraint is similar to the obedience constraint in the literature on Bayesian persuasion (Kamenica and Gentzkow, 2011), but differs since the committee rather than observer has decision rights and the observer can only impose a penalty on the committee for violating the constraint.

[17] Given a conversation (i.e., a strategy profile) the set of histories that can be generated is fully determined by the state (x, y) and so this integration is with respect to the measure on states generated by the conversation history and the priors. Our admissibility restriction automatically holds in any finite environment. But, in the continuum one can exploit the fact that conditional probabilities are not well defined for zero measure events to create measure-theoretic paradoxes. For instance, one can pool good and bad news by constructing a bijection between two sets of unequal measure. Our restriction rules out such paradoxical constructions.

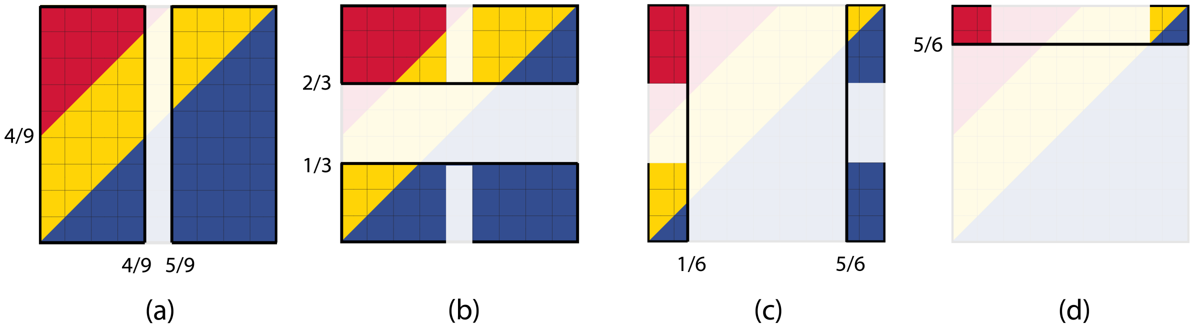


Figure 3: Subversive conversation: spreading protocol

that the proposal is acceptable for the observer, meeting the deniability constraint at this point in time. As we show below, it also creates slack in the deniability constraint for types of the other player that otherwise would not have such slack, allowing the committee to move forward with the conversation without objections from the observer.

Panel (b) depicts what happens in round 2 if player 1 passes in round 1. Player 2 reveals the exact value of y if $y \in [1/3, 2/3]$, following which player 1 recommends the committee's optimal decision given all information available to her. If $y \notin [1/3, 2/3]$, player 2 passes the conversation over to player 1 and we move to round 3. What happens when player 2 reveals y in round 2? Given the history of the conversation, in particular the pooling of disjoint sets created by player 1's round 1 pass, the deniability constraint is satisfied for all possible types that can be revealed by player 2 in this round when player 1 subsequently proposes a decision to accept the proposal.

For instance, when player 2 reveals $y = 2/3$, which is the most troubling information for the observer as y is a cost, the residual part of $\mathcal{B}$ is $[y, 1] - [4/9, 5/9]$ while the residual part of $\mathcal{Y}$ equals $[y - 4/9, y] - [4/9, 5/9]$. This is because round 1 play has revealed $x \notin [4/9, 5/9]$. Notice that if this information was not revealed in round 1 by player 1's pass, the deniability constraint would not be satisfied when player 2 reveals $y = 2/3$ in round 2. In fact, this is true for any $y > 5/9$. By not revealing types $x > 5/9$ and passing 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. Creating this slack comes at a cost however. The round 1 pass also tightens the deniability constraint for lower types of player 2 (types $y \in [1/3, 4/9)$) because for those types it removes states that belong to the agreement set $\mathcal{B}$. But all these lower types of player 2 have slack to spare in the deniability constraint.

In the process of gradual information revelation that we are describing, the players act cooperatively, dynamically creating slack in the deniability constraint for each other and trading slack across types. As each player reveals her signal or passes, she effectively creates a context that reveals which states remain possible and which do not. This context allows the other player to provide information that she could not safely reveal otherwise, or to further refine the context and extend the process of gradual information revelation on to later rounds. The process we are describing is, in fact, a conversation. It is designed for perfect coordination while maintaining deniability, ensuring there are no objections from the observer.

Panels (c) and (d) of Figure 3 continue this line of reasoning. If a decision has not been taken by round 3, player 1 reveals the exact value of x if $x \in [1/6, 5/6]$, passing the conversation back to player 2 otherwise. As shown in panel (c), types $x \in [1/6, 4/9)$ can be safely revealed by player 1 in this round, even though they could not be in round 1, because of the context created player 2's pass in round 2. When $x \notin [1/6, 5/6]$, player 2 passes again and we enter round 4. At this point enough states that are unfavorable from the perspective of meeting the deniability constraint have been eliminated as possible states, while enough favorable states remain possible. The conversation takes one more round. In round 4, player 2 reveals her signal y if it is in $[0, 1/3]$, passing otherwise, as depicted in panel (d). In the former case, player 1 proposes the committee optimal decision in round 5, taking into account her own signal. In the latter case, player 1 simply proposes acceptance (without revealing her signal) in round 5 when x is in $[5/6, 1]$, and recommends rejection otherwise. The deniability constraint will be satisfied in all cases and the conversation ends by round 5 regardless of the state of the world.

Theorem 1(i) below shows that the gradual spreading process depicted in Figure 3 and described above works as a subversive conversation for each $c < 1/2$. A similar conversation also works for $c \geq 1/2$ but with different messages in the initial rounds that take into account the higher conflict between the committee and the observer. The details of the constructions are in the proof.

Theorem 1 *(i) For each c, there exists a subversive conversation. (ii) Any subversive conversation requires at least three rounds.*

The conversation constructed for Theorem 1 highlights how the two players create increasingly refined contexts, cooperatively creating slack in the deniability constraint for each other, in order to safely reveal information that could not be otherwise revealed. Such a subversive conversation allows the players to reach their optimal decision while satisfying the deniability constraint. Regardless of the true state of the world, the conversation ends within a finite amount of time that can be specified in advance.

This uniform bound on the length of the conversation used to prove Theorem 1(i) depends on the conflict c between the committee and the observer. Subversion under higher conflict may require more time than under lower conflict. But, regardless of the construction, part (ii) of Theorem 1 shows that subversion must involve back-and-forth communication. Since $\mathcal{R}$ is nonempty the players need to exchange information in order to determine their optimal decision. But, since $c > 0$, they cannot do so immediately in order to maintain deniability. As shown in the proof in the Appendix, for any c this requires at least three rounds.[18]

[18] If we allow the case $c = 1$ where $\mathcal{R}$ is empty, there is no coordination problem and the committee can end the

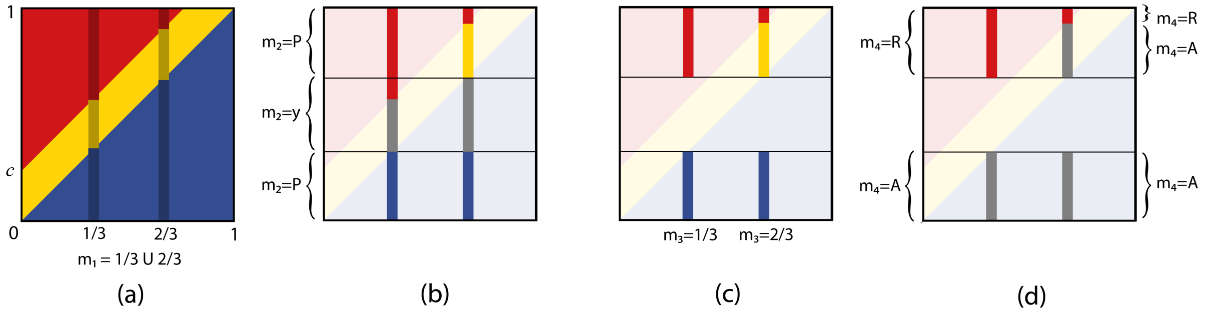


Figure 4: Four-round universal conversation

For our next result, we construct an alternative conversation to that employed to establish Theorem 1(i). This conversation takes at most four rounds to conclude, regardless of the level of conflict c. The key to both constructions is that a message by one player that suitably pools favorable and unfavorable information creates slack in the deniability constraint (DC) for the other player, similar to what was described above when one player said “pass”. The key to saving time with our next construction is that instead of both players engaging in such pooling only one player does so. Figure 4 depicts this construction.

Looking at Figure 4, player 1 first reports the pooled message $m_1 = \{a, 1-a\}$ which reveals that player 1’s signal x is either equal to a or equal to $1-a$, where $a \leq 1/2$. The exact value of x is not revealed at this stage. Figure 4(a) illustrates this first round for the case $a = 1/3$ so that everyone learns x equals either 1/3 or 2/3.[19]

After the pooled message in round 1, player 2 reveals her signal y if $y \in [a, 1-a]$ and player 1 then proposes a decision in round 3, as seen in Figure 4(b). Otherwise, if $y \notin [a, 1-a]$, player 2 reveals this fact and we move on to round 3 when player 1 reveals the exact value of her signal $x \in \{a, 1-a\}$, following which player 2 recommends a decision in round 4. This is depicted in panels (c) and (d).

To see that the deniability constraint is satisfied whenever the players use this conversation, consider first the case where player 2 has revealed her exact signal $y \in [a, 1-a]$ in round 2 following the first message that x belongs to $\{a, 1-a\}$. The two possible states of the world at this stage are either (a, y) or $(1-a, y)$. Since $y \leq 1-a$, we must have $(1-a, y) \in \mathcal{B}$. So the worst case scenario from the perspective of meeting the deniability constraint is when $(a, y) \in \mathcal{Y}$. But even in this case, since the two possibilities (a, y) or $(1-a, y)$ are equally likely, the deniability constraint must be satisfied. It follows that the deniability constraint is satisfied in all cases where the decision is taken in round 2.

conversation in one round simply by proposing acceptance without exchanging messages. While all our results extend to the case $c = 1$, by assuming $c < 1$ we focus on more interesting cases where the committee needs to exchange information in order to coordinate on their own optimal decision.

[19] To aid the reader, the sets $\{a\} \times [0,1]$ and $\{1-a\} \times [0,1]$ are drawn in the figure thicker than they actually are.

Consider next the case where player 2 has revealed $y \notin [a, 1-a]$ following the first message that x belongs to $\{a, 1-a\}$. In round 3, player 1 reveals the exact value of $x \in \{a, 1-a\}$. Suppose player 2 subsequently recommends acceptance in round 4. Given the observed history, the measure of the residual part of $\mathcal{B}$ is equal to a regardless of the exact value of x, while the measure of the residual part of $\mathcal{Y}$ can be seen to be $\max[\min\{x+c, 1\} - (1-a), 0]$. It is straightforward to verify that the former is at least as large as the latter for all possible a, implying that the deniability constraint is met in all cases where the decision is taken in round 4.

Notice that while the final recommendations to accept or reject the proposal depend on c as they must, the structure of the above conversation and all messages other than the final recommendation are independent of c. So the exact same conversation can be used for any c. Call a subversive conversation with this property a *universal* subversive conversation. Our next result shows that there is no other universal subversive conversation that is quicker than the one described above.

Theorem 2 *(i) There exists a four-round universal subversive conversation. (ii) Any universal subversive conversation requires at least four rounds.*

The unifying feature of the constructions in Theorems 1 and 2 is that disjoint types with good and bad news are pooled. As shown in Figure 3, for the conversation used for Theorem 1, the disjoint pooling is created by a player passing the conversation over to the other player in each round. Each such pass creates slack in the deniability constraint for the *other* player for the next round. A pass by a player does not directly benefit the player who is passing. This is the key to why the conversation used for Theorem 2 saves time compared to the one used for Theorem 1.

To see the last point, consider the case where player 1's type $x \in \{1/3, 2/3\}$. In the conversation employed for Theorem 1, such types are revealed in round 3, after the other player has revealed $y \notin [1/3, 2/3]$. Compared to this, the shorter conversation used for Theorem 2 saves time, essentially by interchanging the order of moves. In the first round, player 1 sends a message "$x \in \{1/3, 2/3\}$" without revealing her exact signal. Such an equivocal initial message by player 1, that pools disjoint types (and means either something or its exact mirror image), allows player 2 to safely reveal her own signal y whenever $y \in [1/3, 2/3]$, since it is known at this point that $x \notin (1/3, 2/3)$. Otherwise, player 2 pools disjoint sets herself by revealing $y \notin [1/3, 2/3]$. This allows the first player to clarify the meaning of her first message and reveal if $x = 1/3$ or $x = 2/3$. By interchanging the order of moves in this way, a decision is taken for sure by round 4.

We conclude this section by asking why subversion is always possible in this baseline model. Looking back at the deniability constraint (DC), notice that a necessary condition for subversion is that the agreement set $\mathcal{B}$ is at least as large in measure as the conflict set $\mathcal{Y}$. This *total slack* condition is always met in the baseline model. It allows all points in $\mathcal{Y}$ to be pooled with points in $\mathcal{B}$.

To see the economic interpretation of the total slack condition, consider a centralized expert who has the same preferences as the committee members but who knows both x and y. For the centralized expert to be able to subvert, the total slack condition is necessary and sufficient. As long as there is non-negative total slack, the centralized expert can meet the deniability constraint simply by recommending acceptance without revealing anything else. But the centralized expert can also subvert by recreating any subversive conversation available to the committee. So she can subvert whenever the committee can. This implies the total slack condition is necessary for the committee to subvert. But it is not sufficient. Because its information is decentralized, the committee also needs to exchange information publicly in order to coordinate on its ideal decision. Although the committee does as well as the centralized expert in the baseline model, in general, decentralization of information puts an additional burden on the committee's attempt to subvert. In Section 4 we identify this additional burden. But we turn first to generalizations of the results obtained above for our baseline model.

2.3 Generalizations

In this section we provide a general property of subversive conversations that allows us to relax the special assumptions on priors and preferences that we have relied on so far. Given a state space $\mathcal{S}$ (equal to $[0,1]^2$ in the above baseline model) and independent priors F_x and F_y for x and y (uniform in the baseline model), a game $\Gamma = \{\mathcal{R}, \mathcal{B}, \mathcal{Y}; F_x, F_y\}$ is defined by the pair of priors F_x, F_y and the three (measurable) sets $\mathcal{R}$, $\mathcal{B}$ and $\mathcal{Y}$ that specify the zones of agreement on rejecting and accepting the proposal as well as the zone of conflict, with $\mathcal{R} \cup \mathcal{B} \cup \mathcal{Y} = \mathcal{S}$.

Call a subversive conversation for Γ a *fine subversion* if every time a decision to accept is taken by the committee, 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. The construction used to establish Theorem 1 is a coarse subversion, while that for Theorem 2 is a fine subversion. In a fine subversion one of the players learns the exact state (x, y) but the other player and the observer may remain uncertain. In a coarse subversion both players as well as the observer sometimes remain unsure of the exact state.

Our next result compares two games Γ and Γ' that have the same state space $\mathcal{S}$ and priors F_x, F_y, but differ in the sets $\{\mathcal{R}, \mathcal{Y}, \mathcal{B}\}$ and $\{\mathcal{R}', \mathcal{Y}', \mathcal{B}'\}$. It shows that the property of universality of subversive conversations described in Theorem 2 is, in a sense made precise by the result, a general property.

Theorem 3 *Fix $\mathcal{S}$, F_x, F_y. If a fine subversion exists for some game Γ, then the same conversation is also a fine subversion for any other game Γ' with $\mathcal{Y}' \subset \mathcal{Y}$, $\mathcal{B}' \supset \mathcal{B}$ and $\mathcal{R}' \supset \mathcal{R}$. If a* *coarse subversion exists for some game* Γ*, then the same conversation is also a coarse subversion for any other game* Γ' *with* $\mathcal{Y}' \subset \mathcal{Y}$*,* $\mathcal{B}' \supset \mathcal{B}$ *and* $\mathcal{R}' = \mathcal{R}$*.*

Theorem 3 says that a subversive conversation remains so if we change preferences in a game Γ in order to reduce, in the sense of subsets, the zone of conflict and increase the zone of agreement on the proposal. 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.

The theorem is a consequence of the following observations. Anytime the committee makes a recommendation to accept the proposal in accordance with the fixed conversation, the observed history is the same in Γ and Γ'. Conditional on this observed history, the residual conflict set $\mathcal{Y}'$ is smaller and the residual agreement set $\mathcal{B}'$ is larger in Γ', compared to the corresponding sets in Γ. In the case of a fine subversion, if (without loss of generality) the exact value of x has been revealed and player 2 makes a recommendation to accept the proposal in game Γ, conditional on the revealed value of x and the observed history, the observer learns that $(x, y) \in \mathcal{Y} \cup \mathcal{B}$ and (DC) is satisfied. But then, when the same recommendation is made in Γ', conditional on the same x and observed history, (DC) must also be satisfied. This is because the (residual) agreement set $\mathcal{B}'$ is larger and the (residual) conflict set $\mathcal{Y}'$ is smaller in Γ' compared to Γ.

In the case of a coarse subversion, if at any point in Γ a player makes a coarse recommendation to accept the proposal without either x or y being revealed, it must be that the residual agreement set $\mathcal{R}$ is empty conditional on the observed history, since otherwise the conversation cannot be subversive. Since in this case $\mathcal{R}' = \mathcal{R}$, the same must be true in Γ' after the same history. It follows that the same coarse recommendation will also be subversive and satisfy (DC) in Γ', once again since $\mathcal{Y}' \subset \mathcal{Y}$ and $B' \supset \mathcal{B}$. This establishes Theorem 3.

An immediate consequence of Theorem 3 is that subversive conversations are, in general, not unique. A fine subversive conversation that works for a given level of conflict will also work for a lower level of conflict. Put differently, a fine subversive conversation for a larger conflict set is universal, i.e., it is also a subversive conversation for every possible smaller conflict set. In this sense, the existence of a universal conversation established by Theorem 2 is a general result that extends beyond the baseline model, in the sense made precise by Theorem 3.

Returning to the baseline model, we can use Theorem 3 to extend the domain of our results in a number of ways. Recall from the construction for Theorem 2 that the exact same conversation works for all levels of conflict. Theorem 3 implies that the same conversation also works as a subversive conversation even when the borders between $\mathcal{R}$ and $\mathcal{Y}$ and between $\mathcal{Y}$ and $\mathcal{B}$ are non-linear, as long as the new conflict set $\mathcal{Y}'$ is a subset of the original conflict set $\mathcal{Y}$ and the new agreement set $\mathcal{B}'$ is a superset of the original agreement set $\mathcal{B}$. Defining $\mathcal{U}$ as the region above the observer's decision line $y = x$ in Figures 3 and 4, and $\mathcal{L}$ as the region below, we have the following corollary.

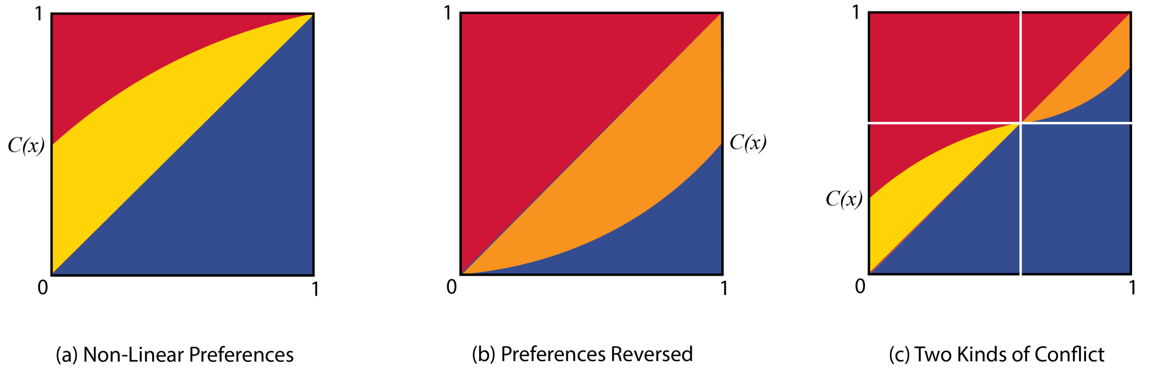


Figure 5: Subversion and subsets

Corollary 1 (Non-linear preferences) *A fine subversive conversation exists for all games* Γ *with* $\mathcal{Y} \subset \mathcal{U}$ *and* $\mathcal{B} \supset \mathcal{L}$.

Figure 5(a) shows the baseline model except the committee's bias c is not constant but depends on the state. Letting $y = C(x)$ be the committee's decision line, the conditions of Corollary 1 apply so a subversive equilibrium exists. Figure 5(b) shows a similar situation except the committee has a bias toward rather than against the status quo while the observer remains unbiased, creating a conflict zone pictured in orange. This situation is just an "upside down" version of the case in 5(a), so subversion follows by applying the exact same logic.

Figure 5(c) shows a case where not just the value but the sign of the conflict is state-dependent, so the committee's decision line $y = C(x)$ crosses the observer's decision line.[20] There are two kinds of deniability constraints, one where the committee recommends acceptance and another where the committee recommends rejection. This situation is not directly covered by Theorem 3 which only allows for one kind of conflict and so one kind of deniability constraint. Nevertheless, the players can use a conversation to divide the state space into subspaces, each of which allows subversion.

Looking at the figure, player 1 first discloses whether x lies above or below the vertical cutoff (i.e., to the right or left of the intersection point). Player 2 then reveals whether y is above or below the horizontal cutoff. If 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 since there is no conflict in these states. And if the state is revealed to be in the lower left or upper right, then the conversation proceeds as it would for the games in Figures 5(a) and 5(b), in each of which subversion is possible via Theorem 3. This same argument holds as long as $C(x)$ is strictly increasing, and no matter how many times $C(x)$ intersects the observer's decision line. We have the following result.

[20] Such preferences are considered by Gordon (2010) in the context of the classic cheap talk model of Crawford and Sobel (1982).

Corollary 2 (Two kinds of conflict) *Suppose the observer's decision line is the diagonal $y = x$ while the committee's decision line is given by a possibly non-linear function $y = C(x)$. Then a fine subversive conversation exists for all strictly increasing $C(\cdot)$.*

The next corollary of Theorem 3 relaxes the uniform distribution assumption of the baseline model. Suppose instead that x and y have possibly non-uniform independent priors F_x and F_y that are continuous and strictly increasing (i.e., invertible). We can then transform variables and specify preferences in terms of the quantiles $q_x = F_x(x)$ and $q_y = F_y(y)$. With this change of variables, the observer's decision line that forms the border between $\mathcal{B}$ and $\mathcal{Y}$ can be written as $q_y = F_y(F_x^{-1}(q_x))$. Similarly, the committee's decision line that defines the border between $\mathcal{R}$ and $\mathcal{Y}$ can be written as $q_y = F_y(F_x^{-1}(q_x) + c) > F_y(F_x^{-1}(q_x))$. It follows that in the space of quantiles $(q_x, q_y) \in [0, 1]^2$, the conflict set of the transformed problem is a subset of the upper triangle $\mathcal{U}$, while the set of agreement on the proposal is a superset of the lower triangle $\mathcal{L}$ if F_x and F_y are identical, and in fact as long as F_x first order stochastically dominates F_y. Since quantiles are uniformly distributed, we have our next corollary.

Corollary 3 (Non-uniform priors) *A fine subversive conversation exists for any priors F_x, F_y such that F_x first-order stochastically dominates F_y.*

Observe that the corollaries above can be used together. For instance, once we are guaranteed the existence of a subversive conversation for arbitrary priors and linear preferences via Corollary 3, we are also guaranteed existence for non-linear preferences that satisfy the conditions of Corollary 1.

Our final corollary to Theorem 3 considers uncertainty about preferences, including higher order uncertainty. For instance, 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 c is either c' or c'' with $c', c'' \in (0, 1)$ and $c' > c''$. Using Corollary 1, a fine subversive conversation for the worst case scenario $c = c'$ will also work when the observer believes instead $c = c'' < c'$. We state this result more generally as follows.

Corollary 4 (Uncertainty about preferences) *A fine subversive conversation exists when the committee believes the game is Γ whereas the observer believes the game is Γ' for any $\mathcal{Y}, \mathcal{Y}' \subset \mathcal{U}$ and $\mathcal{B}, \mathcal{B}' \supset \mathcal{L}$.*

These corollaries show how Theorem 3 can be used to generalize the results of our baseline model, but it is a more general result that applies to arbitrary state spaces. In the next section we consider some variations of the baseline model that differ either in terms of the state space or in terms of preferences.

3 Variations and Extensions

We now turn to several variations on the baseline model. For each we maintain the assumptions of the baseline model except for the specific difference being analyzed.

3.1 Subverting an Ex Ante Skeptical Public

Absent any information from the committee, would the observer support the proposal or the status quo? In the baseline model $\Pr[\mathcal{B}] \geq \Pr[\mathcal{Y}]$ for all c so the observer prefers the proposal without any communication. But often the observer will be skeptical if the cost is more likely to exceed the benefit. The committee then has the added burden of not just concealing bad information from the observer, but also of raising the posterior probability of a good proposal enough to meet the deniability constraint.

To capture such situations we now suppose the random variables x and y have different supports as depicted in the first panel of Figure 6(a). While the benefit x is uniformly distributed over $[0, 1]$, the cost y is uniformly distributed over $[0, 3/2]$. Otherwise, the situation is identical to the baseline model. The observer prefers the alternative if the benefit exceeds the cost, $x > y$, whereas the committee prefers the alternative as long as $x + c \geq y$.

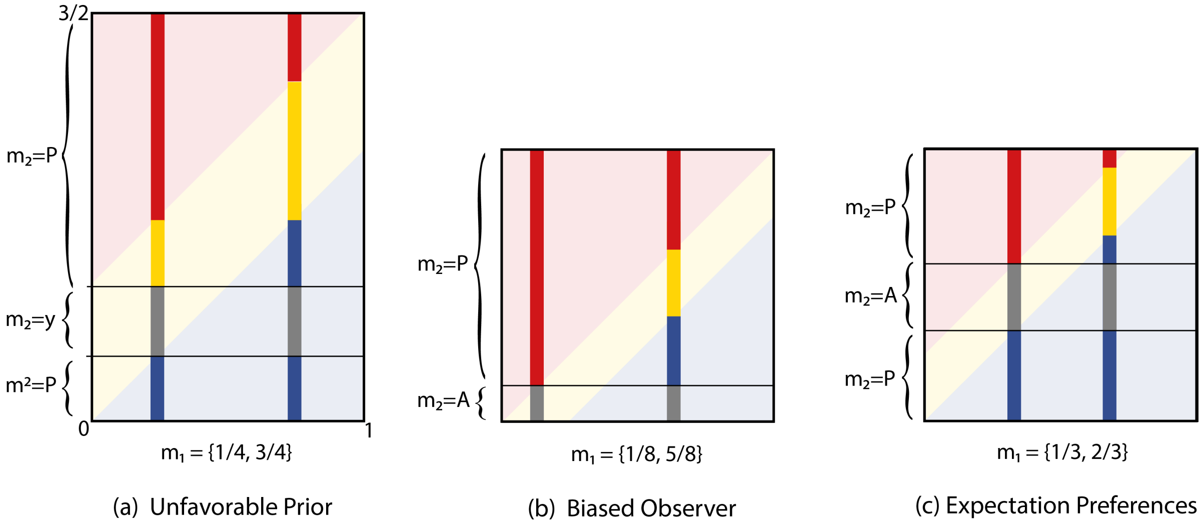
Recall that a centralized expert can subvert if and only if there is non-negative total slack, or $\Pr[\mathcal{B}] \geq \Pr[\mathcal{Y}]$, which is a necessary condition for subversion by the committee. This condition holds for all $c \in (0, 1]$ in the baseline model. For this example with an unfavorable prior, there is non-negative total slack only if the conflict parameter is low enough, $c \leq 1/2$. Figure 6(a) shows the case of $c = 1/2$ where there is zero total slack. Hence the experts must successfully coordinate without any slack at any point in the deniability constraint. If a fine subversion is possible in this case with $c = 1/2$, it is also possible when $c \leq 1/2$, via Theorem 3.

Figure 6(a) depicts a fine subversion for the case $c = 1/2$ that is similar to that of Proposition 1 except that after player 1 reveals x is in $\{a, 1 - a\}$, $a \leq 1/2$, player 2 reveals y for $y \in [a + 1/2, 1/2]$, rather than for $y \in [a, 1 - a]$. Player 1 meets the deniability constraint when he recommends acceptance when any such y is revealed. On the other hand, when player 2 reveals y is not in the interval $[a + 1/2, 1/2]$, this pooling allows player 1 to reveal the exact value of $x \in \{a, 1 - a\}$, and player 2 to recommend acceptance or not.

Proposition 1 (Unfavorable prior) *In the baseline model with support on* $[0, 1] \times [0, 3/2]$ *the committee can subvert for all* $c \leq 1/2$.

3.2 Unbiased Experts Facing a Biased Observer

The baseline model assumes the observer is unbiased while the players are biased toward the proposal. In this section we consider the opposite case where the observer is biased while



(a) Unfavorable Prior

(b) Biased Observer

(c) Expectation Preferences

Figure 6: Variations and extensions

the players are not. Specifically, suppose that $u(x, y) = 1$ if $x \geq y$ and -1 otherwise, while $v(x, y) = 1$ if $x - r \geq y$ and -1 otherwise, for some bias parameter $r > 0$ for the observer. Hence the players prefer the proposal if the net benefit is positive, while the observer only prefers it if the net benefit is at least r.

Looking at Figure 6(b), the problem is not strategically symmetric with the baseline model because now $\Pr[\mathcal{B}] < \Pr[\mathcal{R} \cup \mathcal{Y}]$ so the observer will reject the proposal without sufficient additional information from the players. As in the previous example, the players have the extra burden that they need to share enough information with the observer to overcome his bias against the proposal. Moreover, the committee's coordination needs are now more demanding compared to the baseline model since for every value of x there is a unique interior value of y that is the cutoff for the committee accepting or rejecting the proposal. In addition, for sufficiently low values of the benefit x there is no possible value of the cost y that can persuade the observer that the proposal is worthwhile, and similarly for sufficiently high values of the cost. Compared to the previous example, this puts additional constraints on the information that can be revealed by the players during a conversation.

Is subversion still possible in this case? If the observer is too biased, r sufficiently large, then the total slack condition is violated, so subversion is impossible even for a centralized expert. Figure 6(b) shows the first two rounds of a four-round protocol for the case where $r = 1/4$, which is less than the largest r for which there is non-negative total slack, $r = 1 - 1/\sqrt{2} \approx .2929$. As in the baseline model, player 1 first pools high and low values of x, but in this protocol player 1 pools $x \in \{a, 1/2 + a\}$ for $a \in [0, 1/2]$. Player 2 then accepts the proposal if $y < a$ and otherwise passes. If player 2 has passed, player 1 then reveals the exact value of x, following which player

2 recommends a decision. As seen in the figure for $x \in \{1/8, 5/8\}$, in this manner the players can pool all conflict and agreement states while meeting the (CC) and (DC) constraints. Since this protocol is a coarse subversion, we can use the version of Theorem 3 for coarse subversions to conclude that the same conversation is also subversive for all smaller r. We have the following result.[21]

Proposition 2 (Biased Observer) *In the baseline model with $v(x, y) = 1$ if $x - r \geq y$ and -1 otherwise, and $u(x, y) = 1$ if $x \geq y$ and -1 otherwise, the committee can subvert for all $r \leq 1/4$.*

3.3 Expectation Preferences for the Experts and Observer

So far we have assumed that expected payoffs depend on the probability that the proposal is sufficiently good. Suppose instead that the committee's payoff from accepting the proposal is $x - y + c$ (while that from the status quo is zero) and the observer's payoff from accepting the proposal is $x - y$ (while that from the status quo is zero). Both the committee and the observer now care directly about the magnitude of the net benefit $x - y$. Is subversion possible in this case?

For the committee only ordinal preferences matter. Expectation preferences maintain the same ordinal ranking as the baseline model, keeping unchanged the notion of subversion and the coordination constraint. In contrast, the change in the observer's preferences affects the deniability constraint which changes to $E[x - y | m^{t-1}, \mathcal{B} \cup \mathcal{Y}] \geq 0$. Meeting the deniability constraint now involves computing also the centers of gravity (i.e., conditional expectations) of the residual parts of $\mathcal{B}$ and $\mathcal{Y}$, as opposed to simply their sizes (i.e., measures). Looking back at the universal conversation in Figure 4(b) for the pictured case $x \in \{1/3, 2/3\}$, notice that when $m_2 = y$ is revealed in the range $[a, 1 - a] = [1/3, 2/3]$, the expected value of the project conditional on an x recommendation in the next round is not positive for y in the range above $1/2$, so the message $m_2 = y$ does not meet the deniability constraint and subversion fails.

Consider instead the following variation on this protocol that is coarse rather than fine. Suppose that as in Figure 6(d) in the second round player 2 pools in the range $y \in [a, a + c]$ with an accept recommendation $m_2 = A$, in which case the constraint is still met. Otherwise player 2 passes and in the next round player 1 reveals $m_3 = x$, after which player 2 recommends $m_4 = A$ if $y \leq x + c$ and $m_4 = R$ otherwise, both of which meet the deniability constraint. Extending this adjustment to the protocol for other values of x not pictured, in round 1 player 1 pools $m_1 \in \{a, 1 - a\}$ as before. Then for all $x \in [0, (1 - c)/2] \cup [(1 + c)/2, 1]$ the same pool $m_2 = y \in [a, a + c]$ is used. For values of x closer to $1/2$ the pool region needs to include lower values of y to maintain the deniability constraint, so the pool is $m_2 = y \in [1 - (a + c), a + c]$

[21] More intricate subversive conversations exist for $r > 1/4$ and we conjecture that subversion is possible for all r with non-negative total slack.

for $a + c < 1$ and $m_2 = y \in [0, 1]$ otherwise. With the deniability constraint satisfied, the conversation proceeds as in the original protocol and is subversive.[22]

Proposition 3 (Expectation preferences) *In the baseline model with $v(x, y) = x - y$ and $u(x, y) = x - y - c$, a subversive conversation exists for all c.*

3.4 Other Extensions

The model can be extended in a number of other directions to fit particular situations. We briefly discuss two additional variations with alternative preference specifications, the first for the observer and the second for the committee members. A full treatment of either of these extensions deserves a separate paper.

Alternative Evidentiary Standards. The deniability constraint of the baseline model says that the observer will not object as long as the updated probability that the state lies in $\mathcal{B}$ is at least as large as that it lies in $\mathcal{Y}$. This evidentiary standard corresponds to the notion of "balance of probabilities" (or "preponderance of evidence") that is used for civil cases under U.S. law. Is the possibility of subversion robust to other evidentiary standards?

To check this suppose that for some weight $w > 0$ the observer's preferences are $v(x, y) = 1$ if $x \geq y$ and $-w$ otherwise. In a subversive equilibrium only the committee's ordinal preferences matter so we leave them as in the baseline model. If $w < 1$, then the observer requires a higher burden of proof before he can intervene, compared to the baseline model. For instance, the "reasonable doubt" burden of proof used for criminal cases under U.S. law is a more stringent burden than the balance of probabilities threshold. Since the deniability constraint is easier to satisfy for smaller w and the coordination constraint is unchanged, a conversation that is subversive when the observer faces a lower evidentiary standard, must continue to be subversive under a more demanding burden of proof for the observer. So all of our results for the baseline model where $w = 1$ extend to the $w < 1$ case.

What happens when $w > 1$ and the observer faces a less demanding burden of proof than in the baseline model? Each point in the disagreement set $\mathcal{Y}$ now has a higher weight $w > 1$ compared to any point in the agreement set $\mathcal{B}$. To adjust for this differential weighting, we can use an approach similar to the four-round protocol of the baseline model except in the first round the lower pooling region has positive measure that is equal to $1/w$ the measure of higher pooling region. It can be shown that such a conversation is subversive in this variation of the baseline model as long as c is sufficiently small.

[22] Note that the coarse protocol used for the Biased Observer example also works in that example with expectation preferences. While the fine subversion used for the Unfavorable Prior example does not extend to expectations preferences, a simple coarsening does—instead of player 2 revealing her signal when $y \in [a, a + 1/2]$, she just proposes acceptance without revealing her signal in these states.

Within-Committee Conflicts. In the baseline model the two players in the committee have identical preferences so the only conflict is with the observer. If there are also conflicts of interest between the players then there is no ex-post optimal decision common to all committee members, and even private communication within the committee is subject to incentive problems. So our notion of subversion from the common interest committee of the baseline model needs to be suitably redefined.

A committee with conflicts of interest can subvert if, even in the presence of the observer, it can implement an incentive-compatible ex ante efficient decision rule of the game of sequential private communication between the committee members.[23] When this is the case, the presence of the observer has no effect on committee payoffs and decision rules, just as in the original definition of subversion.

It can be shown that our main result on the possibility of subversion via back and forth conversations is robust to small within-committee conflicts. Because of these conflicts, it is incentive compatible for the committee to engage only in coarse information transfer (similar to the interval partitional equilibria of Crawford and Sobel, 1982), even in the absence of the observer. In the presence of the observer, such coarse information transfer eases the coordination needs of the committee while also allowing the committee to better conceal the exact state, compared to the case of a fine subversion by a common interest committee.

4 The Limits of Subversion

In our baseline model, the committee can always subvert. Since there is non-negative total slack, so can the centralized expert. Therefore, the committee does as well as the centralized expert. We turn now to question of when, in general, is it impossible for the committee to subvert. There are two cases to consider.

In the first case, there is negative total slack. This implies the centralized expert cannot subvert and so the committee cannot either. We show that in this case the committee with decentralized information does at least as well as the centralized expert, and in some cases strictly better. In the second case, there is non-negative total slack which implies the centralized expert can subvert. For this case, we provide a class of examples where subversion by the committee is impossible because of insufficient concordance between the committee's preferences and those of the observer. We also relate our results to Hall's (1935) classic marriage theorem. We show that the committee may fall short of the centralized expert exactly because it has to coordinate while holding decentralized information. For simplicity, throughout this section we restrict attention to discrete type examples similar to the introductory example.

[23] Such decision rules are analyzed in Chakraborty and Yilmaz (2017). While in general there are multiple such decision rules, their results imply that when the within-committee conflict is small enough, there is a unique such rule.

4.1 Negative Total Slack

Consider a game with negative total slack, i.e, $\mathcal{B}$ is smaller in measure than $\mathcal{Y}$. In such a situation, the centralized expert cannot subvert. What can she do? In our cheap talk environment there are only two kinds of equilibrium outcomes (or decision rules) that can be implemented by the centralized expert. Either there is an informative equilibrium where the centralized expert must get her ideal decisions (i.e., she subverts), or there is an uninformative babbling equilibrium. That there is a babbling equilibrium is a well known fact for cheap talk games. That any informative equilibrium with a centralized expert must result in the centralized expert getting her ideal decisions is also a well-known fact for binary decision cheap talk games with a single expert. In any equilibrium where the proposal is both accepted and rejected with positive probability, in a way that depends on the expert's information and without any incentive for the observer to intervene, the centralized expert has an incentive to recommend acceptance exactly when she prefers to, and recommend rejection otherwise. It follows the centralized expert can credibly communicate only what she prefers and her recommendations are guided entirely by her own preferences.

When there is negative total slack, the centralized expert cannot be informative. The uninformative babbling outcome is the only outcome for the centralized expert. In the babbling equilibrium, the proposal is always rejected. The committee can clearly match the babbling outcome. Can it do strictly better? We conclude our discussion of this case of negative total slack by presenting an example with an equilibrium that is not subversive but still better for the committee than babbling (and by Blackwell's theorem for the observer as well). Panel (a) of Figure 7 depicts this example.[24]

In the situation depicted in panel (a), there is negative total slack so subversion by the committee is impossible. But the following is a cheap talk equilibrium. Player 1 recommends rejection when her signal corresponds to the left column and recommends acceptance when the benefit is higher and her signal corresponds to the right column. Player 2 says nothing. It is straightforward to check that the observer is willing to accept the proposal when it is recommended (i.e., the deniability constraint is satisfied), and that no player has an incentive to deviate from the prescribed behavior.

The centralized expert cannot mimic the informative equilibrium for the committee described above, precisely because she holds both pieces of information. When she holds the information corresponding to the bottom-left conflict point she has an incentive to engage in a "double deviation" and recommend acceptance via pretending that her signal corresponds to the right column. In addition to the incentive constraints faced by each committee member, the centralized expert also faces an incentive constraint that must prevent such double deviations. This means it

[24] For each of the four examples depicted in Figure 7, the benefit x is increasing as we move right and the cost y is decreasing as we move down, just like all our other examples.

is more difficult for her to be credible compared to the committee. Decentralized information is strictly better than centralized information because decentralization reduces the number of incentive constraints and so raises credibility.[25] Equivalently, public communication in front of the observer is strictly better for the committee than insisting on secure private communication prior to recommending a decision. While subversion is impossible, public communication does a better job of persuading the observer.[26]

4.2 Non-negative Total Slack

We turn now to the second kind of situation that leads to a failure of subversion by the committee. We suppose there is non-negative total slack so that the centralized expert can subvert. When is subversion impossible for the committee? Subversion involves pooling, or matching, conflict points in $\mathcal{Y}$ with agreement points in $\mathcal{B}$. Since the committee's information is decentralized, for a fine subversion the committee is further constrained to match agreement and conflict types that lie in the same row or column. Necessary conditions for the existence of subversive conversations must take into account these constraints.

Figures 7(b) and 7(c) consider two examples similar to our baseline model, each with non-negative total slack. For the example in (b), it is straightforward to verify that fine subversive conversations exist.[27] The example in (c) differs from that in panel (b) only in the observer's preferences. We argue below that no subversive conversation, coarse or fine, exists for the example of panel (c).

To see why subversion is impossible in the example of panel (c), notice first that $\mathcal{R}$ is non-empty, ruling out an initial summary recommendation by one player to accept the proposal without exchanging information. Notice also from the shape of $\mathcal{R}$ that for each possible value of x, the maximum value of y for which the committee prefers the proposal depends on the exact value of x. Similarly, for each value of y, the minimum value of x for which the committee prefers acceptance (but the observer does not) depends on the exact value of y. This implies that at some stage of the conversation, either player 1 or player 2 must reveal her exact signal (or allow it to be deduced), in order to determine the committee optimal decision in all states.

However, whenever a player reveals her signal and the other player subsequently proposes

[25] Player 1 does not have the same incentive as the centralized expert to engage in the double deviation simply because she does not know that the state of the world belongs to the bottom row. Notice further that when player 1 recommends a rejection in the equilibrium played by the committee, the other player prefers to accept the proposal for sufficiently favorable draws of her signal, but doing so will provoke observer objections. Precisely because of the decentralized nature of information, the players can use the threat of objections by the observer to gain credibility and sustain the equilibrium.

[26] Of course, if contrary to the cheap talk environment that is our focus, the centralized expert had commitment she could always mimic the committee. So the committee can never do strictly better than a centralized expert *with commitment* and, in general, does strictly worse.

[27] One example works as follows. Player 1 first reveals the middle column or passes. If she passes, player 2 can reveal the second from bottom row or she can pass. If player 2 passes, player 1 can then reveal either of the two remaining columns.

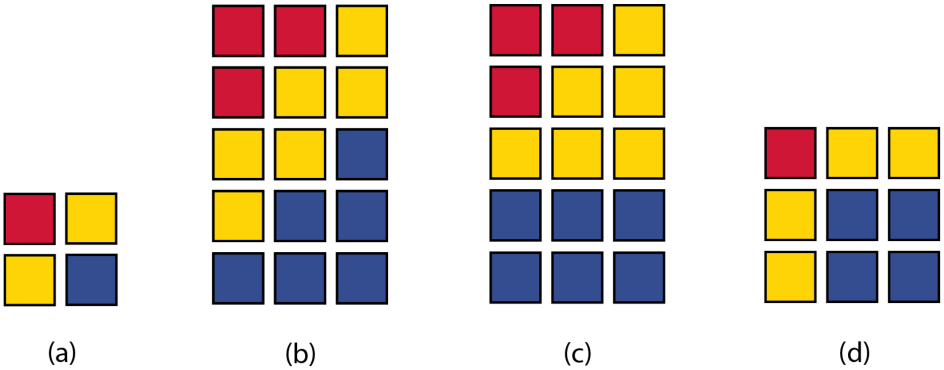


Figure 7: Failures of subversion

acceptance, it must be player 1 who reveals her signal and player 2 who proposes acceptance. It cannot be the other way round. Player 1 cannot pool a conflict point with an agreement point that lies in the same row since there is no agreement point that lives in the same row as a conflict point in panel (c). Yet such pooling is necessary whenever acceptance is proposed because there is zero total slack in this example. To conclude the argument, observe that when player 1 reveals her highest signal (corresponding to the rightmost column) and player 2 proposes acceptance, the deniability constraint cannot be met. Since there is an initial imbalance between conflict and agreement points in this column, this imbalance will remain no matter how many pairs of conflict and agreement points have been paired off prior to this stage.[28]

Subversion of any kind is impossible in the example of Figure 7(c). The economic intuition behind this non-existence result is given by the interaction of two key features of the example. First, because of the shape of $\mathcal{R}$, the committee faces a demanding coordination burden. They need to publicly reveal at least one signal in order to determine their optimal decision. This feature is shared with the example of panel (b). The second key feature, and a crucial difference with panel (b), is that in panel (c) the observer only cares about the cost y and not about the benefit x. He is unwilling to trade off unfavorable information about the cost in exchange for favorable information about the benefit. Consequently, when player 1 reveals the best possible news about the benefit, and player 2 subsequently proposes acceptance, the deniability constraint will not be met because no amount of good news about x can compensate for the possibility of bad news about y, from the perspective of the observer. This argument extends beyond the specific example of panel (c) to any example that shares these two key features.

The mathematical underpinning for the non-existence result in panel (c) is provided by Hall's

[28] Since the committee holds decentralized information, a conflict point in the rightmost column cannot be paired with an agreement point that belongs to a different column, without also removing an agreement point from this column via pairing it with a conflict point in a different column. This implies that the initial imbalance in the rightmost column will remain no matter how many pairs of conflict and agreement point have been disposed off before player 1 reveals her signal corresponding to this column.

(1935) marriage theorem. 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. All women are equally acceptable for each man. Hall's theorem states that a necessary and sufficient condition to match or pair up the women to the men 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 among the women. For instance, if two or more women find one and the same man acceptable, it is impossible to match all these women.

To apply this theorem to a discrete type version of our model, let each conflict point $(x, y) \in \mathcal{Y}$ be identified as a "woman" and each agreement point $(x, y) \in \mathcal{B}$ as a "man". Because of the decentralized nature of the committee's information, in a fine subversion the committee can only match conflict and agreement points that lie in the same row or column. To reflect this constraint, for each $w \in \mathcal{Y}$, let $H(w) \subset \mathcal{B}$ be all those agreement points in $\mathcal{B}$ that lie in the same row or column as w.

Intuitively, each possible conflict point w is a point of concern for the observer whenever the committee recommends accepting the proposal. To satisfy the deniability constraint in a fine subversion, the committee has to justify its recommendations by matching or pooling each w with some element of $H(w)$; and do this for each conflict point $w \in \mathcal{Y}$. Hall's condition states that there be sufficient diversity of possible justifications, across all possible subsets of $\mathcal{Y}$. Hall's condition is necessary for a fine subversion. For if a fine subversion exists, it gives rise to a matching of each conflict point $w \in \mathcal{Y}$ to some agreement point $m \in H(w) \subset \mathcal{B}$. But then Hall's condition must be satisfied since otherwise we would violate Hall's theorem.

To rule out non-existence because of trivial failures of Hall's condition we impose two restrictions, that the sets $H(w)$ are always non-empty and that every element of $\mathcal{B}$ belong to some $H(w)$.[29] Under these two restrictions, Hall's condition implies, but is more demanding than, the total slack condition. Since the total slack condition is necessary and sufficient for the centralized expert to subvert, Hall's condition identifies extra constraints that must be satisfied for the committee to be able to subvert.

In the example in Figure 7(c), for the three conflict points in the right most column, the union of their acceptable agreement points are the two agreement points in that column. This imbalance is a violation of Hall's condition. It immediately rules out a fine subversion because it is impossible to match three conflict points with two agreement points and meet the deniability constraint. Combined with the additional feature described above, arising from the shape of $\mathcal{R}$ that makes it necessary for the committee to exchange precise information about one signal at

[29] Since in a problem with zero total slack, we could equally define $\mathcal{B}$ as the set of women and $\mathcal{Y}$ as the set of men, these two restrictions are essentially the same and mirror-images of each other. In terms of our figures, they say that every conflict point in $\mathcal{Y}$ has an agreement point in $\mathcal{B}$ in either the same row or the same column, and vice versa.

some stage of the conversation, this violation of Hall's condition also rules out the existence of a coarse subversion. Whenever player 1 reveals her signal corresponding to this column, as she must, there must be an imbalance between conflict and agreement points, making it impossible to satisfy the deniability constraint.

In the example of panel Figure 7(b), $\mathcal{R}$ has the same shape and so the committee faces the same coordination burden as in the example of panel (c). The key difference is that the observer also cares about both signals in panel (b), just like the committee. So he is willing to trade off bad news for one signal against good news about the other signal. Because of this concordance of preferences between the committee and the observer, there is greater diversity of possible justifications for each conflict point—some conflict points have acceptable agreement points in the same row and also in the same column. This diversity gives both players the freedom to reveal signals or to propose a decision, pooling conflict and agreement points across rows or across columns. In a suitably designed conversation, it lets each player hide unfavorable information and allow the other player to reveal favorable information that offsets the effect of her own bad information. Fine subversive conversations exist for the example in panel (b). It follows that Hall's condition must be satisfied in that example.

Although necessary for a fine subversion, Hall's condition is not a sufficient condition for the existence of a subversive conversation. Figure 7(d) illustrates this. In (d), Hall's condition is satisfied but subversion is impossible for the committee. To see why, notice first that since $\mathcal{R}$ is non-empty, no committee member can simply recommend an action at the outset and get his optimal decision. The committee needs to exchange at least some information in order to coordinate on their ideal decision. But anytime a player reveals a row or a column (or a subset of rows or columns), there is an imbalance between the agreement and conflict points in either that or the complementary part of the state space.

For instance, if player 1 starts the conversation by recommending acceptance for the right-most column, it is 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 in this part of the state space and there is negative slack. More generally, the deniability constraint must be violated either in the part of the state space that is revealed by the first move, or in the complementary part, no matter what this first move is and who makes it. This rules out the existence of a subversive conversation.[30]

Hall's condition is not sufficient because the committee holds decentralized information. To see this, perform the following thought experiment. Consider again the centralized expert who holds both pieces of information, but suppose now that she is constrained to match conflict

[30] Hall's condition is sufficient for existence if at most one player has more than two possible signals. So, if we employ the construction of Theorem 2 in discrete type versions of our model, it is necessary and sufficient to check Hall's condition on the residual state space(s) created by player 1's first move of pooling two signals. The same is true for the construction used for Proposition 1.

points with agreement points that lie in the same row or column, just like the committee is for a fine subversion. Call this incarnation of the centralized expert, the *constrained* centralized expert.

It is not difficult to see Hall's condition is necessary and sufficient for the constrained centralized expert to subvert. Since Hall's condition is satisfied in panel (d), the constrained centralized expert can subvert in that example, even though the committee cannot. Her advantage over the committee comes from the fact that she knows both pieces of information and so she can condition her messages on both. The constrained centralized expert can match a specific conflict point with an agreement point, while leaving other agreement points in the same row/column unmatched. For instance, in panel (d), she can match the conflict point in the middle column with the top agreement point in the same column, matching the remaining agreement point in that column with the conflict point that is in the middle row, and so on. This is possible only because she knows both pieces of information.

In contrast, in the case of the committee, whenever coordination requires an exact signal to be revealed, the player who reveals her signal cannot condition on information privately held by the other player. If a recommendation to accept the proposal is made subsequently, not only must conflict points in the revealed row or column be matched with the same number of agreement points in that row/column in order to satisfy the deniability constraint, but all other hitherto unmatched agreement points in the same row/column must also be eliminated from consideration at the same time. This wastes slack from the perspective of the deniability constraint.

To minimize this wastage, the committee has to engage in a conversation. A conversation is a process by which the committee members gradually gather more information in a carefully constructed sequence, safely matching conflict points with agreement points whenever they recommend a decision, wasting no more slack than what is available. The matching problem facing the committee is an inherently dynamic one. Such dynamic considerations are absent from the constrained centralized expert's problem. We postpone for future research the task of a general characterization of the dynamic matching process that is the problem of subversion by the committee.

5 Related Literature

The literature has not previously considered communication between informed players with the same preferences who want to share information with each other while also hiding information from a third party with different preferences. 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.

Chakraborty and Yilmaz (2017) consider a sequential 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 and they do not consider the possibility of subversion under public information exchange. However, the approach of Chakraborty and Yilmaz (2017) can be utilized to extend our results to cases of within-committee conflicts.

The key feature of subversive communication in our model – the necessity of a back and forth conversation – is novel to the literature. While a role for multiple rounds of communication arises in the literature on communication between two players with different preferences and one-sided private information (e.g., Forges, 1990; Aumann and Hart, 2003; Krishna and Morgan, 2004; and Chen, Goltsman, Horner, and Pavlov, 2017), this is based on jointly controlled lotteries. The back and forth conversations in our environment with two-sided private information are not reliant on such lotteries. Our model also differs from that of Meyer-ter-Vehn, Smith, and Bognar (2017) who model coarse communication in the form of repeated voting by two players in a debate setting with costly delay.

Since each expert is communicating to each other and also to the third party who may intervene, the paper is related to the literature on cheap talk with multiple audiences starting with Farrell and Gibbons (1989).[31] And since each expert has some information of relevance for the decision, it is related to the literature on communication to receivers with private information (e.g., Watson, 1996). Private information in sender-receiver games can sometimes allow for pooling of disjoint types (e.g., Guo and Shmaya, 2019; Harbaugh and To, 2020). Such pooling of good and bad news arises in our constructions for the distinct reason that it delays disclosure of information by one player until the other player has revealed more news that may put the bad news in a more favorable light.

Our approach belongs to the general literature on communication and information design. In many environments cheap talk can be persuasive in that it induces the receiver to take an action that is more favorable to the sender (Crawford and Sobel, 1982; Chakraborty and Harbaugh, 2010), even if it need not do as well as when the sender has commitment power (Kamenica and Gentzkow, 2011; Lipnowski and Ravid, 2019). Subversive cheap talk attains the ex post optimal, full information outcome for the experts. So whenever a subversive conversation exists, it is optimal for the senders under commitment. Since each expert must be truthful in order not to mislead the other expert, our subversive conversations are persuasive also in verifiable information settings where message spaces depend on types (e.g., Milgrom, 1981).

[31] In Kolb and Conitzer (2020) a third party monitors sender-receiver communications and takes actions to undermine the sender's credibility in a repeated game.

These results also relate to the literature on communication within organizations, and in particular to the question of whether authority should be delegated to lower level managers that have more information (e.g., Dessein, 2002; Alonso, Dessein, Matouschek, 2008). A potential concern is that if communication between lower level managers is not secure then the central authority has an ex post incentive to intervene, thereby making the promise of delegation less credible. In this paper the observer does equally well from delegation or from listening to the experts since the experts can still attain their best outcome even if their communications are learned by the observer. Hence the observer loses nothing from credibly delegating to the experts. However, this also raises the important open question of whether the observer can do even better by enforcing an alternative communication design, i.e., who can speak to whom with what messages when, that makes subversion difficult or impossible.

Glazer and Rubinstein (2004) consider optimal rules of persuasion, from the perspective of a single receiver facing a single sender, when the receiver can obtain a limited amount of evidence on his own. Our focus is on sender optimal outcomes, when information is decentralized among multiple senders, for the case of unverifiable information. 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 expert has private information, such pooling (and separation) must be designed not only to persuade the receiver but also to share enough information publicly and coordinate on the experts' ideal decisions.

Finally, our research is connected to the literature on pragmatics in logic, linguistics and philosophy. Grice (1967) introduces the *cooperative principle* for conversations between two players with common interests: do whatever is necessary to achieve the purpose of your talk and don't do anything that will frustrate that purpose. While we share the common interest assumption, we employ an explicit game theoretic model, unlike Grice.[32] The key driver of our results is the possibility that committee communication will be scrutinized by a wider public with different interests. It is because of this scrutiny that the committee must engage in conversations that dynamically pool and reveal information in order to achieve the twin goals of coordination and concealment. Absent the scrutiny, there is no need for a back and forth conversation.

6 Conclusion

This paper addresses a new problem in the sender-receiver game literature and in doing so provides a new understanding of the role of conversations. We show that the two experts can and must use a back and forth conversation to obtain their preferred policy in every state. Even if the conversation is public or is nominally private but leaked with some chance, the exact

[32] See Borgers (2002) for game theoretic arguments in favor of conversations, motivated by Grice's maxim of relevance, and based on the notion that sending and receiving messages are both costly, unlike what we assume.

reason for the decision remains uncertain. The experts thereby maintain plausible deniability that their recommendation was influenced by bias rather than just the facts.

When communication between two experts is public or might be made so, the process of communication matters. Subversiveness requires that some things can and must be said only at a suitable time and some things can never be said by some experts. During the conversation, each expert must create a context that allows the other expert to either take a decision or to provide further clarification that allows the first expert to provide more information. By sharing only what needs to be shared conditional on the information revealed so far, the needs to both coordinate and conceal are thereby reconciled. The experts can successfully determine and implement their ideal policy even when a fully-informed observer or the public would prefer a different policy.

7 Appendix

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 as follows

$$m_t = \begin{cases} z & \text{if } z \in \left[z_t^L, z_t^H\right], \\ \neg \left[z_t^L, z_t^H\right] & \text{otherwise;} \end{cases} \tag{1}$$

where $z = x$ if t is odd and $z = y$ if t is even. The interpretation is as follows. In an odd round t, when expert 1 speaks, he reveals x perfectly when $x \in \left[z_t^L, z_t^H\right]$ and otherwise reveals that x does not belong to the interval $\left[z_t^L, z_t^H\right]$. If expert 1 reveals x in round t, $m_t = x$, then expert 2 takes her (and the committee's) ex post optimal decision

$$m_{t+1} = \begin{cases} A & \text{if } (x, y) \in \mathcal{B} \cup \mathcal{Y}, \\ R & \text{otherwise;} \end{cases} \tag{2}$$

after taking into account the revealed value x and her private information y. If instead, expert 1 reveals that x does not belong to $\left[z_t^L, z_t^H\right]$ in round t, then the conversation moves to the next round where it is expert 2's turn to speak. The procedure is symmetric in an even round t when it is expert 2's turn to speak and $z = y$. When a expert whose turn it is to speak does not reveal her signal, indicating instead that it does not belong to the interval $\left[z_t^L, z_t^H\right]$, we will say that she passes the conversation to the other expert or simply "passes".

We will refer to the intervals $\left[z_t^L, z_t^H\right]$ as "cuts". They fully identify the conversation that we construct. To see how these cuts are constructed, let

$$T = \left\lceil \log_2 \left(1 + \frac{1}{1-2c} \right) \right\rceil . \tag{3}$$

Since $c > 0$, we must have $T \geq 2$. First, for $t = 1, ..., T-2$, let

$$z_t^L = \frac{1 - \left(2^t - 1\right)(1-2c)}{2} \text{ and } z_t^H = \frac{1 + \left(2^t - 1\right)(1-2c)}{2}. \tag{4}$$

We will refer to these cuts as the opening cuts and the corresponding phase of the conversation as the opening game. Next, set $z_0^L = 0$, and for $t = T-1$, let

$$z_{T-1}^L = \max \left\{ \frac{1 - \left(2^{T-1} - 1\right)(1-2c)}{2}, \frac{z_{T-2}^L}{2} \right\} \text{ and } z_{T-1}^H = 1 - z_{T-1}^L, \tag{5}$$

if $T-1$ is odd, and

$$z_{T-1}^L = 1 - z_{T-1}^H \text{ and } z_{T-1}^H = \min \left\{ \frac{1 + \left(2^{T-1} - 1\right)(1-2c)}{2}, \frac{1 - z_{T-2}^H}{2} \right\}, \tag{6}$$

if $T-1$ is even. Moreover, for $t = T$, let

$$z_T^L = z_{T-1}^L \text{ and } z_T^R = 1 - z_T^L. \tag{7}$$

We will refer to these cuts as midgame cuts and the corresponding phase of the conversation as the midgame. Finally, for $t > T$, let

$$z_t^L = \frac{z_{t-2}^L}{2}, z_t^H = \frac{1 + z_{t-2}^H}{2} \tag{8}$$

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) in accordance with (1). Then given a 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 $x \in [z_t^L, z_t^H]$

$$\begin{aligned}
& \Pr\left[x \geq y | m^t, m_{t+1} = A\right] \\
= \ & \Pr x \geq y | x, y \notin \left(z_{t-1}^L, z_{t-1}^H\right), x, x - y + c \geq 0] \\
= \ & \frac{x}{x + c - \left(z_{t-1}^H - z_{t-1}^L\right)} \\
\geq \ & \frac{z_t^L}{z_t^L + c - \left(z_{t-1}^H - z_{t-1}^L\right)} \\
= \ & \frac{1}{2},
\end{aligned}$$

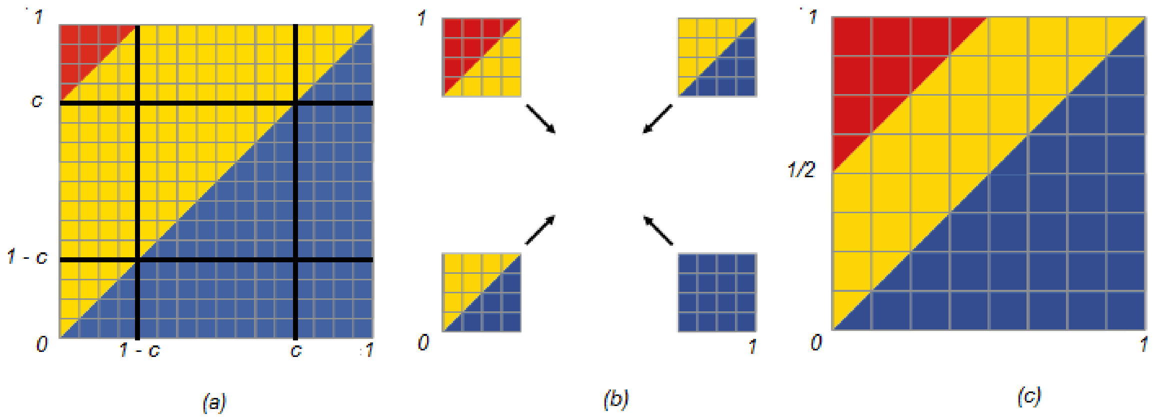


Figure 8: Opening rounds for high conflict

using (4). An identical argument applies for the case where y is revealed and expert 1 takes the decision. This establishes (DC) for all $t < T - 1$. The same argument also applies for $t = T - 1$, when the cuts for round $T - 1$ are given by the first term in the max/min defined in (5) and (6). Identical arguments also establish that the deniability constraint is satisfied for $t \geq T$ and we omit the details. This completes the proof of Case 1.

Case 2 ($c \geq 1/2$):

Figure 8(a) depicts the opening moves for the first two rounds when $c > 1/2$. In $t = 1$, expert 1 recommends accepting the proposal if $x \in [1 - c, c]$, without revealing the exact value of x. Conditional on this message, the zone of conflict $\mathcal{Y}$ is equal (in measure) to the zone of agreement $\mathcal{B}$. Such a recommendation will therefore satisfy the deniability constraint. If $x \notin [1 - c, c]$ then expert 1 passes in $t = 1$. Subsequently, expert 2 makes a recommendation accept in $t = 2$ without revealing her exact signal if $y \in [1 - c, c]$; otherwise she passes. As in $t = 1$, this recommendation also satisfies (DC) and implements the committee's ideal decision for these states.

Panel (b) depicts the residual state space after experts 1 and 2 pass in the first two rounds without taking a decision. The states $[1 - c, c] \times [0, 1]$ and $[0, 1] \times [1 - c, c]$ have been removed in panel (b) and it depicts the residual zones of agreement and conflict consisting of the four square-shaped areas depicted in the figure. For the continuation game, this residual state space is identical for strategic purposes with the state space where we paste these four squares together. This is depicted in panel (c). This residual state space is identical to the case where $c = 1/2$. Hence, it is without loss of generality to focus on the case $c = 1/2$ for the rest of our construction.

Figure 9 depicts the structure of a subversive conversation for the case $c = 1/2$. Panel (a) shows the state space $[0, 1]^2$, the zones of agreement $\mathcal{R}$ and $\mathcal{B}$ as well as the zone of conflict $\mathcal{Y}$. The remaining panels depict the structure of a subversive conversation that the committee can employ. The conversation starts with expert 1 revealing whether or not $x \in [1/4, 3/4]$ as

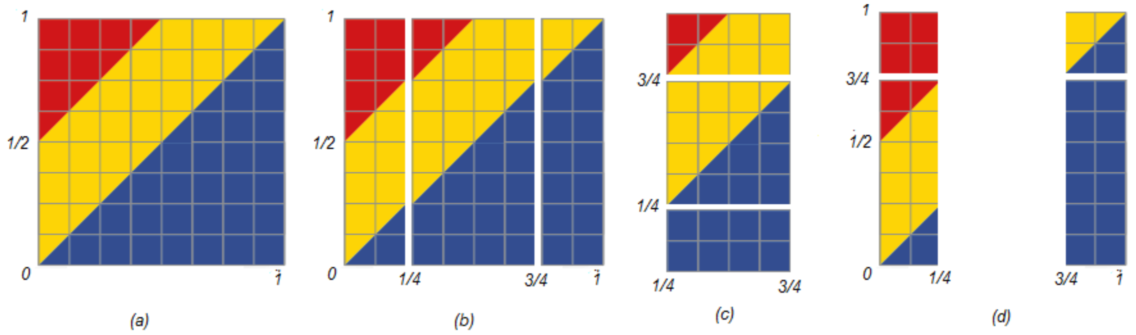


Figure 9: Subversion under large conflict

depicted in panel (b). If $x \in [1/4, 3/4]$ the conversation moves to panel (c) and otherwise it moves to panel (d).

Consider panel (c) first. If $y \in [1/4, 3/4]$, expert 2 recommends the proposal in $t = 2$. Since the sets $\mathcal{B}$ and $\mathcal{Y}$ are equal (in measure) conditional on the observed history of messages, the deniability constraint is satisfied. On the other hand, if $y \notin [1/4, 3/4]$, then expert 2 says so and we move to $t = 3$. In $t = 3$, expert 1 perfectly reveals his signal $x \in [1/4, 3/4]$. Subsequently, expert 2 recommends accepting the proposal if $(x, y) \in \mathcal{B} \cup \mathcal{Y}$ and the status quo otherwise. Neither recommendation will be overruled by the observer given the information revealed by the observed history. This information consists of the exact value of $x \in [1/4, 3/4]$ and the facts $y \notin [1/4, 3/4]$ and $(x, y) \in \mathcal{B} \cup \mathcal{Y}$ that the observer can deduce from the observed history. That the deniability constraint is satisfied can be seen from panel (c) by noting that the $\mathcal{Y}$ is at most at large as $\mathcal{B}$, for each value of x that can be reveled at this stage, conditional on states of the world that are still possible given the observed history.

If $x \notin [1/4, 3/4]$ then expert 1 says so in $t = 1$ and we move to the situation depicted in panel (d) where the states $[1/4, 3/4] \times [0, 1]$ have been deleted from the state space. In this case, in $t = 2$, expert 2 reveals the exact value of y if $y \leq 3/4$ and otherwise simply reveals $y > 3/4$ without revealing the exact value. When y is revealed and expert 1 proposes acceptance, the deniability constraint is satisfied as can be seen from the figure. If on the other hand $y > 3/4$ and expert 2 simply reveals this fact, then expert 1 recommends rejection if $x < 1/4$ and acceptance if $x > 3/4$. Once again, the deniability constraint is satisfied since the residual measure of $\mathcal{Y}$ is no larger than that of $\mathcal{B}$ conditional on the observed history. 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. In round 2, player 2 must make a recommendation. Note that player 1 cannot perfectly reveal $x \in [0, c)$ and satisfy the deniability constraint (DC). However, the coordination constraint (CC) cannot be satisfied in round 2 if player 2 does not know exactly the value of $x \in [0, \min\{c, 1 - c\})$, which is a positive measure set when $c < 1$, and thus we have a contradiction. ■

Proof of Theorem 2: The proof of part (i) follows from discussion in the text.

For the proof of part (ii), assume by way of contradiction that a universal, three-round subversive conversation exists, i.e., a 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.

In round 3, player 1 must make a recommendation. This means that player 2 must reveal enough information in round 2 to make subversion possible in all states of the world. Suppose $y = 1$. If player 1 is to satisfy (CC) 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 the recommendation.[33]

Consider the following set of histories, $H \subset \mathcal{H}_3^*$: 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 in round 3 says "Accept".

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 $H := \left\{m^t : m_1 \in M_1, m_2 = 1, m_3 = \text{A}\right\}$. The deniability constraint (DC) integrated over this set H must be satisfied by an admissible conversation $\int_{m^t \in H} \left\{\Pr[\mathcal{B}|m^t] - \Pr[\mathcal{Y}|m^t]\right\} \; d\nu\left(m^t\right) \geq 0$. 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 $(1,1) \in \mathcal{B}$ and thus $\int_{m^t \in H} \Pr[\mathcal{B}|m^t \in H] = 0$. For any $c > 1/2$, we have that $\int_{m^t \in H} \Pr[\mathcal{Y}|m^t \in H] \geq \Pr[x \in [1 - c, 1/2)] = c - 1/2 > 0$ and thus the deniability constraint fails, which is a contradiction. ■

Proof of Theorem 3: Follows from the discussion in the text. ■

Proof of Propositions 1–3: Follow from the discussion in the text. ■

References

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

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

[33] More generally, the argument works similarly for a positive measure set of y, if subversion was to be defined as the players achieving their first-best outcome almost everywhere.

[3] Baccara, Mariagiovanna and Heski Bar-Isaac. 2008. “How to Organize Crime,” *The Review of Economic Studies*, 75(4): 1039-–1067.

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

[5] Borgers, Tilman (2002). Comment on *Economics and Language* in same by Ariel Rubinstein, Cambridge University Press, Cambridge, U.K.

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

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

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

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

[10] Crawford, Vincent P., 2003. “Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions,” *American Economic Review*, 93(1): 133–149.

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

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

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

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

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

[16] Gordon, Sidartha. 2010. “On Infinite Cheap Talk Equilibria,” working paper.

[17] Grice, Paul H. (1967), “Logic and Conversation,” reprinted in *Studies in the Way of Words*, Paul Grice (ed.), Harvard University Press, Cambridge, MA, 1989.

[18] Guo, Yingni, and Eran Shmaya. 2019. “The Interval Structure of Optimal Disclosure,” *Econometrica*, 87(2): 653–675.

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

[20] Harbaugh, Richmond and Theodore To. 2020. “False Modesty: When Disclosing Good News Looks Bad,” *Journal of Mathematical Economics*, 83: 43–55.

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

[22] Kolb, Aaron and Vincent Conitzer. 2020. “Crying about a Strategic Wolf: A Theory of Crime and Warning,” *Journal of Economic Theory*, in press.

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

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

[25] Lipnowski, Elliott and Doron Ravid. 2019. “Cheap Talk with Transparent Motives,” working paper.

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

[27] Morgan, J. and Phillip C. Stocken. 2003. “An Analysis of Stock Recommendations,” *RAND Journal of Economics*, 34(1): 183–203.

[28] Padlipsky, P.A., D.W. Snow, and P.A. Karger. 1978. “Limitations of End-to-End Encryption in Secure Computer Networks,” Project No. 672B, prepared for Deputy for Technical Operations, Electronic Systems Division, Air Force Systems Command, USAF, Hanacom Air Force Base, Massachusetts.

[29] Shannon, Claude. 1949. “Communication Theory of Secrecy Systems,” *Bell System Technical Journal*, 28(4): 662–715.

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