Mediated Talk

Andreas Blume* Ernest K. Lai† Wooyoung Lim‡

August 14, 2024

Communicating through a third party, a mediator, can shield an information provider (sender) from unwelcome inferences made by an information user (receiver): the mediator can garble the information he receives from the sender before passing it on to the receiver. This makes it possible for the sender to influence the beliefs of the receiver without those beliefs becoming too extreme.

Consider the following example of a sender-receiver game: A sender (she) with private information about the state of the world communicates with a receiver (he) who takes a payoff relevant action. There are two equally likely states of the world, s_1 and s_2. Payoffs depend on the state of the world and the receiver's action $a \in \{a_1, a_2, a_3, a_4, a_5\}$ as indicated in Figure 1: the first entry in each cell is the sender's payoff and the second entry the receiver's payoff from that state-action pair. The sender privately learns the state of the world, while the receiver remains uninformed.

	a_1	a_2	a_3	a_4	a_5
s_1	0,10	1,0	5,9	3,4	2,7
s_2	1,0	0,10	3,4	5,9	2,7

Figure 1: Benefit of garbling information

Without communication, it is uniquely optimal for the receiver to take action a_5 for an expected payoff of 7. The sender's payoff in this case is 2.

Suppose instead that the sender can talk to the receiver: After privately learning the state, the sender sends a message $m \in \{m_1, m_2\}$ to the receiver. Doing so might provide the receiver with useful information. If, for example,

*Department of Economics, University of Arizona. *Email:* ablume@arizona.edu

†Department of Economics, Lehigh University. *Email:* kwl409@lehigh.edu

‡Department of Economics, The Hong Kong University of Science and Technology. *Email:* wooyoung@ust.hk

the sender sent message m_1 in state s_1 and message m_2 in state s_2, the receiver could perfectly infer the state and take action a_i in state s_i for a (receiver) payoff of 10. This strategy profile, however, is not incentive compatible for the sender: she would have an incentive to misreport the state to receive a payoff of 1 instead of 0.

What if instead the sender only provided partial information? Suppose, for example, the sender sent message m_1 with probability 2/3 in state s_1 and sent message m_2 with probability 2/3 in state s_2. Then, upon receiving message m_1 the receiver would assign probability 2/3 to state s_1 and upon receiving message m_2 the receiver would assign probability 2/3 to state s_2. The receiver's unique best reply to message m_1 would be action a_3 and to message m_2 it would be action a_4. Both sender and receiver would be strictly better off than without communication.

While the provision of partial information just described is (in expectation) beneficial to both parties, it also suffers from not being incentive compatible: given the receiver's strategy, the sender would strictly prefer to send message m_1 in state s_1 and to send message m_2 in state s_2 rather than to randomize.

This is where a nonstrategic mediator can help. Suppose the mediator can be trusted to pass on any received message with probability 1/3 and otherwise to send each of the two messages with equal probability. The effect is that upon receiving message m_i the mediator will send message m_i to the receiver with probability 2/3. If, in addition, the sender sends message m_i to the mediator when the state is s_i, the receiver will believe that the probability of state s_i is 2/3 upon receiving message m_i and will find it optimal to behave exactly as in the case of partial information provision. The difference is that we have solved the sender's incentive compatibility problem. The sender, given the mediation rule and the receiver's strategy, finds it to be strictly optimal to send message m_i in state s_i. In this example, being able to rely on a mediator is akin to the sender being able to commit to a randomized strategy.

Blume, Lai and Lim (2023) compare direct with mediated communication in a laboratory experiment. They use an incentive structure that rules out influential communication with direct talk while permitting influential efficiency-improving communication with mediated talk, similar to the above example. They observe a significant difference of behavior in mediated versus direct talk. Whereas with direct talk senders eventually settle down on predominantly sending a message that is independent of the state, with mediated talk modal sender behavior is to send different messages for different states.

The potential role of using garbling to improve information transmission was recognized by Warner (1965) in the context of survey design. He was

interested in eliciting sensitive information (about drug use, tax compliance, employee theft, etc). His idea was to let responders condition their answers on a privately observed random event (e.g, the outcome of die roll) rather than relying on direct reporting about a sensitive trait.

Suppose a survey respondent either has a "stigmatizing trait" s or a "regular trait" r. With direct questioning the respondent may be unwilling to answer truthfully if asked "Do you have trait s?" Now suppose that instead of direct questioning the respondent is given a six-sided die to roll privately (or another comparable randomizing device) and instructed to answer the question "Do you have trait s?" whenever the die shows one of the numbers $1, \ldots, 4$ and otherwise to answer the question "Do you have trait t?" Warner reasoned that this provides some privacy protection to the survey respondent. Even if the responder is always truthful, a "yes" answer is no clear indication of the responder having the stigmatizing trait. If this privacy protection is sufficient to induce truth telling by survey respondents, it becomes possible to obtain reliable estimates of the prevalence of a sensitive trait in a population. Versions of this procedure are known as the randomized response technique (RRT).

Ljungqvist (1993) formalizes Warner (1965)'s survey design problem as a mechanism design question. Importantly, he postulates a payoff function for survey respondents that includes a truth-telling incentive: while survey respondents have payoffs that are decreasing in the audience's belief that they have the stigmatizing trait, they also receive a psychological reward whenever they tell the truth. If the randomizing device is such that conditional on truth telling the receiver's posterior moves only a little in response to an answer, the truth-telling reward will be greater than the loss from the shift in the audience's beliefs. The mechanism design problem then reduces to maximizing the informativeness of the procedure subject to satisfying a truth-telling constraint.

Blume, Lai and Lim (2019) examine Warner's idea in a laboratory experiment using a payoff function that is inspired by Ljungqvist (1993). To give Warner's randomized response technique its best shot at making a difference, Blume et al. (2019) provide experimental subjects with monetary rewards for truth telling. The rewards are designed to be insufficient to induce truth-telling with direct questioning but make truth-telling incentive compatible with RRT-garbling.

As in Blume et al. (2023), Blume et al. (2019) find that garbling does change behavior in the direction predicted by theory. Experimental subjects are more willing to give truthful answers with garbling than with direct questioning. Blume et al. (2019) do point out however that the RRT environment has multiple equilibria and that observed behavior is more consistent with partial truth-telling equilibria than complete truth-telling equilibria. This

undermines the ability to make inferences about population proportions that are based on the assumption of RRT inducing truth-telling.

In the theory literature, the observation that mediators can be used to profitably garble information transmission has a long tradition. Forges (1985) gives an example of a sender-receiver game that has no influential equilibria with direct talk, while there are efficiency improving influential equilibria with mediated talk. Myerson (1991) illustrates the effect of communicating through a noisy channel with an example that is accompanied by a story about sending messenger pigeons. In the example, the sender has two options. She can either send or not send a messenger pigeon, each of which counts as a different message. If no pigeon is sent, none is received and if one is sent it is not received with probability 1/2, owing to the perils of the journey. A key consequence is that if no pigeon arrives the receiver is uncertain about whether or not one has been sent: having not send a pigeon, the sender can plausibly deny that this was the case since the pigeon could have been sent but got lost. As in Forges's example, there is no influential equilibrium in the game in which pigeons always arrive, while there is an efficiency improving influential equilibrium if sent pigeons fail to arrive half of the time.

Blume, Board and Kawamura (2007) demonstrate that the effect noted by Forges and Myerson extends to the class of sender-receiver games analyzed by Crawford and Sobel (1982). They modify the direct-talk game of CS by letting the sender communicate through a noisy channel. With some probability the channel lets messages pass through unchanged. Otherwise, it passes on a message that is drawn from a uniform distribution over the message space. Blume et al. (2007) show that for almost every level of conflict for which there is influential communication in CS, there exists a level of noise and a corresponding equilibrium that improves on the best CS equilibrium. Furthermore, with the appropriate level of noise one can obtain the efficiency bound from using mediated talk in this environment that was established by Goltsman, Hörner, Pavlov and Squintani (2009).

The structure of optimal equilibria in Blume et al. (2007) has much in common with the equilibrium structure in Myerson's example (and also in the experimental implementation in Blume et al. (2023)). Messages sent by high types are sometimes replaced by messages that only low types would voluntarily send. This induces more favorable receiver responses to low types' messages, making it attractive for a larger set of types to send those messages. This percolates through the entire equilibrium structure and reduces senders' incentives to strategically distort their messaging behavior.

As we noted earlier, employing a nonstrategic mediator to garble message is a form of commitment. The Bayesian persuasion literature (Kamenica and Gentzkow, 2011) postulates a stronger form of a commitment: the uninformed sender publicly commits to an *information structure* that maps

the states of the world into distributions over signals, or, to use the sender-receiver-game language, prior to observing the state of the world the sender commits to a strategy that maps states into distributions over messages. The stronger variety of commitment afforded by being able to commit to a strategy prior to observing the state of the world rather than having to rely on a mediator (sometimes strictly) increases the sender's maximal equilibrium payoff.

Salamanca (2021) compares mediated talk with direct talk and Bayesian persuasion, focussing on ex ante sender-optimal equilibria. He notes that incentive compatibility requirements become more onerous as one moves from Bayesian persuasion, to mediated, and finally to direct talk. As a result, there are instances of incentive structures for which the sender's maximal ex ante equilibrium payoff from mediation lies strictly above the maximal payoff from direct talk and strictly below the maximal payoff from Bayesian persuasion.

Fréchette, Lizzeri and Perego (2022) compare the value to the sender from various degrees of being able to commit to a garbling scheme in the lab. With partial commitment, there is positive probability that the sender can surreptitiously revise an initially chosen information structure. When messages are not verifiable (the case considered here), such revision amounts to replacing the signal from the information structure by a cheap talk message as in Min (2021) and Lipnowski, Ravid and Shishkin (2022). At the extremes this environment includes direct talk and Bayesian persuasion. Fréchette et al. (2022) find that with non-verifiable messages increasing commitment raises the amount of information transmitted, consistent with theory. Departing from the theoretical prediction, they also find that a fraction of senders is "commitment blind" and behave as if they are unable to commit.

Some recent experimental work on mediation goes beyond the sender-receiver paradigm. Casella, Friedman and Perez Archila (2020) experimentally investigate the impact of garbling on conflict resolution. Their experiment takes the model of Hörner, Morelli and Squintani (2015) to the lab. They find that, consistent with theory, mediation improves truth-telling but, contrary to theory, does not raise the probability of peaceful resolution. Chassang and Zehnder (2019), building on a model of Chassang and Padró i Miquel (2019), conduct an experiment on whistleblowing in organization in which they compare direct questioning, Warner's randomized response technique and communication via a non-strategic mediator. In line with theory, they find that mediation improves information transmission. Randomized response, which in their environment is predicted to be ineffective, does better than expected.

References

Blume, Andreas, Ernest K Lai, and Wooyoung Lim, "Eliciting private information with noise: the case of randomized response," *Games and Economic Behavior*, 2019, *113*, 356–380.

_ , _ , and _ , "Mediated talk: An experiment," *Journal of Economic Theory*, 2023, *208*, 105593.

_ , Oliver J Board, and Kohei Kawamura, "Noisy talk," *Theoretical Economics*, 2007, *2* (4), 395–440.

Casella, Alessandra, Evan Friedman, and Manuel Perez Archila, "Mediating conflict in the lab," Technical Report, National Bureau of Economic Research 2020.

Chassang, Sylvain and Christian Zehnder, "Secure Survey Design in Organizations: Theory and Experiments," Technical Report, National Bureau of Economic Research 2019.

_ and Gerard Padró i Miquel, "Crime, intimidation, and whistleblowing: A theory of inference from unverifiable reports," *The Review of Economic Studies*, 2019, *86* (6), 2530–2553.

Crawford, Vincent P and Joel Sobel, "Strategic information transmission," *Econometrica: Journal of the Econometric Society*, 1982, pp. 1431–1451.

Forges, Françoise, "Correlated equilibria in a class of repeated games with incomplete information," *International Journal of Game Theory*, 1985, *14*, 129–149.

Fréchette, Guillaume R, Alessandro Lizzeri, and Jacopo Perego, "Rules and commitment in communication: An experimental analysis," *Econometrica*, 2022, *90* (5), 2283–2318.

Goltsman, Maria, Johannes Hörner, Gregory Pavlov, and Francesco Squintani, "Mediation, arbitration and negotiation," *Journal of Economic Theory*, 2009, *144* (4), 1397–1420.

Hörner, Johannes, Massimo Morelli, and Francesco Squintani, "Mediation and peace," *The Review of Economic Studies*, 2015, *82* (4), 1483–1501.

Kamenica, Emir and Matthew Gentzkow, "Bayesian persuasion," *American Economic Review*, 2011, *101* (6), 2590–2615.

Lipnowski, Elliot, Doron Ravid, and Denis Shishkin, "Persuasion via weak institutions," *Journal of Political Economy*, 2022, *130* (10), 2705–2730.

Ljungqvist, Lars, "A unified approach to measures of privacy in randomized response models: A utilitarian perspective," *Journal of the American Statistical Association*, 1993, *88* (421), 97–103.

Min, Daehong, "Bayesian persuasion under partial commitment," *Economic Theory*, 2021, *72* (3), 743–764.

Myerson, Roger B, *Game theory*, Harvard university press, 1991.

Salamanca, Andrés, "The value of mediated communication," *Journal of Economic Theory*, 2021, *192*, 105191.

Warner, Stanley L, "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, 1965, *60* (309), 63–69.