*Research Reports*
Publications of the Helsinki Center of Economic Research, No. 2019:2
*Dissertationes Oeconomicae*

MICHELE CRESCENZI

# ESSAYS ON
# INFORMATION AND KNOWLEDGE IN GAME THEORY

ISSN 2323-9786 (print)
ISSN 2323-9794 (online)

ISBN 978-952-10-8750-9 (print)
ISBN 978-952-10-8751-6 (online)

Doctoral dissertation, to be presented for public examination with the permission of the Faculty of Social Sciences of the University of Helsinki, in the Lecture Room, Economicum, Arkadiankatu 7, on the 16th of August, 2019 at 12 o'clock.

# Abstract

This dissertation consists of three essays on information and interactive knowledge in game theory. In the first essay, we study how a consensus emerges in a finite population of rational individuals who are asymmetrically informed about the realization of the true state of the world. Agents observe a private signal about the state and then start exchanging messages. Generalizing previous models of rational dialogues, we dispense with the standard assumption that the state space is either finite or a probability space. We show that a class of rational dialogues can be found that always lead to consensus provided that three main conditions are met. First, everybody must be able to send messages to everybody else, either directly or indirectly. Second, communication must be reciprocal. Finally, agents need to have the opportunity to participate in dialogues of transfinite length.

In the second essay, we provide a syntactic construction of correlated equilibrium. For any finite game, we study how players coordinate their play on a signal by means of a public strategy whose instructions are expressed in some natural language. Language can be ambiguous in that different players may assign different truth values to the very same formula in the same state of the world. We show that, absent any ambiguity, self-enforcing coordination always induces a correlated equilibrium of the underlying game. When language ambiguity is allowed, self-enforcing coordination strategies induce subjective correlated equilibria. Our analysis provides a justification for heterogeneous beliefs in strategic play.

In the final essay, we study the problem of a Sender who wants to persuade a two-member committee to take a certain action. Contrary to previous models, we assume that the Sender is uncertain about committee members' preference parameters. We provide a full characterization of the Sender's optimal persuasion strategy in two different contexts. In the first case, the Sender is allowed to elicit information by asking committee members to report their preference types. In the second, the Sender is not allow to do so. We show how the Sender's optimal persuasion strategy depends on the prior probability distribution over preference types. If the prior is informative enough, the Sender may find it optimal to persuade only a strict subset of type profiles. Finally, we show that uncertainty always entails a loss to the Sender with respect to the benchmark case with commonly known preferences.

# Acknowledgments

I owe many thanks to the people who have helped in the preparation of this dissertation. My supervisor Hannu Vartiainen has played an indispensable role throughout the entire process. During many conversations over the last five years, I have had the opportunity to appreciate and benefit form his generosity, curiosity, and deep knowledge of economic theory. He has helped me to think better, more clearly, and he has always been encouraging along the way. I am grateful for how much I have learned from him. I wish to thank the preliminary examiners Hannu Salonen and Mark Voorneveld. Their careful reading of the manuscript and their detailed comments and suggestions have significantly improved this dissertation. I also thank Mark Voorneveld for accepting with enthusiasm to act as the opponent in my public examination. Many thanks to Klaus Kultti for reading my work, providing helpful feedback on it, and for several enjoyable conversations on academic and non-academic matters. I wish to thank Juuso Välimäki for having introduced me to the literature on Bayesian persuasion and for his constructive comments on my work. I also thank the discussants and participants in the workshops of the Finnish Doctoral Programme in Economics and in several other conferences in Finland and abroad for their suggestions and thought-provoking questions.

Venturing into the intellectual inquiries of academic research is impossible without the prosaic comfort of money. I thank the Faculty of Social Sciences of the University of Helsinki and the OP Group Research Foundation for generously funding my doctoral studies. I also benefited from the Chancellor's travel grant and travel grants from the Doctoral School in Humanities and Social Sciences.

Finally, I wish to thank all my friends and colleagues at Economicum, especially Olena Izhak, Sara Yi Zheng, Yin Ming, and Min Zhu. Their kindness and laid-back attitude have made Economicum a great place where to do research.

*Helsinki, June 2019*

# Contents

# Chapter 1

# Introduction

Information and knowledge are foundational concepts in game theory. Every game involves interactive reasoning. Indeed, the game is typically assumed to be common knowledge. This means that everybody knows the game which is being played, everybody knows that everybody knows the game which is being played, and so on. No matter what the adopted solution concept is, the very fact that a game is being played implies that there is an infinite hierarchy of propositions about states of knowledge. Furthermore, a game is made of many components, and knowledge may refer to each and every aspect that is included in the description of the game. For instance, a player may know who her opponents are, but she may not know what the set of available actions of, say, player $i$ is. What a player knows or does not know is a function of her information.

This dissertation contributes to the understanding of the interplay between knowledge and information in game theory and, more generally, in interactive rationality. It consists of three self-contained chapters. The first chapter examines how common knowledge can be acquired through communication and how it leads to consensus. The second chapter studies how ambiguity in natural language induces differential information in games. The third paper examines how information can be selected and manipulated for strategic motives. The first two chapters belong to the area called interactive epistemology, where interactive reasoning about knowledge and beliefs are at the center of stage. The last chapter belongs to the area of information design, where the strategic manipulation of information is examined. In the subsequent section, we give a brief overview of the theoretical frameworks in these areas within which we conduct our analysis.

## 1.1 The framework

### 1.1.1 Interactive epistemology

There are two main frameworks to represent interactive knowledge and beliefs: the event-based model and the syntactic model. We use the first in Chapter 2, and the latter in Chapter 3. We now give a brief overview of these models. The material we are about to discuss is standard and is adapted from Fagin et al. (2004) and Maschler et al. (2013). The standard framework in game theory for modeling interactive knowledge is the *event-based* model introduced by Aumann (1976). The model has a simple structure and it consists of two elements: a state space $\Omega$ and a profile of information partitions $(H_i)_{i\in I}$, where $I$ is the set of players. The state space is a set containing the possible states. A state is a complete description of all the relevant aspects of the world. Information partitions represent players' information about the prevailing state of the world. If two states belong to the same partition cell, then the player cannot distinguish these two states. Differently put, she does not have enough information to distinguish the occurrence of one world from the occurrence of the other. Players reason about events, which are subsets of the state space. We say that $i$ knows $E \subseteq \Omega$ in state $\omega$ if $H_i(\omega) \subseteq E$, where $H_i(\omega)$ is the cell of the information partition $H_i$ containing state $\omega$. In words, $i$ knows $E$ if $E$ occurs in each and every state that $i$ considers as possible based on the information she has at $\omega$. For each player, one can thus define a knowledge operator $K_i : 2^\Omega \longrightarrow 2^\Omega$ such that $K_iE := \{\omega \in \Omega : H_i(\omega) \subseteq E\}$. In words, the event $K_iE$ stands for "$i$ knows that $E$" and is the (possibly empty) subset of states where $i$ knows that $E$.

It is a standard requirement that the knowledge operator $K_i$ satisfies the following properties, also known as *S5 System*:

1. $K_i\Omega = \Omega$: the player knows what the state space is. This also captures the fact that the player is logically omniscient, i.e. she knows all the theories (or tautologies) in the system.

2. $K_iE \cap K_iF = K_i(E \cap F)$: knowing $E$ and knowing $F$ is the same as knowing the conjunction $E$ and $F$.

3. $K_iE \subseteq E$: this is called the axiom of knowledge. It says that agents can only know events that are true.

4. $K_i(K_iE) = K_iE$: this is called the axiom of positive introspection. It says that, if $i$ knows $E$, then she also knows that she knows $E$.

5. $(K_iE)^c = K_i((K_iE)^c)$, where the superscript denotes the set-theoretic complement in $\Omega$. This is called the axiom of negative introspection. It says that, if $i$ does not know $E$, then she also knows that she does not know $E$.

Since the definition of knowledge is information-based, there is a close connection between properties of knowledge and properties of information. More specifically, for any information partition, the corresponding knowledge operator satisfies the S5 system above. In addition, one can show that, for any operator satisfying the S5 system, there exists a partition that induces that operator.

The model presented so far allows us to talk about higher-order and interactive knowledge in a natural way. For instance, the event that $i$ knows that $j$ knows $E$ is $K_iK_jE$; $i$ knows that $j$ knows that $i$ knows that $j$ knows $E$ is expressed as $K_iK_jK_iK_jE$; and so on. Importantly, one can construct arbitrarily long chains describing interactive reasoning of any order. Therefore, the concept of common knowledge is well defined. We say that $E$ is common knowledge in state $\omega$ if, for *every* finite sequence of players $i_1, \ldots, i_j$, we have that $\omega \in K_{i_1}K_{i_2} \cdots K_{i_{j-1}}K_{i_j}E$.

The above definition captures our intuition that common knowledge presupposes an infinite sequence of statements: everybody knows $E$, everybody knows that everybody knows $E$, and so on. But this appeal to our intuition is also a weakness because one needs to check infinitely many objects to assess whether a certain event is common knowledge. An equivalent, yet more compact, representation is provided by Aumann (1976). Let $M$ be the meet, i.e. the finest common coarsening, of the information partitions $(H_i)_{i\in I}$. Then the event $E$ is common knowledge at $\omega$ if and only if $M(\omega) \subseteq E$, where $M(\omega)$ is the cell of the meet containing $\omega$.

One can use the event-based framework to represent not only knowledge but also beliefs. The model is the same as in the case of knowledge with the proviso that $\Omega$ is now required to be a probability space. Then one can define the belief operator as $B_iE := \{\omega \in \Omega : \mu(E|H_i(\omega)) = 1\}$, where $\mu$ is a probability measure over $\Omega$. The event $B_iE$ stands for "$i$ believes that $E$". The interpretation is that $B_iE$ contains every state of the world where, based on the information that $i$ has at that state, she ascribes probability 1 to the event $E$. The belief operator shares all the $S5$ properties of knowledge except for the axiom of knowledge. That is, it is not necessarily true that $B_iE \subseteq E$. This means that, while people can only know true facts, they may believe in events that turn out to be false. Similarly to the case of knowledge, we can talk about higher-order and interactive beliefs in a natural way. In particular, one can construct arbitrarily long chains of events, called belief hierarchies, which describe beliefs, beliefs about beliefs, beliefs about beliefs about beliefs, and so on. As is usually done in applications, these belief hierarchies can be equivalently represented in

type spaces *à la* Harsanyi. The reason is that belief hierarchies are rather complex objects to work with. Instead of describing belief hierarchies in full detail, Harsanyi's idea is to describe them implicitly using a more elementary set of types. For instance, in Chapter 4 we use a finite set of types to represent the (infinitely long) belief hierarchies that are relevant to our analysis.

The *syntactic* approach to knowledge and beliefs is the standard framework in fields like logic, computer science, and philosophy. The fundamental component of the model is a set of primitive propositions $\Phi$. Then a language is formed by taking primitive propositions and closing off under negation, conjunction, and modal operators $K_1, \ldots, K_n$. In this case, the argument of $K_i$ is a formula and not an event. Intuitively, the language contains sentences through which agents reason about the world. The truth value of each formula is determined by a semantic model. The most common semantics are Kripke structures. A Kripke structure consists of a state space, a profile of information partitions, and, contrary to the event-based approach, an interpretation function $\pi : \Omega \times \Phi \longrightarrow \{true, false\}$. The latter allows us to determine whether any given primitive proposition is true or false at any given state of the world. By structural induction, the assignment of truth values can be extended to any other non-primitive formula in the language. To express common knowledge, one needs to augment the language with the operator $CK$, which stands for "it is common knowledge that". To express beliefs, one needs to augment the language with probability formulas, i.e. sentences that allow players to use probabilities in their reasoning. In addition, the state space needs to be a probability space.

The event-based and the syntactic models are two distinct representations of interactive knowledge. These representations are essentially equivalent. More specifically, for any syntactic model there exists an event-based model such that any formula in the former is true if and only if the corresponding event in the latter holds. Conversely, for any event-based model, one can always construct a syntactic model such that an event in the former holds if and only if the corresponding formula in the latter is true. However, there is a sense in which the syntactic model is a richer framework than the event-based model is. The richness lies in the fact that a formal language is part of the model. This allows us to talk formally and explicitly about players' reasoning.

### 1.1.2 Information design

In the model of interactive knowledge and beliefs that we have just introduced, information is taken as a given. More specifically, agents are endowed with an initial stock of information, hence knowledge, the origin of which is left out of the model. The recent literature on

information design examines how information can be strategically acquired and exchanged when potentially conflicting interests are present.

It is convenient to introduce information design by making a comparison with the classical literature on mechanism design. In the latter, one typically asks the following question: Given an economic environment, and given a certain distribution of information, what are the rules of the game that allow us to achieve a certain distribution of outcomes? In information design, the starting point is different. Given an economic environment, and given the rules of the game, what are the information structures that allow us to attain a certain distribution of outcomes? While both approaches seek to understand how social outcomes can be attained, they differ in what the designer, or planner, is allowed to do. In mechanism design, the designer's choice variable is a game form; in information design, it is an information structure.

We now introduce the basic framework for studying information design. The literature on this topic was initiated by Kamenica and Gentzkow (2011) and Bergemann and Morris (2016a). The material in this subsection is standard and is adapted from Bergemann and Morris (2019) and Taneva (2019). The fundamental object is a finite game of incomplete information, which we represent as a pair $(G, S)$. The first component describes the so-called payoff structure of the game, namely the set of agents $I$, the profile of available action sets $(A_i)_{i\in I}$, a set of states $\Theta$, and payoff functions $u_i : A \times \Theta \longrightarrow \mathbb{R}$, where $A = \times_{i\in I} A_i$. Players share a common prior $\mu$ over $\Theta$. The component $S$ describes the information structure of the game. More specifically, it includes a profile of signal realizations $(T_i)_{i\in I}$ and a function $\pi : \Theta \longrightarrow \Delta(T)$, where $\Delta(T)$ is the set of probability distributions over $T = \times_{i\in I} T_i$. Intuitively, $\Theta$ is the set containing the payoff-relevant parameters about which players are uncertain. At the ex-ante stage, their information about the state is represented by a common prior over this set. At the interim stage, each agent receives information about the true state by means of the information structure. Once the true state has been determined by nature, each player $i$ observes a signal in $T_i$. The probability with which profiles of signals are observed as a function of the true state is captured by the map $\pi$.

Absent any design problem, the above representation describes a standard game of incomplete information. Now suppose that, for a fixed $G$, a designer wants to choose the information structure so as to induce a particular outcome distribution. The designer's behavior is represented by a decision rule $\sigma : T \times \Theta \longrightarrow \Delta(A)$. In words, a decision rule sends recommendations on how to play the game that are contingent on the true state of the world and the profile of signal realizations. Each player observes her action recommendation $a_i$ privately, and the designer knows both the true state of the world and which signals are being observed. Clearly, players might have the incentive to disobey the designer's recommendations. Therefore, one needs to identify the set of decision rules so that nobody has

such an incentive. Formally, we say that a decision rule $\sigma$ is obedient if, for every player $i$, every $t_i \in T_i$, and every $a_i, a_i' \in A_i$, we have that

$$\begin{aligned} \sum_{a_{-i},t_{-i},\theta} u_i((a_i, a_{-i}), \theta)\sigma(a|t,\theta)\pi(t|\theta)\mu(\theta) \geq \\ \sum_{a_{-i},t_{-i},\theta} u_i((a_i', a_{-i}), \theta)\sigma(a|t,\theta)\pi(t|\theta)\mu(\theta). \end{aligned} \tag{1.1}$$

Every obedient decision rule is a Bayes Correlated Equilibrium as introduced in Bergemann and Morris (2016a). They show that this solution concept is a superset of all the main notions of correlated equilibrium for games with incomplete information considered in Forges (1993, 2006). The reason why this is the case is that, in a Bayes Correlated Equilibrium, the designer can condition her action recommendations on both the true state of the world and the actual signal realizations that players observe.

The set of obedient decision rules identifies the set of implementable allocations. The task of designing information thus amounts to choosing the decision rule that the designer prefers among all the obedient ones. As (Bergemann and Morris, 2016a, Theorem 1) show, any obedient rule implicitly defines an information structure for which there exists a Bayesian Nash Equilibrium of the underlying game that induces the same outcome distribution as the chosen rule. In case of multiple equilibria, it is standard to assume that the players coordinate over the equilibrium that the designer has selected. More stringent solution concepts can be used instead of Bayes Correlated Equilibrium. For example, if the designer cannot condition her recommendations on the true signals $t_i$, she can elicit that information from players. Since constraints in (1.1) guarantee obedience only, additional constraints need to be imposed on $\sigma$ so as to guarantee truthful reporting. Irrespective of the solution concept, we emphasize that information design is always a two-step procedure. First, one identifies the set of implementable allocations through Bayes Correlated Equilibrium or more restrictive solution concepts. Second, one chooses the implementable allocation(s) that maximizes the designer's objective function.

## 1.2 Contribution

In this section we give a brief overview of the research questions we address and motivate their relevance. The first chapter of this thesis contributes to the literature on Aumann's agreement theorem and common knowledge acquisition through communication. The seminal papers are Aumann (1976) and Geanakoplos and Polemarchakis (1982), respectively. Aumann's agreement theorem says that rational people sharing a common prior cannot

agree to disagree. More specifically, if their posterior beliefs about a certain event are common knowledge, then those beliefs must be the same. The question arises as to how to attain a state of affairs where common knowledge holds. The insight put forward by Geanakoplos and Polemarchakis (1982) is that common knowledge can be attained through communication. If people announce their beliefs, and concurrently update their information in light of others' announcements, then a consensus will eventually emerge. That is, beliefs will become common knowledge and therefore they will be the same for every agent.

The above results have been generalized along several dimensions. Bacharach (1985) was the first to show that Aumann's agreement theorem is an extremely general result. Not only does it hold for beliefs, but it also holds for any function that maps information sets to messages and that satisfies the sure thing principle. In addition, the theorem goes through even when the state space is not a probability space. As we mentioned earlier, all we need to define knowledge is a set and information partitions, no further structure is required. Bacharach's generalization ensures that, provided the message function is well-defined, the agreement theorem holds in extremely general spaces. But we lack such a degree of generality in the literature on common knowledge acquisition and communication initiated by Geanakoplos and Polemarchakis (1982). More specifically, all the analyses that study convergence to consensus, e.g. Parikh and Krasucki (1990) and Krasucki (1996), assume that information partitions are finite, even when the underlying state space is infinite, or that the state space is a probability space. But, as we already remarked, neither finiteness nor probability are strictly necessary for the agreement theorem to hold. Therefore, we fill this gap by asking the following question: Is it possible for rational people to converge to common knowledge and consensus through dialogues when the underlying state space is not assumed to be a probability space and when information partitions are not necessarily finite? We show that it is indeed possible provided that dialogues of transfinite length are allowed. Our contribution thus highlights that, at least from a conceptual viewpoint, common knowledge acquisition through communication and convergence to consensus are extremely general properties. On a methodological level, our main contribution is to provide a new framework that is general enough to accommodate for dialogues of transfinite length.

In the second chapter we provide a syntactic construction of the correlated equilibrium introduced by Aumann (1974). Correlated equilibrium is a solution concept that captures correlated play. By making use of a correlating device, players have the opportunity to select strategies that are not statistically independent of each other. Correlating devices are theoretical constructs that subsume implicit opportunities of communication and coordination that players have at their disposal. For a given game, the set of correlated equilibria identifies all the possible equilibrium outcomes that can possibly arise from all the implicit and un-

modeled communication opportunities. Being theoretical constructs, correlating devices may have no natural interpretation. Our contribution in this chapter is to provide a construction of correlated equilibrium that has a more natural interpretation. Our analysis is motivated by the following observation. When people agree to coordinate their actions, they typically do so by means of some contract or agreement expressed in some natural language. But natural languages are ambiguous, i.e. the map from sentences to meanings that they induce is not necessarily commonly known. We argue that it is this interpretive uncertainty that acts as a correlating device. Using the syntactic approach, we model explicitly the language that players use to communicate. This allows us to separate messages from their meaning. The logic we use is that of Halpern and Kets (2015), in which the interpretation of formulas is player-dependent. In a nutshell, our model consists of a group of players who agree on a public strategy that tells them how to play a given game. We show that the ambiguity of the language that players use to communicate and reason is able to induce every correlated equilibrium of the underlying game.

The third chapter of this dissertation contributes to the literature on information design. In the standard model, a designer chooses an information structure to induce one or more agents to take a certain action. This problem has been studied under a variety of assumptions on what the designer can or cannot condition her recommendations upon. Relevant analyses include Bergemann and Morris (2016a), Alonso and Câmara (2016), Chan et al. (2019), and Kolotilin et al. (2017). In all these studies, agents' preferences are assumed to be commonly known. Consequently, the designer can effortlessly send different messages to different types of agents with absolute certainty about their preferences. Our contribution is to study an information design problem under the assumption that preferences are not commonly known. To motivate our analysis, one can consider the following example. Suppose a prosecutor wants to persuade a jury to convict the defendant in a court of law. It is not hard to imagine that the juror may be uncertain about the composition of the jury she is addressing. More specifically, she does not know if jurors are relatively tough or lenient. If the prosecutor knew the exact jury composition, she would tailor her persuasion strategy accordingly. Presumably, this would allow her to induce a more beneficial (to her) outcome. We conduct our analysis by identifying the designer's optimal persuasion strategy in a model where outcomes are chosen by a two-member, heterogeneous committee. We show that the optimal strategy crucially depends on the informativeness of the prior distribution over preference types. In addition, we show how uncertainty about preferences always entails a loss to the designer with respect to the benchmark case with commonly known preferences. This shows that the assumption of complete information about preferences that is commonly made in the literature is not without loss of generality.

## 1.3 Summary of the Essays

### 1.3.1 Chapter 2: Learning to agree over large state spaces

In the first paper of this dissertation, we study the problem of common knowledge acquisition and consensus. The model consists of a finite set of agents exchanging messages according to a well-defined message function $f$. Agents are like-minded and $f$ satisfies the sure-thing principle. A rational dialogue between agents takes place as follows. A directed graph $G$ describes who sends a message to whom in one round of communication. At the end of the round, everybody updates her information by taking the join (coarsest common refinement) between her information partition and the partition induced by the messages she receives. This process naturally defines a function $g$ from the set of profiles of information partitions to itself. We thus use this function to construct a dialogue of arbitrary, and possibly transfinite length. More specifically, we define a dialogue as a sequence obtained by iterating the function $g$ transfinitely often, starting from some profile of initial information partitions.

Our result is to give sufficient conditions under which a dialogue leads to a consensus, i.e. a state of affairs where, at any state of the world, everybody sends the same message. The emergence of consensus turns out to depend crucially on the properties of the communication graph $G$. For any graph, we first show that the (non-empty) set of fixed points of the message function $f$ is always a subset of the fixed points of the function $g$ induced by $G$. Since $g$ is also increasing, this means that a dialogue is always a well-defined sequence that will eventually be constant. However, a consensus need not hold when the sequence becomes constant. Loosely speaking, learning stops at some point, but we cannot be sure that agents agree at that point. We then show that, if the graph $G$ satisfies two properties, then the sets of fixed points of $f$ and $g$ coincide. This means that, not only will learning stop at some point, but players will also agree at that point. The two properties of $G$ are the following. First, we require that $G$ contains a spanning subgraph that is strongly connected: for every pair of distinct agents $i$ and $j$, there is a directed path from $i$ to $j$ and a directed path from $j$ to $i$. The second property is that the spanning subgraph in $G$ is symmetric: if there is a directed edge from $i$ to $j$, then there is also a directed edge from $j$ to $i$. These two conditions capture the fact that everybody must be able to talk with everybody else, and that communication must be reciprocal. Under these assumptions about $G$, we are able to establish a theorem that goes roughly as follows: In any event-based model of interactive knowledge, every rational dialogue leads to a consensus. Finally, we show that the cardinality of the least index ordinal at which a consensus holds cannot be greater than $n$ times the cardinality of the state space, where $n$ is the number of agents.

### 1.3.2 Chapter 3: Coordination through ambiguous language

In the second paper, we give a syntactic construction of correlated equilibrium. The language through which players communicate is expressive enough to talk about signals, beliefs, expected payoffs, and choices in a given finite game with simultaneous moves. Before playing the game, agents receive information expressed in formulas. The interpretation of formulas is captured by an epistemic probability structure in which truth values are assigned relative to a player. This implies that players may disagree on the interpretation of the signals they are receiving. Coordination is achieved by means of a public strategy, which is a set of conditional formulas telling players how to play the game as a function of observed signals. We make assumptions about the interpretation of signals so that it is always the case that, according to any given player, everybody receives one, and only one, signal per state. Since the coordination strategy maps signals to actions, this means that, according to any player, everybody chooses one, and only one, action in each state.

Our results give a characterization of the probability distributions over action profiles induced by self-enforcing coordination strategies. By self-enforcing we mean that nobody has the incentive to choose an action different from that prescribed by the public strategy. We examine two cases separately. In the first, we assume that the language is not ambiguous, so that everybody assigns the same truth value to any given formula. We show that any self-enforcing coordination strategy induces a correlated equilibrium distribution of the underlying game. Conversely, for any correlated equilibrium distribution of the underlying game, one can always find an unambiguous language, a set of signals, and a self-enforcing coordination strategy, that induce that equilibrium distribution. The two results together suggest that our model can be interpreted as a syntactic version of the standard, event-based construction of correlated equilibrium.

In the second case, we allow the language to be ambiguous. More specifically, each player has her own interpretation function which assigns truth values to primitive propositions in every state of the world. We show that any self-enforcing coordination strategy now induces a subjective correlated equilibrium. Conversely, for every subjective correlated equilibrium of the underlying game, one can always find a (possibly ambiguous) language, a set of signals, and a self-enforcing coordination strategy, that induce that equilibrium. These results suggest that ambiguity in natural language provides a justification for heterogeneous and possibly inconsistent beliefs about strategic play.

### 1.3.3 Chapter 4: Persuading a committee with privately known preferences

In the final paper of this dissertation, we solve the decision problem of a designer who wants to persuade a two-member committee to take a certain action. Contrary to existing models, we assume that the committee members' preferences are not commonly known. The underlying environment is binary: there are two states of nature, each player has two actions at her disposal, and each player's set of types is binary. Preferences are not perfectly aligned. The designer wants the committee to take the same action irrespective of the true state of the world, whereas committee members prefer one action in one state, and the other action in the other state. As we mentioned, players can be of two types: low types require relatively little evidence to choose the designer's preferred alternative, whereas the high types are harder to persuade.

We study two cases separately. In the first case, the designer can elicit private information. She asks committee members to report their preference types and then sends action recommendations based on these reports and the true state of nature. In the second case, information elicitation is not allowed. The designer does not ask committee members to send any information at all and she sends action recommendations that are contingent on the true state of the world and her prior distribution over types. In either case, we solve the designer's choice problem both when a unanimous consent is needed to implement her preferred option and when a single approval is sufficient.

We show that the designer's optimal decision rule has some qualitative features that are invariant to the different cases we examine. More specifically, the optimal rule crucially depends on the informativeness of the prior probability distribution over types. When the designer is confident enough that no committee member is of the high type, she tailors her strategy entirely to low types without persuading high type members. The reason is that incentive constraints require that, in order to persuade high types, low types vote for the designer's preferred alternative with lower probability. Thus the expected gain from persuading the high types is more than compensated by the expected loss from the low types. When the prior is such that the designer is confident enough that one committee member, but not both of them, is of the high type, she finds it optimal not to persuade the committee to adopt her preferred policy in the case in which they both declare to be of the high type. Finally, when the prior is such that a committee of high types only is more likely, then the designer finds it optimal to induce both members to vote for her preferred policy irrespective of what information they choose to report. While these qualitative features of the optimal persuasion strategy are shown to hold across all the cases we examine, we show

that non-unanimous decision making and the possibility of eliciting information are both beneficial to the information designer.

# Chapter 2

# Learning to agree over large state spaces

## 2.1 Introduction

A classic result of Aumann (1976) shows that rational people sharing a common prior cannot agree to disagree. If their posterior beliefs about a certain event are common knowledge, then these beliefs must be the same. But what if, as it is often the case, the common knowledge assumption does not hold? Starting from a situation of disagreement, how can people arrive at a state of common knowledge and, therefore, agree? A possible answer, originally put forward by Geanakoplos and Polemarchakis (1982), is that people can achieve a consensus through dialogues. If everyone announces her beliefs, or other types of messages that depend on one's information in a sufficiently regular way, and concurrently updates her information in light of others' announcements during a dialogue, then a consensus will eventually emerge.

Our goal is to examine how general this emergence of consensus is. More specifically, we ask the following question: Does a dialogue between like-minded and rational people lead to consensus when the underlying set of states of the world is arbitrarily large? We know from existing results that, provided that messages are derived from a sufficiently regular function, there are essentially two (not mutually exclusive) cases where a dialogue ends up with a consensus. The first is when people's information about the state of the world is represented by a *finite* partition, even if the underlying state space is infinite. The second case is when the state space is a *probability space* and, consequently, exchanged messages are posterior probabilities. But one can argue that these two cases do not exhaust all possible situations that one could be interested in. First of all, the restriction to finite information partitions seems hard to justify when the underlying state space is infinite. Secondly, the problem of common knowledge acquisition and consensus is not necessarily confined to probability spaces. Indeed, the very definition of knowledge is independent from probabilities, and one can safely talk about interactive knowledge even without beliefs.

In this paper, we attempt to overcome both of the limitations we have just mentioned. More specifically, we study a model of dialogues where information partitions are not assumed to be finite and, at the same time, the underlying state space is not necessarily a probability space. Our main result is that, if two main conditions are met, it is always possible for rational and like-minded people to engage in a dialogue that ends up with a consensus on the value of a sufficiently regular function. The first condition that needs to hold is on the richness of the communication structure. Intuitively, everyone should be able to talk to everyone else during a dialogue, either directly or indirectly. And communication should be reciprocal: if agent $i$ sends her message to $j$ at some point during a dialogue, then $j$ has to send a message back to $i$. The second condition is about the length of feasible dialogues. More specifically, we allow dialogues to have transfinite length. When the state space is infinite, it might be the case that agents need to exchange infinitely many messages before reaching a consensus. Therefore, we should make sure that a dialogue lasts sufficiently long to accommodate this possibility. The first condition is essentially equivalent to that already explored in finite models. As for the second condition, to the best of our knowledge this is the first paper that allows transfinite dialogues in problems of common knowledge acquisition and consensus.

On a methodological level, we approach the problem from a somewhat different perspective than existing papers. The standard approach is to define dialogues starting from a communication protocol, i.e. a sequence of ordered pairs of agents that indicate who talks to whom and when. Every protocol induces a graph over the set of agents, and conditions ensuring consensus are found by studying the properties of this graph. Our approach takes the opposite route. The primitive object is a graph which describes who talks to whom during one round of communication. This graph induces a self-function in the set of profiles of information partitions: intuitively, it maps the information that agents have at the beginning of the communication round to the refined information they have after having talked. By iterating this function "transfinitely often" we can generate a sequence of profiles of partitions that capture all the information that is generated during the dialogue. And the dialogue is the one in which every round of communication takes place according to the graph that we initially fixed. While the two approaches are essentially equivalent, we believe that our way to frame the problem allows us to "solve" the model in a more compact way.

The paper is organized as follows. In the remainder of this section, we offer an example to illustrate why transfinite dialogues are needed and then discuss the related literature. The model is presented in section 2.2 and results are illustrated in section 2.3. A discussion of some of the assumptions we make and of the robustness of our results is contained in section 2.4.

### 2.1.1 Example

Consider a simple decision-making problem under uncertainty. There are two agents, Ann and Bob. The set of possible states of the world is the set of natural numbers $\mathbb{N}$. The true state is $x$. The set of feasible decisions is $D = \mathbb{N} \cup \{0\}$. For each agent, payoffs are determined by the following function:

$$u(d, x) = \begin{cases} 1 & \text{if } d = x \\ 0 & \text{if } d = 0 \\ -3 & \text{otherwise.} \end{cases}$$

In words, taking action $d \neq 0$ in state $x$ yields a reward if the action and the state match. If they don't, the decision maker incurs a loss. In every state, the safe option of choosing $d = 0$ is always available. Decisions are made independently. There are no payoff externalities: the payoff accruing to Ann is independent of what Bob does, and vice versa.

At the ex ante stage, information about the state is represented by a prior probability distribution over $\mathbb{N}$. The prior is such that, for every $k \in \mathbb{N}$, the probability of $x = k$ is $\frac{1}{2^k}$. If decisions were to be made at the ex ante stage, it is clear that both Ann and Bob would choose the safe option $d = 0$. In addition, the fact that Ann chose $d = 0$ would not reveal any new information to Bob, and vice versa.

Things change when decision makers are no longer symmetrically informed. Suppose that, after the actual state is determined, agents observe different partitional signals about $x$ before making their choice. More specifically, each agent is endowed with an information partition over $\mathbb{N}$. When Nature selects $x = k$, either agent learns that the true state lies in his or her partition block containing $k$. Let information partitions be as follows:

$$\pi_A = \{\{1\}, \{2, 3\}, \{4, 5\}, \dots\}$$
$$\pi_B = \{\{1, 2\}, \{3, 4\}, \{5, 6\}, \dots\}.$$

Observe that, for each information set $\{k, k+1\}$ of two elements, posterior beliefs are such that

$$\mathsf{Prob}(x = k | \{k, k+1\}) = \frac{2}{3}$$
$$\mathsf{Prob}(x = k+1 | \{k, k+1\}) = \frac{1}{3}.$$

It is then straightforward to verify that, when $x \neq 1$, the unique optimal choice is $d = 0$ for both Ann and Bob. But when $x = 1$, Ann chooses $d = 1$ and Bob selects $d = 0$. Thus there

is a state in which agents act differently, i.e. they disagree. But if Bob *knew* that they were disagreeing, he would learn new information about the state and would revise his decision accordingly.

To be more specific about the learning process we have just mentioned, suppose that agents are now allowed to communicate and revise their decisions sequentially. During each stage, Ann and Bob declare their actions. Then each of them revises his or her information in light of the other's announcement. Finally, they can change their decisions.

During the *first* stage, we have the following scenario. If the state is $x = 1$, Ann already knows this. Bob only knows that the true state must be either $x = 1$ or $x = 2$. Also he knows that Ann would choose $d = 1$ when $x = 1$ and $d = 0$ otherwise. By learning that Ann chose $d = 1$, he concludes that $x = 1$ and so he changes his action to $d = 1$. Similarly, when $x = 2$, Bob infers from Ann's picking $d = 0$ that $x \neq 1$. Thus he concludes that $x = 2$ and changes his action to $d = 2$. If $x \neq 1, 2$, neither agent can learn anything new about the state. In sum, at the end of the first stage, Ann and Bob have the following information partitions:

$$\pi_A^1 = \{\{1\}, \{2,3\}, \{4,5\}, \dots\}$$
$$\pi_B^1 = \{\{1\}, \{2\}, \{3,4\}, \{5,6\}, \dots\}.$$

In the *second* stage, both agents revise their information on the basis of what they communicate during that stage and of what they learned from the previous stage. Therefore, when $x = 1$, they already know this. When $x = 2$, Ann only knows that either $x = 2$ or $x = 3$. Also she knows that Bob would choose $d = 2$ in the former case and $d = 0$ in the latter. By learning that Bob chose $d = 2$ at the end of the previous stage, she infers that true state must be $x = 2$, so changing her decision to $d = 2$. Similarly, if $x = 3$, Ann infers from Bob's picking $d = 0$ in the first stage that $x \neq 2$. Thus she concludes that $x = 3$ and selects $d = 3$. When $x \geq 4$, neither agent can learn anything new about the state. In sum, at the end of the second stage, Ann and Bob have the following information partitions:

$$\pi_A^2 = \{\{1\}, \{2\}, \{3\}, \{4,5\}, \dots\}$$
$$\pi_B^2 = \{\{1\}, \{2\}, \{3,4\}, \{5,6\}, \dots\}.$$

It is easy to see that, for any state $k$, it takes $k$ stages of communication for both players to learn that the actual state is indeed $k$. But there is always a state where they disagree, the reason being that their actions are not commonly known at that state. Is it possible to achieve a consensus at *every* state in $\mathbb{N}$? Equivalently, can Ann and Bob learn to agree

over the entire state space? We show later on that the answer is affirmative if the dialogue they are participating in has order type $\omega + 1$, and not just $\omega$ as it is commonly assumed in existing models.

### 2.1.2 Related literature

The paper contributes to the vast literature on common knowledge and agreement initiated by Aumann (1976), a survey of which can be found in Bonanno and Nehring (1997). Papers that are closer to ours are those focusing on dialogues and convergence to consensus. Geanakoplos and Polemarchakis (1982) introduce dialogues in a two-player model with a finite state space where messages exchanged during the dialogue are posterior beliefs about a fixed event. Bacharach (1985) and Cave (1983) show that a consensus can be reached not only when people communicate posterior beliefs but also when they communicate the values of any function satisfying a condition akin to the sure thing principle from decision theory. Bacharach (1985) assumes that initial information partitions are finite, whereas Cave (1983) assumes that the state space is countably infinite. Washburn and Teneketzis (1984), Nielsen (1984), and Bergin (1989) study convergence to consensus but they all confine their attention to the probabilistic case only.

Dialogues between more than two agents and with private communication are introduced by Parikh and Krasucki (1990) and further examined in Krasucki (1996) and Heifetz (1996). Assuming finite partitions, they show how convergence to a commonly known consensus is guaranteed if communication takes place according to a protocol whose graph is strongly connected and symmetric.

Our paper is also related to the common learning model of Cripps et al. (2008). They study (approximate) common knowledge acquisition for two agents who privately observe a sequence of exogenous signals. In our model, we let agents observe external private signals only once, i.e. at the beginning of a dialogue. As a consequence, people learn from the messages that they endogenously choose to exchange during a dialogue.

Mueller-Frank (2013) provides a framework from learning in social networks in an environment similar to ours. However, his analysis is confined to countably infinite information partitions and he uses choice correspondences instead of choice (message) functions as we do.

Finally, both Aumann and Hart (2003) and Parikh (1992) study dialogues of transfinite length. In the former, a simultaneous-move game is played after a countable sequence of cheap talk messages are exchanged. In the latter, knowledge acquisition is studied using Kripke structures. The analysis is confined to the countable case and interactive discovery

systems are introduced instead of message functions.

## 2.2 Model

### 2.2.1 Setup

Our object of study is an environment $\mathcal{E} = (I, X, A, f, G)$ where:

- $I = \{1, \ldots, n\}$, with $n \geq 2$, is a finite set of agents;
- $X$ is a nonempty set of states of the world;
- $A$ is a nonempty set of messages;
- $f : \mathcal{X} \longrightarrow A$ is a message function, where $\mathcal{X}$ is the set of non-empty subsets of $X$;
- $G$ is a directed graph whose set of nodes is $I$. Abusing notation, we write $G$ to indicate both the graph and its set of edges $G \subseteq I \times I$.

Information about the state is represented by partitions. The set of all partitions of $X$ is $\Pi$, with typical elements $\pi$, $\pi'$, etc. Given a state $x \in X$ and a partition $\pi$ of $X$, the block of the partition containing $x$ is denoted by $\pi(x)$. The set $\Pi$ is partially ordered by the relation $\leq$ such that, for any two partitions $\pi$ and $\pi'$, we have $\pi \leq \pi'$ if and only if $\pi$ is a coarsening of $\pi'$, i.e. every block of $\pi$ can be written as the union of some blocks of $\pi'$. We use $\pi \vee \pi'$ to denote the join (coarsest common refinement) of $\{\pi, \pi'\}$, and $\bigvee \{\pi_h : h \in H\}$ for the join of the indexed family $\{\pi_h : h \in H\}$. Similarly, we use $\bigwedge \{\pi_h : h \in H\}$ to indicate the meet (finest common coarsening) of the family $\{\pi_h : h \in H\}$. Recall that $\Pi$ is a complete lattice.

When agent $i$'s information is represented by a partition $\pi_i \in \Pi$, we say that $i$ has information $\pi_i$. The definition of knowledge is standard. Given a state $x \in X$ and an event $E \subseteq X$, we say that agent $i$ knows $E$ in state $x$ if $\pi_i(x) \subseteq E$. We say that $E$ is common knowledge at $x$ if $\bigwedge \{\pi_i : i \in I\}(x) \subseteq E$.

### 2.2.2 Messages, communication, and learning

Agents are allowed to exchange messages. The message function $f$ determines how agents send messages as a function of their information. The graph $G$ determines who sends a message to whom. We do not make any particular assumption about the content of messages. We interpret a message just as a function of agents' private information. For example, a message can be a posterior belief about a certain event as in Geanakoplos and Polemarchakis

(1982); it can be an action as in Example 2.1.1; or it can be a string of symbols in some formal language.

**Messages.** When $i$ has information $\pi_i$, we use the function $f_i : X \longrightarrow A$ to indicate what message $i$ sends at any given state $x$. Since no confusion should arise, we save on notation by dropping the dependence of $f_i$ on $\pi_i$. We assume the following condition.

**Assumption 1** (Like-mindedness)**.** *For every $i \in I$, and for every partition $\pi_i \in \Pi$, if $i$ has information $\pi_i$, then $f_i(x) = f(\pi_i(x))$ for every $x \in X$.*

Like-mindedness captures the fact that agents share the same view of the world. If any two agents have the same information in a given state, then they must send the same message in that state. Consequently, agents' sending different messages is solely due to asymmetric information and not to, say, different subjective states or other forms of fundamental disagreement. Notice that, in every state $x$ and for every player $i$, the message that $i$ sends when $x$ is the true state is a function of the smallest event that $i$ knows at $x$, i.e. $\pi_i(x)$.

Another implication of Assumption 1 is that, for every $x, x' \in X$, if $\pi_i(x) = \pi_i(x')$, then $f_i(x) = f_i(x')$. This reflects full rationality. If an agent transmitted different messages in different states belonging to the same information block, then she would realize that those states are not indistinguishable after all, and so she would assign them to different information blocks. In addition, every agent always knows the message she is transmitting.

We also make the following assumption about $f$.

**Assumption 2** (Sure thing principle (STP))**.** *For any $S \in \mathcal{X}$, and for any partition $\{S_h : h \in H\}$ of $S$, if $f(S_h) = a$ for all $h \in H$ then $f(S) = a$.*

We use the same formulation as Bacharach (1985). The condition is also known as union consistency[1]. Intuitively, the STP says that if an agent sends message $a$ when she knows that the state is in $S \subseteq X$, and she sends again message $a$ when she knows that the state is in $S'$, with $S \cap S' = \emptyset$, then she must send the same message $a$ when she knows that the state is in $S \cup S'$. Bacharach (1985) shows that the STP is satisfied by "just about any plausible theory of rational decision". Two relevant examples of message functions satisfying this principle are: 1) the function that, as in Example 2.1.1, maps information sets to Bayes rational choices; 2) the function that, as in Geanakoplos and Polemarchakis (1982), maps information sets to posterior beliefs about a certain event[2].

[1] See Section 2.4 for a comparison between the sure thing principle as defined in Bacharach (1985) and the union consistency of Cave (1983).

[2] See Moses and Nachum (1990) for a critique of the STP in epistemic models, and Samet (2010) and Tarbush (2016) for possible ways to address their critique.

We can now define working partitions. For a given individual signal function $f_i$, we let $W_i$ be the corresponding working partition. For every $x \in X$, the block of $W_i$ containing $x$ is $W_i(x) := \{x' \in X : f_i(x') = f_i(x)\}$. In words, $W_i(x)$ corresponds to the event "$i$ emitted signal $a$" for some $a \in A$. Therefore, one can also interpret $W_i(x)$ as the information conveyed to any agent $j \neq i$ who receives $i$'s message $f_i(x)$ in state $x$. The fact that $W_i$ is a partition reflects the lack of any sort of ambiguity about the interpretation of messages. Since no confusion should arise, we save again on notation by dropping the dependence of $W_i$ on the underlying information partition $\pi_i$. Finally, notice that $W_i$ is necessarily a coarsening of $\pi_i$.

**Communication and learning.** Communication between agents takes place according to the graph $G$. If $(i, j) \in G$, then there is a directed edge from $i$ to $j$ and we say that $i$ sends a message to $j$. To describe how a receiver updates her information upon receiving a message, we introduce a function $g$ constructed as follows.

Let $\Pi^n$ be the $n$-fold Cartesian product of $\Pi$. An element of $\Pi^n$ is an indexed collection $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$ of partitions of $X$. We endow this space with the product order:

$$(\pi_1, \ldots, \pi_n) \leq (\pi'_1, \ldots, \pi'_n) \iff \pi_i \leq \pi'_i \text{ for all } i \in I.$$

Notice that $\Pi^n$ is a complete lattice. For each $i \in I$, we define the (possibly empty) set $S(i) := \{j \in I : (j, i) \in G\}$. In words, $S(i)$ is the subset of agents that send a message to $i$. We can now define the function $g : \Pi^n \longrightarrow \Pi^n$ as follows:

$$g_i\left((\pi_1, \ldots, \pi_n)\right) = \begin{cases} \bigvee \{\pi_i \vee W_j\}_{j \in S(i)} & \text{if } S(i) \neq \emptyset \\ \pi_i & \text{otherwise,} \end{cases} \tag{2.1}$$

where we write $g_i(\boldsymbol{\pi})$ to denote the $i$th component of $g(\boldsymbol{\pi})$.

The function $g$ captures the following process of communication and learning. Suppose agents have information $\boldsymbol{\pi}$. Then they exchange messages according to the communication graph $G$. How should they revise their information in light of the new information they receive? If $i$ does not receive any message, her information will clearly stay the same. But if she receives a message from $j$, and if the state is $x$, she reasons as follows[3]: "I know that the true state must be in $\pi_i(x)$, and I know that $j$ has information $\pi_j$. Now, $j$ sent me a message $f_j(x)$, and I know that he would have sent that message if and only if the true state had been contained in $W_j(x)$. Therefore, I can conclude that the true state must be

[3]This is the learning process introduced in Parikh and Krasucki (1990) and later amended by Weyers (1992). In particular, Weyers (1992) shows that fully rational agents update their entire information partition and not just the partition block containing the true state of the world.

in $\pi_i(x) \cap W_j(x)$." By repeating this line of reasoning at any state, we obtain that $i$'s new information partition after receiving a message from $j$ is the join $\pi_i \vee W_j$. With more than one sender, $i$ refines her information by taking into account the working partition of every $j \in S(i)$.

We assume that communication does not take place just once. Agents are allowed to engage in *dialogues* of arbitrary length. Formally, given a profile $\boldsymbol{\pi}^{\mathbf{0}}$ of initial information partitions, a dialogue starting from $\boldsymbol{\pi}^{\mathbf{0}}$ is the sequence $(g^\alpha : \alpha \in \mathsf{Ord})$ constructed recursively as follows:

$$\begin{aligned} g^0 &:= \boldsymbol{\pi}^{\mathbf{0}}, \\ g^{\alpha+1} &:= g\left(g^\alpha\right) \text{ for every ordinal } \alpha, \\ g^\lambda &:= \bigvee \{g^\alpha : \alpha < \lambda\} \text{ for every limit ordinal } \lambda. \end{aligned}$$

In words, a dialogue is a sequence in $\Pi^n$ starting from an initial profile $\boldsymbol{\pi}^{\mathbf{0}}$ and constructed by iterating "transfinitely often" the function $g$ induced by $G$. Notice that a profile $\boldsymbol{\pi}$ uniquely determines the profile of messages transmitted at every state. Thus it is without loss of generality to define a dialogue as a sequence of partitions and not, as it would be more natural, as a sequence of messages.

The fact that we construct a dialogue from $g$ can also be interpreted as follows. The graph $G$ describes one *round* of communication; the corresponding function $g$ maps profiles of information partitions that agents have at the beginning of this round of communication to profiles of partitions that are refined in light of the messages exchanged during the communication round. Thus a dialogue is nothing other than the transfinite repetition of this round of communication: the element $g^\alpha$ tells us what information agents have at the end of the $\alpha$th round of communication.

Finally, we remark that the initial profile $\boldsymbol{\pi}^{\mathbf{0}}$ can be thought of as exogenous information, whereas partitions $g^\alpha$, with $\alpha > 0$, can be thought of as endogenous information. That is, $\boldsymbol{\pi}^{\mathbf{0}}$ captures the information content of a privately observed signal about the state that we do not explicitly model. Nature acts only once and determines what realization of this signal agents observe. Subsequent information partitions are endogenously determined by the communication and learning process described above. As we discuss in more detail in section 2.4, the whole structure of the model is common knowledge. In particular, it is commonly known who talks with whom and when, how partition blocks are mapped to messages, and how information is updated.

## 2.3 Results

In this section we study properties of the communication structure that lead to consensus. We first give a full characterization of consensus for the static case, i.e. for a fixed profile of information partitions; we then find conditions under which dialogues lead to consensus.

### 2.3.1 Consensus

Let $\boldsymbol{\pi} \in \Pi^n$ be the profile of agents' information partitions. Then we say that a *consensus* holds if, for all $i, j \in I$, we have that $f_i = f_j$. Our definition describes consensus in a global sense. That is, we require that, for all $i, j \in I$, $f_i(x) = f_j(x)$ for every state $x \in X$. If agents agree at some state $x$ but not necessarily at every state, then we say that a *partial* consensus holds at $x$.

Our first result is a full characterization of (global) consensus.

**Proposition 1.** *Suppose agents have information $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$. Then the following are equivalent:*

*a) For all $x \in X$, the profile of messages that is sent at $x$, i.e. the event*

$$E(x) = \{x' \in X : f_1(x') = f_1(x), \ldots, f_n(x') = f_n(x)\},$$

*is common knowledge at $x$*

*b) For all $i, j \in I$, $f_i = f_j$*

*c) For all $i, j \in I$, $W_i = W_j$.*

*Proof.* The implication $a) \Rightarrow b)$ follows from Theorem 3 in Bacharach (1985). $b) \Rightarrow c)$ follows immediately from the definition of the working partition. To show $c) \Rightarrow a)$, fix a state $x \in X$. By the definition of the working partition, for every $i \in I$, the event $\{x' \in X : f_i(x') = f_i(x)\}$ is the same as $W_i(x)$. Let $W(x) := \cap_{i \in I} W_i(x)$. Since every $W_i$ is a coarsening of $\pi_i$, and since $W_i(x) = W_j(x)$ for all $i, j \in I$ by assumption, we have that, for all $i \in I$,

$$\pi_i(x) \subseteq W_i(x) = W(x).$$

Therefore, for all $i \in I$,

$$\pi_i(x) \subseteq \bigwedge \{\pi_i : i \in I\}(x) \subseteq W(x).$$

□

A global consensus is equivalent to the knowledge configuration where, at every state, the profile of messages that are being sent at that state is common knowledge. And when a profile of messages is commonly known, then those messages must be the same. The latter statement is nothing other than the generalized version of Aumann's agreement theorem established in Bacharach (1985).

We can also say that a global consensus cannot hold without it being common knowledge that it holds. This is not necessarily true for a partial consensus. In Example 2.1.1, both Ann and Bob send message 0 in state 2. But this is not common knowledge and not even mutual knowledge. Since Bob knows that the state can be either 1 or 2, he doesn't know which message Ann is going to send to him.

Notice that a consensus does not imply that agents have the same information partitions. Furthermore, we emphasize that the equivalence in Proposition 1 crucially relies on the STP. Without it[4], one can only conclude that $b) \Rightarrow c) \Rightarrow a)$. Consequently, it would no longer be impossible to agree to disagree.

We conclude this subsection with a corollary that will prove useful in establishing subsequent results.

**Corollary 1.** *If $f_i \neq f_j$, then $\pi_i < \pi_i \vee W_j$ or $\pi_j < \pi_j \vee W_i$.*

*Proof.* By contrapositive, suppose that neither $\pi_i < \pi_i \vee W_j$ nor $\pi_j < \pi_j \vee W_i$ hold. Since it is always the case that $\pi_i \leq \pi_i \vee W_j$ and $\pi_j \leq \pi_j \vee W_i$, we must have both $\pi_i = \pi_i \vee W_j$ and $\pi_j = \pi_j \vee W_i$. This implies that $W_i$ is a coarsening of $\pi_j$ and $W_j$ is a coarsening of $\pi_i$. Combining this with the fact that each working partition is a coarsening of the underlying information partition, we have that both $W_i$ and $W_j$ are common coarsenings of $\pi_i$ and $\pi_j$. Therefore, for every $x \in X$, $W_i(x) \cap W_j(x)$ is common knowledge at $x$ between $i$ and $j$. Thus it follows from Proposition 1 that $f_i = f_j$. □

The interpretation is straightforward. If $i$ and $j$ disagree at some state, then it must be the case that either $i$ can (strictly) refine her information by receiving a message from $j$, or $j$ can refine his information by receiving a message from $i$, or both. In other words, when two agents are disagreeing, at least one of them can learn some new information from the other.

### 2.3.2 Dialogues leading to consensus

We now examine conditions under which dialogues lead to consensus. Formally, for a given communication graph $G$, we say that the corresponding dialogue $(g^\alpha : \alpha \in \mathsf{Ord})$, starting

[4]See Section 2.4 for a proof.

from initial information $\boldsymbol{\pi}^{\mathbf{0}}$, leads to a consensus if, for every $i, j \in I$, $f_i^\alpha = f_j^\alpha$ for some $\alpha \in \mathsf{Ord}$, where we write $f_i^\alpha$ to indicate the individual message function associated with the $i$th component of $g^\alpha$. In other words, a dialogue leads to consensus if the sequence $(g^\alpha : \alpha \in \mathsf{Ord})$ contains a profile $g^\alpha \in \Pi^n$ at which everybody agrees.

The structure of the communication graph $G$ clearly affects $g$ and, consequently, the corresponding dialogue $(g^\alpha : \alpha \in \mathsf{Ord})$. We make the following preliminary observation.

**Remark.** *For any $G$, the function $g$ is inflationary but need not be monotone*[5].

*Proof.* It follows immediately from (2.1) that $g$ is inflationary. The following example shows that $g$ need not be monotone. Let $X = \{x, y, w, z\}$, $I = \{1, 2\}$, and let the message function $f$ be such that $f(\{x\}) = f(\{x, y\}) = a$, and $f(S) = b$ for any other non-empty subset $S$ of $X$. In addition, suppose the communication graph is $G = \{(1, 2), (2, 1)\}$. Now take the following elements of $\Pi^2$:

$$\begin{aligned} \boldsymbol{\pi} &= (\pi_1, \pi_2) = (\{X\}, \{\{x, y\}, \{w, z\}\}) \\ \boldsymbol{\pi'} &= (\pi'_1, \pi'_2) = (\{X\}, \{\{x\}, \{y\}, \{w\}, \{z\}\}) . \end{aligned}$$

Thus we have

$$\begin{aligned} g(\boldsymbol{\pi}) &= (\{\{x, y\}, \{w, z\}\}, \{\{x, y\}, \{w, z\}\}) \\ g(\boldsymbol{\pi'}) &= (\{\{x\}, \{y, w, z\}\}, \{\{x\}, \{y\}, \{w\}, \{z\}\}) . \end{aligned}$$

Therefore, $\boldsymbol{\pi} \leq \boldsymbol{\pi'}$ but $g(\boldsymbol{\pi}) \not\leq g(\boldsymbol{\pi'})$. □

A consequence of $g$'s being inflationary is that the sequence $(g^\alpha : \alpha \in \mathsf{Ord})$ is increasing, i.e. for every $\alpha, \beta \in \mathsf{Ord}$, $\beta < \alpha$ implies $g^\beta \leq g^\alpha$. The argument is straightforward and is by induction on $\alpha$.

We define for later use the following subsets of $\Pi^n$:

$$\mathsf{Cons}(f) := \{\boldsymbol{\pi} \in \Pi^n : f_i = f_j \text{ for all } i, j \in I\}$$

and

$$\mathsf{Fix}(g) := \{\boldsymbol{\pi} \in \Pi^n : g(\boldsymbol{\pi}) = \boldsymbol{\pi}\} .$$

In words, $\mathsf{Cons}(f)$ is the set of partition profiles at which a global consensus holds, whereas $\mathsf{Fix}(g)$ is the set of fixed points of $g$.

[5]Let $P$ be a poset. Then a function $h : P \longrightarrow P$ is monotone (or order-preserving) if $x \leq y \implies h(x) \leq h(y)$; $h$ is inflationary (or increasing) if, for all $x \in P$, $x \leq h(x)$.

**Proposition 2.** *For any $G$, we have $\emptyset \neq \mathsf{Cons}(f) \subseteq \mathsf{Fix}(g)$.*

*Proof.* To show that $\mathsf{Cons}(f)$ is nonempty, take any profile $(\pi_1, \ldots, \pi_n) \in \Pi^n$ such that $\pi_i = \pi_j$ for every $i, j \in I$. By like-mindedness, any such a profile is always contained in $\mathsf{Cons}(f)$.

To show the inclusion $\mathsf{Cons}(f) \subseteq \mathsf{Fix}(g)$, take $(\pi_1, \ldots, \pi_n) \in \mathsf{Cons}(f)$. By Proposition 1, $(\pi_1, \ldots, \pi_n)$ is such that $W_i = W_j$ for any $i, j \in I$. Therefore, since $W_i$ is a coarsening of $\pi_i$, we have that $\pi_i \vee W_j = \pi_i$ for any $i, j \in I$. Thus $(\pi_1, \ldots, \pi_n) \in \mathsf{Fix}(g)$. □

Notice that the non-emptiness of $\mathsf{Fix}(g)$ could alternatively be proved by invoking the Bourbaki-Witt fixed point theorem, see (Roman, 2008, Theorem 12.7).

It is clear that if a dialogue $(g^\alpha : \alpha \in \mathsf{Ord})$ contains a fixed point at $g^\alpha$, then it stays constant at any $\beta > \alpha$. However, it is not necessarily the case that the dialogue leads to a consensus. In order for this to be the case, we need to make sure that the communication structure in $G$ is sufficiently rich. We thus make the following assumption.

**Assumption 3.** *The communication graph $G$ contains a spanning subgraph[6] $G'$ such that:*

*a) $G'$ is strongly connected: for every distinct $i, j \in I$, there exists a directed path in $G'$ from $i$ to $j$ and a directed path from $j$ to $i$;*

*b) $G'$ is symmetric: for every $i, j \in I$, if $(i, j) \in G'$, then $(j, i) \in G' \subseteq G$.*

Strong connectedness says that no one is excluded from communication, i.e. everyone communicates with everybody else, either directly or indirectly. Symmetry means that communication is reciprocal. When the communication structure satisfies these two properties, the following equivalence holds.

**Proposition 3.** *Let $G$ satisfy Assumption 3. Then $\mathsf{Cons}(f) = \mathsf{Fix}(g)$.*

*Proof.* By Proposition 2, it is enough to show that $\mathsf{Fix}(g) \subseteq \mathsf{Cons}(f)$. Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$ be a fixed point of $g$. Suppose by way of contradiction that $\boldsymbol{\pi}$ does not induce a consensus. Hence there are distinct $i$ and $j$ in $I$ such that $f_i \neq f_j$. By strong connectedness, there exists a directed path in $G$ from $i$ to $j$: that is, for some integer $K \geq 1$, there is a path $i_0, i_1, \ldots, i_K$ in $G$ such that $i_0 = i$ and $i_K = j$. Since $i$ and $j$ disagree, this path must contain an edge $(i_k, i_{k+1})$ such that $i_k$ and $i_{k+1}$ disagree, for some $k \in \{0, \ldots, K-1\}$. By symmetry, $(i_{k+1}, i_k) \in G$. By Corollary 1, we have

$$\pi_k < \pi_k \vee W_{k+1} \quad \text{or} \quad \pi_{k+1} < \pi_{k+1} \vee W_k,$$

[6]Recall that a spanning subgraph of $G$ is a subgraph $G' \subseteq G$ with the same set of vertexes as $G$.

and using this in (2.1) we obtain

$$\pi_k < g_k(\boldsymbol{\pi}) \quad \text{or} \quad \pi_{k+1} < g_{k+1}(\boldsymbol{\pi}),$$

so contradicting the hypothesis that $\boldsymbol{\pi}$ is a fixed point of $g$. □

In light of Proposition 3, looking for dialogues leading to consensus is the same as looking for fixed points of $g$. Intuitively, we know from Corollary 1 that, in case of disagreement between $i$ and $j$, learning can take place in either direction. Assumption 3 makes sure that communication between $i$ and $j$ is reciprocal, so that it can never be the case that $i$ and $j$ disagree without having the possibility of exchanging messages between each other. The importance of reciprocity in communication has been already pointed out by Krasucki (1996), and Example 2 in Parikh and Krasucki (1990) shows how a consensus may never emerge if one dispenses with it.

We can now establish the main result. We write $\alpha^*$ to denote the least ordinal $\alpha$ such that $g^{\alpha+1} = g^\alpha$. Then we have the following.

**Theorem.** *Let $\mathcal{E} = (I, X, A, f, G)$ be an environment satisfying Assumptions 1-3. For any profile $\boldsymbol{\pi}^{\mathbf{0}}$ of initial information partitions, the dialogue $(g^\alpha : \alpha \in \mathsf{Ord})$ induced by $G$ and starting from $\boldsymbol{\pi}^{\mathbf{0}}$ always leads to a consensus. Furthermore, $|\alpha^*| \leq n|X|$.*

*Proof.* Since $g$ is inflationary, and since $\Pi^n$ is a complete lattice, it follows from (Roman, 2008, Theorem 12.9) that the sequence $(g^\alpha : \alpha \in \mathsf{Ord})$ starting from $\boldsymbol{\pi}^{\mathbf{0}}$ is always well-defined, increasing, and contains one, and only one, fixed point of $g$. By Proposition 3, the dialogue $(g^\alpha : \alpha \in \mathsf{Ord})$ induces a consensus.

In order to show that $|\alpha^*| \leq n|X|$, take the subsequence $(g^\alpha : \alpha \leq \alpha^*)$. Since $(g^\alpha : \alpha \in \mathsf{Ord})$ is increasing and $g^{\alpha^*}$ is a fixed point of $g$, the subsequence $(g^\alpha : \alpha \leq \alpha^*)$ is strictly increasing, i.e. for all $\alpha, \beta \leq \alpha^*$,

$$\beta < \alpha \Longrightarrow g^\beta \leq g^\alpha \quad \text{and} \quad g^\beta \neq g^\alpha. \tag{2.2}$$

Now define the image of $(g^\alpha : \alpha \leq \alpha^*)$ as $\mathsf{Im} := \{g^\alpha : \alpha \leq \alpha^*\}$. Since $(g^\alpha : \alpha \leq \alpha^*)$ is strictly increasing, $\mathsf{Im}$ is a well-ordered subset of $\Pi^n$ having order type $\alpha^* + 1$. Furthermore, for every $i \in I$, let $\mathsf{Im}_i := \{g_i^\alpha : \alpha \leq \alpha^*\}$ be the $i$th projection of $\mathsf{Im}$. Notice that $\mathsf{Im}_i$ is a well-ordered subset of $\Pi$ having order type at most $\alpha^* + 1$. Furthermore, by Lemma 3.1 in Avery et al. (2018), $|\mathsf{Im}_i| \leq |X|$. Now construct a map $\phi : \alpha^* + 1 \longrightarrow \bigsqcup_{i=1}^n \mathsf{Im}_i$, where $\bigsqcup$ denotes disjoint union, as follows. Let $\phi(0) := \pi_1^0$, and for every $0 < \alpha \leq \alpha^*$,

$$\phi(\alpha) := \min_i \left\{ g_i^\alpha : g_i^\beta \leq g_i^\alpha \quad \text{and} \quad g_i^\beta \neq g_i^\alpha \quad \text{for all} \quad \beta < \alpha \right\}. \tag{2.3}$$

In words, $\phi$ maps each ordinal $\alpha$ less than or equal to $\alpha^*$ to an individual partition $g_i^\alpha$ that is a strict refinement of all partitions $g_i^\beta$ having index less than $\alpha$. Without loss of generality, in case of multiple individual partitions satisfying (2.3), we take the one with the lowest (agent) index. It follows from (2.2) that $\phi$ is well-defined and injective. Therefore we have

$$|\alpha^*| \leq |\alpha^* + 1| \leq \left| \bigsqcup_{i=1}^{n} \mathsf{Im}_i \right| \leq n|X|.$$

□

Notice that, when $X$ is an infinite set, we have $|\alpha^*| \leq n|X| = |X|$.

We now provide an example to show what a dialogue of transfinite length looks like in a three-agent environment.

### 2.3.3 Example

Suppose the state space is $X = \mathbb{N}$, the set of players is $I = \{A, B, C\}$, and the profile $\boldsymbol{\pi^0}$ of initial information partitions is the following:

$$\begin{aligned}
\pi_A^0 &= \{\{1,3,6\},\{2,4\},\{5\},\{7,9\},\{11,13\},\dots,\{8,10\},\{12,14\},\dots\} \\
\pi_B^0 &= \{\{1,3,6,8\},\{2,4\},\{5,7\},\{9,11\},\dots,\{10,12\},\{14,16\},\dots\} \\
\pi_C^0 &= \{\{1,5,9,13\dots\},\{2,6,10,14,\dots\},\{3,7,11,15,\dots\},\{4,8,12,16,\dots\}\}
\end{aligned}$$

or, more formally:

$$\begin{aligned}
\pi_A^0 &= \{\{1,3,6\},\{2,4\},\{5\}\} \cup \{\{7+4k_1, 9+4k_1\} : k_1 \geq 0\} \cup \{\{8+4k_2, 10+4k_2\} : k_2 \geq 0\} \\
\pi_B^0 &= \{\{1,3,6,8\},\{2,4\}\} \cup \{\{5+4k_1, 7+4k_1\} : k_1 \geq 0\} \cup \{\{10+4k_2, 12+4k_2\} : k_2 \geq 0\} \\
\pi_C^0 &= \{\{1+4k_1 : k_1 \geq 0\},\{2+4k_2 : k_2 \geq 0\},\{3+4k_3 : k_3 \geq 0\},\{4+4k_4 : k_4 \geq 0\}\}.
\end{aligned}$$

The message function is:

$$f(S) = \begin{cases} k & \text{if } S = \{k\} \text{ for some } k \in \mathbb{N} \\ 0 & \text{otherwise.} \end{cases}$$

The communication graph is $G = \{(A,B),(B,A),(B,C),(C,B)\}$.

At the first iteration of $g$ we obtain the following information partitions:

$$g_A^1 = \pi_A^0$$

$$g_B^1 = \{\{1,3,6,8\},\{2,4\},\{5\},\{7\},\{9,11\},\dots,\{10,12\},\{14,16\},\dots\}$$
$$g_C^1 = \pi_C^0.$$

At $\alpha = 2$ we get

$$g_A^2 = \{\{1,3,6\},\{2,4\},\{5\},\{7\},\{9\},\{11,13\},\dots,\{8,10\},\{12,14\},\dots\}$$
$$g_B^2 = g_B^1$$
$$g_C^2 = \{\{5\},\{7\},\{1,9,13\dots\},\{2,6,10,14,\dots\},\{3,11,15,\dots\},\{4,8,12,16,\dots\}\}.$$

Continuing this way, at the first limit ordinal we have

$$g_A^\omega = \{\{1,3,6\},\{2,4\},\{5\},\{7\},\{9\},\{11\},\{13\},\dots,\{8,10\},\{12,14\},\dots\}$$
$$g_B^\omega = \{\{1,3,6,8\},\{2,4\},\{5\},\{7\},\{9\},\{11\},\{13\},\dots,\{10,12\},\{14,16\},\dots\}$$
$$g_C^\omega = \{\{1\},\{3\},\{5\},\{7\},\{9\},\{11\},\{13\},\dots,\{2,6,10,14,\dots\},\{4,8,12,16,\dots\}\}.$$

Similarly, at the second limit ordinal we have

$$g_A^{\omega\cdot 2} = \{\{2,4\},\{1\},\{3\},\{5\},\{6\},\dots\}$$
$$g_B^{\omega\cdot 2} = \{\{2,4\},\{1\},\{3\},\{5\},\{6\},\dots\}$$
$$g_C^{\omega\cdot 2} = \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\dots\}.$$

In words, $C$ always knows the true state; $A$ and $B$ knows what the true state is expect for when $x = 2$ or $x = 4$. But at $\alpha = \omega \cdot 2 + 1$, $B$ learns to distinguish between $x = 2$ and $x = 4$ by communicating with $C$. Then at $\alpha = \omega \cdot 2 + 2$, $A$ learns this piece of information from $B$. Therefore, at $\alpha = \omega \cdot 2 + 2$ we finally get

$$g_A^{\omega\cdot 2+2} = g_B^{\omega\cdot 2+2} = g_C^{\omega\cdot 2+2} = \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\dots\} = \pi_A^0 \vee \pi_B^0 \vee \pi_C^0,$$

so that $\alpha^* = \omega \cdot 2 + 2$. Notice that $X$ and $\alpha^*$ have the same cardinality, but their order types are different.

## 2.4 Discussion

1. Proposition 1 and subsequent results hinge upon the STP. If one dispenses with it, the equivalence $a) \Leftrightarrow b) \Leftrightarrow c)$ in Proposition 1 breaks down and one can only conclude that $b) \Rightarrow c) \Rightarrow a)$. Notice that the arguments used in the main text to prove both $b) \Rightarrow c)$

and $c) \Rightarrow a)$ do no rely on the STP. Now let us consider two elementary examples to show how the converse implications do not need to hold. First, let us show how $b)$ does not follow from $a)$ or from $c)$. Set $X = \{x, y\}$ and let the message function be such that $f(\{x\}) = f(\{y\}) = a$ and $f(\{x, y\}) = b$. Suppose that there are two agents whose information partitions are $\pi_1 = \{\{x\}, \{y\}\}$ and $\pi_2 = \{\{x, y\}\}$. It is common knowledge at every state what messages 1 and 2 are sending, but clearly $f_1 \neq f_2$. Notice that we also have $W_1 = W_2 = X$.

It remains to show that $c)$ does not follow from $a)$. Set $X = \{x, y, z\}$ and let the message function be such that

$$f(\{x\}) = f(\{y\}) = a$$
$$f(\{z\}) = f(\{x, y\}) = b.$$

Suppose that there are two agents with the following information partitions:

$$\pi_1 = \{\{x\}, \{y\}, \{z\}\}$$
$$\pi_2 = \{\{x, y\}, \{z\}\}.$$

At every state, the profile of messages is common knowledge but $W_1 \neq W_2$.

2. We use the same formulation of the STP as Bacharach (1985). In several papers in the literature on common knowledge and consensus, the union consistency of Cave (1983) is used instead of the STP.

**Definition** (Union consistency (UC))**.** *The message function $f : \mathcal{X} \longrightarrow A$ is union consistent if $S, S' \in \mathcal{X}$, $S \cap S' = \emptyset$, and $f(S) = f(S') = a$ imply $f(S \cup S') = a$.*

While the two conditions are equivalent in the finite case, this is no longer true for infinite information partitions. The reason is simple: UC applies to finite collections of disjoint sets, while the STP applies to collections of arbitrary cardinality. Consequently, UC is no longer sufficient for a consensus to hold in the infinite case. This can easily be seen in the following example.

Set $X = \mathbb{N}$ and let the message function $f$ be the following:

$$f(S) = \begin{cases} 0 & \text{if } |S| < \infty \\ 1 & \text{otherwise.} \end{cases}$$

It is straightforward verifying that $f$ satisfies UC but not STP. Take any two disjoint sets $S$ and $S'$ in $\mathcal{X}$ such that $f(S) = f(S')$. By the definition of $f$, they must be both finite or both infinite; in the former case, their union is a finite set, and $f(S \cup S') = 0$; if the latter case, the union is an infinite set, and so $f(S \cup S') = 1$. To see that STP does not hold, take the finest partition of $X$, i.e. $\pi = \{\{1\}, \{2\}, \dots\}$. Each block in $\pi$ is finite, so that $f(S) = 0$ for every $S \in \pi$; but $f(X) = 1$. To see that it is possible to agree to disagree, suppose there are two agents having information partitions $\pi_1 = \{\{1\}, \{2\}, \dots\}$ and $\pi_2 = \{X\}$. It is then immediate that $W_1 = W_2$ and $f_1 \neq f_2$.

3. Our analysis assumes that the communication structure is common knowledge. The graph $G$ is commonly known, and so is the dialogue that it gives rise to. This is crucial in order to have a well-defined learning process. When $i$ receives a message from $j$, she knows exactly whom $j$ talked with in the past and, consequently, she can infer what information $j$ learned from that history, even if $i$ does not necessarily know the actual message that $j$ sent to others or received in some states. Roughly speaking, a commonly known communication structure implies that the informational content of any given message is not ambiguous, so making it possible for people to learn. In the case in which the common knowledge assumption is relaxed, learning is not well-defined in our framework and thus convergence to consensus is not guaranteed. As Koessler (2001) and Tsakas and Voorneveld (2011) show, one needs to enlarge the state space so as to include any possible history of communication. In so doing, uncertainty about the communication structure can be dealt with in the enlarged state space. In other words, the fact that we keep the state space fixed throughout a dialogue is a direct consequence of having a commonly known communication structure.

   We also remark that the communication channel is assumed to be faultless and fully reliable. That is, when $i$ sends a message to $j$, that message is delivered to $j$ with absolute certainty, and it is common knowledge that it is so. In other words, we rule out the possibility that a message never reaches the intended recipient and also the possibility that a recipient gets a different message than what was sent by the sender. Should the communication channel be unreliable, we would be in a situation akin to the email game of Rubinstein (1989), where convergence to consensus is not guaranteed to hold.

4. Communication is not strategic. Adding strategic motives to our analysis is likely to alter our results substantially. In a different yet related setting, Anderlini et al. (2011) show that, while agents with common interests are able to aggregate their information

in a full learning equilibrium, no such an equilibrium can be sustained when interests diverge.

5. Since our goal is to establish whether a consensus emerges or not, we observe that our assumption of pairwise and simultaneous (within a round) communication is without loss of generality. As long as strong connectedness and symmetry in Assumption 3 are preserved, any modification of the communication network could affect only the speed of convergence or the fixed point contained in the dialogue. That is, a communication graph leads to a unique fixed point, but different graphs may lead to different fixed points, all of which induce a consensus.

   We emphasize that the need for two-way communication arises from the fact that we want to find a class of dialogues that induce a consensus for *every* possible profile of information partitions. It is not hard to find cases where a consensus can be reached even with unilateral communication. For instance, suppose that the initial information partitions are totally ordered. It suffices to let the agent with the finest partition send a message to everyone else and a consensus immediately ensues.

## 2.5 Conclusion

We have studied the problem of knowledge acquisition and convergence to consensus in a finite population of like-minded, fully rational individuals. We have showed that such a convergence is always possible provided that dialogues of transfinite length are allowed and that the communication structure is sufficiently rich. More generally, our results suggest that, at least in principle, knowledge acquisition between fully rational individuals is well-defined irrespective of the cardinality of the state space.

# Chapter 3

# Coordination through ambiguous language

## 3.1 Introduction

Correlated equilibrium is a solution concept that captures the impact of communication on strategic interaction. It does so without modeling explicitly the communication process in which players are involved. Differently put, correlated equilibrium "express[es] an assumption that players have implicit communication opportunities, in addition to the strategic options explicitly described in the game model" (Myerson, 1991, p. 245). All *implicit* communication opportunities are subsumed into canonical correlating devices which send private recommendations on how to play the game. But such devices, and the corresponding "equilibria[,] may have no natural interpretation" (Osborne and Rubinstein, 1994, p. 47).

Our goal in this paper is to provide an alternative construction of correlated equilibrium which, we believe, has a more natural interpretation than the canonical one. The main idea behind our construction is that correlated play can be induced by the ambiguity of the natural language through which players communicate. By ambiguity, we mean interpretive uncertainty stemming from the fact that words or sentences can have a plurality of meanings. Let us consider an example. Suppose a central banker delivers the following public speech: "If the GDP growth is sustained, then interest rates will be kept constant; otherwise they will be lowered". Firms listen to the speech in order to decide on their investments, which depend on future interest rates. But what is the true content of the banker's statement? More specifically, how should the antecedent "if the GDP growth is sustained" be interpreted? Is there a threshold $x$ such that if the actual growth rate $y$ is greater than $x$ then it is really the case that "the GDP growth is sustained"? One can argue that such a threshold $x$ does

exist but, unless its value is explicitly stipulated in some contract or convention, it is not necessarily unique. A firm $i$ may think that the threshold is $x_i$, whereas firm $j$ believes that it is $x_j \neq x_i$. In addition, either firm does not know what threshold the other is using to classify the growth rate as sustained or not. In sum, even if the banker's speech is public, it may convey differential information to those who hear it. Differently put, uncertainty about the interpretation of an ambiguous statement acts like a correlating device that sends private messages to players, so inducing differential information.

The backbone of our analysis is the following process of communication and coordination. In a given simultaneous-move game, players receive information about the prevailing state of the world at the pre-play stage. Information is not payoff relevant, and it may come in different varieties. It can be a public speech, a private signal, a sunspot, etc. Players have the opportunity to condition their play on the information received in the pre-play stage. They do so by means of a coordination strategy, which is a public list of conditional statements on how to play the game. In the banker's example, a coordination strategy can contain the following statements: "if interest rates will be kept constant, then only firm $i$ invests" and "if interest rates will be lowered, everybody invests". Firms publicly agree to follow the action recommendations contained in the coordination strategy. But due to language ambiguity, firms do not know with certainty how others will interpret the strategy recommendations and, therefore, how they will react to the banker's speech. As in the standard case, this uncertainty can sustain equilibrium payoffs that are outside the convex hull of Nash equilibrium payoffs of the underlying game.

Contrary to the standard construction of correlated equilibrium, we separate messages from their meaning. We do so by modeling explicitly the language through which players communicate. A language is a set of well-defined formulas that describe every relevant aspect of the world. To capture ambiguity, we use the logic of Halpern and Kets (2015). In it, truthfulness of formulas is defined relative to a player. Consequently, there can be states of the world where different players give different truth values to the very same formula. Players can disagree on a subset of formulas, namely those constructed as the conjunction or negation of primitive propositions, whereas the interpretation of probability formulas, i.e. beliefs, is the same for everybody. This means that every player is sophisticated enough to understand that others might be using different information partitions to form their beliefs.

Our main contribution is to provide a syntactic construction of correlated equilibrium. We consider two cases. In the first, we model the communication and coordination process illustrated in the central banker's example under the assumption that language is not ambiguous. We show that, for any finite game, any self-enforcing coordination strategy induces an objective correlated equilibrium distribution of the underlying game. In addition, any

objective correlated equilibrium distribution of the underlying game can be induced by some coordination strategy in some unambiguous epistemic structure capturing players' interpretations of formulas in the language. In the second case, we allow language to be ambiguous. We obtain the same characterization as in the unambiguous case with the proviso that equilibrium distributions are now subjective correlated equilibria. We thus show that language ambiguity provides a justification for heterogeneous beliefs about strategic play.

We illustrate the model in Section 3.2. It consists of three parts: the syntax (how formulas are formed), the semantics (how meaning to formulas is assigned), and the coordination process. Results are presented in Section 3.3, where the two cases of common-interpretation and ambiguous epistemic structures are treated separately.

### 3.1.1 Related literature

Correlated equilibrium is introduced in Aumann (1974). A reformulation of it in a decision-theoretic framework is provided in Aumann (1987). Our analysis is related to the following strands of the literature.

First, a classical literature initiated by Forges (1988, 1990) and Bárány (1992) studies whether and how correlated equilibrium can be obtained in a decentralized manner, i.e. without the help of a mediator. In our analysis, a mediator is not strictly necessary in that the information that players receive in the pre-play stage can be interpreted as a sunspot *à la* Cass and Shell (1983).

Lehrer (1996), Lehrer and Sorin (1997), and Di Tillio (2004) study public mediated talk in which correlation is achieved trough a machine that receives private inputs and sends out public recommendations. If we assume that the information in our model is provided by a mediator, then communication is always one-way, i.e. from the mediator to the players. Under this interpretation, correlation is achieved through uncertainty about the messages sent by the mediator. Players do not need to exchange messages with each other, nor do they need to send reports to the mediator.

Blume and Board (2013) examine strategic interaction under the assumption that players differ in their "language competence", i.e. their ability to use language. They model language explicitly. Other analyses aimed at modeling ambiguity (or vagueness) in natural language include Lipman (2009) and De Jaegher (2003). However, none of these papers use the syntactic approach as we do.

Our work is also related to the literature on epistemic foundations of solution concepts initiated by Aumann and Brandenburger (1995). The main goal of this literature is to find epistemic conditions that give rise to standard solution concepts. The approach is to model

explicitly how players reason about the game and, in particular, how they reason about the rationality of their opponents. Recent contributions in which correlated equilibrium is studied include Bach and Perea (2018) and Barelli (2009). Rather than rational play, the focus of our analysis is on how players reason about the realization of extraneous signals, and how this reasoning is affected by language ambiguity. All the contributions mentioned so far are carried out from a set-theoretic perspective. But another branch of the research program on epistemic foundations uses techniques from modal logic, as is done in Lorini and Schwarzentruber (2010) and Galeazzi and Lorini (2016). An extensive overview is provided in De Bruin (2010). To the best of our knowledge, our analysis would be the first to use modal logic to examine ambiguity about the interpretation of extraneous signals in games. As we already mentioned, we build on the logic of Halpern and Kets (2015). In particular, the syntax (Section 3.2.1) and the semantics (Section 3.2.2) are theirs.

## 3.2 Model

Let $G = (I, (A_i, u_i)_{i \in I})$ be a finite game with simultaneous moves. The set of players is $I = \{1, \ldots, n\}$. For every $i \in I$, $A_i$ is a non-empty, finite set of actions, and $u_i : \times_{j \in I} A_j \longrightarrow \mathbb{R}$ is the corresponding payoff function. As is standard, we define $A := \times_{i \in I} A_i$ and, for any $i$, $A_{-i} := \times_{j \neq i} A_j$.

Players coordinate their play in $G$ on the realizations of a payoff-irrelevant signal. In the pre-play stage, they agree on a list of instructions that tell them how to play the game conditional on signal observations. Players' reasoning about the game and the signals is captured by a formal language, which we are going to model explicitly. We describe the syntax in subsection 3.2.1, the semantics in subsection 3.2.2, and the coordination strategy in subsection 3.2.3.

### 3.2.1 Syntax

The fundamental object is a non-empty, countable set $\Phi$ of primitive propositions, with typical elements $p, q, \ldots$. Propositions in $\Phi$ describe non-epistemic aspects of the world. A *language* $\mathcal{L}(\Phi)$ is a set of well-formed formulas constructed from $\Phi$ through syntactic rules. Since no confusion should arise, from now on we omit the reference to $\Phi$ and write $\mathcal{L}$. The formulas contained in $\mathcal{L}$ determine the expressiveness of the language, i.e. the set of epistemic and non-epistemic aspects of the world that players can reason about. We construct $\mathcal{L}$ according to the following syntax:

- If $p \in \Phi$, then $p$ is a formula in $\mathcal{L}$;

- *Negation*: If $\varphi \in \mathcal{L}$, then $\neg\varphi$ ("not $\varphi$") is a formula in $\mathcal{L}$;
- *Conjunction*: If $\varphi, \psi \in \mathcal{L}$, then $\varphi \wedge \psi$ ("$\varphi$ and $\psi$") is a formula in $\mathcal{L}$;
- *Probability formulas*: If $\varphi_1, \ldots, \varphi_k \in \mathcal{L}$ and $b_1, \ldots, b_k, c \in \mathbb{R}$, then, for every $i \in I$,

$$b_1 \mathsf{pr}_i(\varphi_1) + \cdots + b_k \mathsf{pr}_i(\varphi_k) \geq c$$

  is a formula in $\mathcal{L}$. The intended reading of $\mathsf{pr}_i(\varphi) \geq x$ is "the probability that player $i$ ascribes to formula $\varphi$ is at least $x$";
- *Modal operator* $\mathsf{CB}$: If $\varphi \in \mathcal{L}$, then $\mathsf{CB}\varphi$ ("it is commonly believed that $\varphi$") is a formula in $\mathcal{L}$.

Probability formulas allow players to reason about beliefs and expected payoffs. We also want $\mathcal{L}$ to be sufficiently rich to describe how agents play the game $G$ and how they interpret the signals that they observe. Hence we assume that, for every $i \in I$, and for every $a_i \in A_i$, there is a primitive proposition $\mathsf{pl}_i a_i$ in $\Phi$. The intended reading of $\mathsf{pl}_i a_i$ is "$i$ chooses $a_i$" or, equivalently, "$i$ plays $a_i$". We assume that all these propositions describing choices are distinct elements, i.e. if $\mathsf{pl}_i a_i = \mathsf{pl}_j b_j$, then $i = j$ and $a_i = b_j$. Let $\Phi_G$ be the finite subset of $\Phi$ containing all such propositions about choices in $G$. In order to describe signals, let $\Phi^*$ be the set obtained by closing off $\Phi \backslash \Phi_G$ under negation and conjunction. Notice that formulas in $\Phi^*$ describe non-epistemic aspects of the world that are not payoff-relevant. We assume that there is a finite subset $\Sigma \subseteq \Phi^*$ of signals. Furthermore, if $\sigma \in \Sigma$, then $\{\mathsf{rec}_i\, \sigma : i \in I\} \subseteq \Phi$. The intended reading of $\mathsf{rec}_i\, \sigma$ is "$i$ has received signal $\sigma$".

We make use of the following **abbreviations**:

- *Implication*: $\varphi \implies \psi$ ("$\varphi$ implies $\psi$") is an abbreviation for $\neg(\varphi \wedge \neg\psi)$;
- *Belief operator*: $\mathsf{B}_i\varphi$ ("$i$ believes that $\varphi$") is an abbreviation for

$$(\mathsf{pr}_i(\varphi) \geq 1) \wedge (-\mathsf{pr}_i(\varphi) \geq -1)\,;$$

- *Mutual belief operator*: $\mathsf{EB}\varphi$ ("everybody believes that $\varphi$") is an abbreviation for $\wedge_{i \in I} \mathsf{B}_i\varphi$. In addition, we define $\mathsf{EB}^m\varphi$ ("$\varphi$ is $m$th-order mutual belief") recursively: $\mathsf{EB}^1\varphi = \mathsf{EB}\varphi$, and $\mathsf{EB}^m\varphi = \mathsf{EB}\left(\mathsf{EB}^{m-1}\varphi\right)$ for $m \geq 2$;

- $U_i(a_i)$ is the abbreviation for the probability formula

$$\sum_{(a_1,\ldots,a_{i-1},a_{i+1},\ldots,a_n)\in A_{-i}} u_i(a_1,\ldots,a_n)\,\mathsf{pr}_i\left(\mathsf{pl}_1 a_1 \wedge \cdots \wedge \mathsf{pl}_{i-1}a_{i-1} \wedge \mathsf{pl}_{i+1}a_{i+1} \wedge \cdots \wedge \mathsf{pl}_n a_n\right). \tag{3.1}$$

  The intended reading of $U_i(a_i)$ is "the expected payoff to $i$ from playing $a_i$". In order for this intended reading to be meaningful, the probabilities that $i$ ascribes to formulas in (3.1) must be non-negative and sum up to one. Under the assumptions we make in Subsections 3.2.2 and 3.2.3, it is always the case that these probabilities are well-defined, so making the reading of $U_i(a_i)$ as expected payoff unproblematic.

- $a_i^* = a_i$ is the abbreviation for

$$\wedge_{a_i'\in A_i}\left(U_i(a_i) \geq U_i(a_i')\right).$$

  The intended reading of $a_i^* = a_i$ is "$a_i$ is utility-maximizing".

- $\mathsf{rat}_i$ is the abbreviation for

$$\wedge_{a_i\in A_i}\left(\mathsf{pl}_i a_i \implies (a_i^* = a_i)\right). \tag{3.2}$$

  The intended reading of $\mathsf{rat}_i$ is "$i$ is rational". Notice that (3.2) is equivalent to saying that $i$ never chooses an action that is not utility-maximizing.

### 3.2.2 Semantics

We need a semantic model to assign meaning to formulas in $\mathcal{L}$. That is, we need a consistent set of rules to determine whether any given formula is true or false. The semantic model we use is an *epistemic probability structure* in which the interpretation of primitive propositions is player-dependent. Formally, an epistemic probability structure $M$ over $\Phi$ is a tuple $(\Omega, \mu, \{\pi_i\}_{i\in I}, \{H_i\}_{i\in I})$, where:

- $\Omega$ is a non-empty, finite set of states or possible worlds;
- $\mu$ is a common prior on (the power set of) $\Omega$;
- $\pi_i : \Omega \times \Phi \longrightarrow \{0,1\}$ is agent $i$'s interpretation function. Agent $i$ deems proposition $p$ as true in state $\omega$ if $\pi_i(\omega, p) = 1$, and false otherwise;
- $H_i$ is agent $i$'s information partition over $\Omega$, with typical element $h_i$. We write $h_i(\omega)$ to indicate the cell containing the states that $i$ considers as possible when the true state

is $\omega$. We make assumptions on how information partitions are determined by signals at the end of this subsection.

The finiteness of $\Omega$ is without loss of generality since we are confining ourselves to finite games. The player-dependent interpretation function $\pi_i$ captures language ambiguity: in a given world, different agents may assign different truth values to the very same primitive proposition. If $\pi_i = \pi_j$ for all $i, j \in I$, then we say that $M$ is a *common-interpretation* structure. The latter corresponds to the standard case without ambiguity where the interpretation of every formula is player-independent. If $M$ is not a common-interpretation structure, then we call it *ambiguous*.

Agents update beliefs through Bayes's rule. Given any event $E \subseteq \Omega$, agent $i$'s posterior belief about $E$ at $\omega$ is

$$\mu\left(E|h_i(\omega)\right) = \frac{\mu(E \cap h_i(\omega))}{\mu(h_i(\omega))}.$$

To ensure that posteriors are always well-defined, we assume that $\mu(h_i(\omega)) > 0$ for every state $\omega \in \Omega$ and every player $i \in I$.

Meaning to formulas in a structure $M$ is given inductively. The expression $(M, \omega, i) \vDash \varphi$ means that $\varphi$ holds at $\omega$ according to player $i$ in structure $M$. In addition, the intension of a formula $\varphi$ to player $i$ is $[[\varphi]]_i := \{\omega \in \Omega : (M, \omega, i) \vDash \varphi\}$, i.e. the set of states where $i$ deems $\varphi$ as true in structure $M$. Meaning to formulas is given as follows:

- If $p$ is a primitive proposition in $\Phi$, then $(M, \omega, i) \vDash p$ iff $\pi_i(\omega, p) = 1$;
- $(M, \omega, i) \vDash \varphi \wedge \psi$ iff $(M, \omega, i) \vDash \varphi$ and $(M, \omega, i) \vDash \psi$;
- $(M, \omega, i) \vDash \neg\varphi$ iff $(M, \omega, i) \nvDash \varphi$;
- $(M, \omega, i) \vDash b_1 \mathsf{pr}_j(\varphi_1) + \cdots + b_k \mathsf{pr}_j(\varphi_k) \geq c$ iff

$$b_1\, \mu\left([[\varphi_1]]_j|h_j(\omega)\right) + \cdots + b_k\, \mu\left([[\varphi_k]]_j|h_j(\omega)\right) \geq c; \tag{3.3}$$

- $(M, \omega, i) \vDash \mathsf{B}_j\varphi$ iff $\mu\left([[\varphi]]_j|h_j(\omega)\right) = 1$;
- $(M, \omega, i) \vDash \mathsf{CB}\varphi$ iff $(M, \omega, i) \vDash \mathsf{EB}^k\varphi$ for $k = 1, 2, \ldots$.

We emphasize that meaning to a formula is always given relative to a player. Due to language ambiguity, there can be states where different players assign different meaning to the very same formula. Formally, there can be states and formulas such that $(M, \omega, i) \vDash \varphi$ and $(M, \omega, j) \vDash \neg\varphi$ for some $i$ and $j$. However, players are fully sophisticated in that

they understand that others are using different information partitions to update beliefs[1]. Consequently, everybody agrees on the interpretation of probability formulas. As can be seen from (3.3), according to player $i$, agent $j$ assigns probability at least $c$ to a formula $\varphi$ if and only if the set of worlds where $\varphi$ holds according to $j$ has probability at least $c$ according to $j$. When the interpretation of a formula $\varphi$ is player-independent at a state $\omega$, we simplify notation and write $(M, \omega) \vDash \varphi$ instead of $(M, \omega, i) \vDash \varphi$ for all $i \in I$. In addition, we write $M \vDash \varphi$ when $(M, \omega, i) \vDash \varphi$ for every $\omega \in \Omega$ and every $i \in I$. In this case, we also say that $\varphi$ is *valid* in $M$.

We now make two assumptions about the interpretation of signals and information partitions.

**Assumption 4.** *For every $i, j \in I$,*

- *the collection*
$$\{[[\mathsf{rec}_i\, \sigma]]_j : \sigma \in \Sigma \text{ and } [[\mathsf{rec}_i\, \sigma]]_j \neq \emptyset\}$$
*is a partition of $\Omega$;*
- *for every $\sigma, \sigma' \in \Sigma$, if $[[\mathsf{rec}_i\, \sigma]]_j = [[\mathsf{rec}_i\, \sigma']]_j \neq \emptyset$, then $\sigma = \sigma'$.*

The assumption says that, according to any player, everyone receives one, and only one, signal at every state. Because of ambiguity, the event of $i$'s receiving signal $\sigma$ can be interpreted differently by different agents. For instance, it could be the case that $(M, \omega, i) \vDash \mathsf{rec}_i\, \sigma \wedge \neg \mathsf{rec}_i\, \sigma'$ and $(M, \omega, j) \vDash \mathsf{rec}_i\, \sigma' \wedge \neg \mathsf{rec}_i\, \sigma$, where $\sigma \neq \sigma'$. For ease of reference, we write $\sigma_{i,\omega}$ to denote the necessarily unique signal that $i$ thinks she is observing at state $\omega$.

**Assumption 5.** *For every $i \in I$ and every $\omega \in \Omega$,*

$$h_i(\omega) = [[\mathsf{rec}_i\, \sigma_{i,\omega}]]_i.$$

The assumption says that a player's information is determined by the signal she thinks she is observing. More specifically, the worlds that $i$ considers as possible at $\omega$ are all those where $i$ thinks that she is observing the same signal as in $\omega$. Notice that $\omega \in h_i(\omega)$ and, if $\omega' \in h_i(\omega)$, then $\sigma_{i,\omega} = \sigma_{i,\omega'}$.

The following example is meant to illustrate how one can use the main concepts introduced so far to capture ambiguity.

[1] This is the *innermost-scope semantics* of Halpern and Kets (2014).

**Example 1** There are two agents: $A$(nn) and $B$(ob). Suppose there is a primitive proposition $p \in \Phi$, whose intended reading is "the air temperature is extreme". According to Ann, temperatures are extreme if they are at most $x_A$ or at least $y_A$. According to Bob, temperatures are extreme if they are at most $x_B$ or at least $y_B$. Suppose $x_A < x_B < y_A < y_B$. The set of possible states of the world is represented in Table 3.1.

| State | Temperature | Ann | Bob |
|---|---|---|---|
| $\omega_1$ | $x_A$ | $\mathsf{rec}_A\, p$ | $\mathsf{rec}_B\, p$ |
| $\omega_2$ | $y_A$ | $\mathsf{rec}_A\, p$ | $\mathsf{rec}_B\, \neg p$ |
| $\omega_3$ | $x_B$ | $\mathsf{rec}_A\, \neg p$ | $\mathsf{rec}_B\, p$ |
| $\omega_4$ | $\frac{x_B+y_A}{2}$ | $\mathsf{rec}_A\, \neg p$ | $\mathsf{rec}_B\, \neg p$ |

Table 3.1: States of the world

Each state is a complete description of all the epistemic and non-epistemic aspects of the world. In state $\omega_1$, the actual temperature is $x_A$. Therefore, the proposition $p$ is deemed as true by both Ann and Bob. But in state $\omega_2$, $p$ is true according to Ann and false according to Bob. Their disagreement stems from language ambiguity. Since $p$ can be given a plurality of meanings, different agents may interpret it differently. We emphasize that, in ambiguous structures, there can be primitive propositions whose interpretation is not ambiguous at all. For instance, suppose that also the primitive proposition $q$ is in $\Phi$, where $q$ stands for "the air temperature is $x_A$". This proposition is unambiguous, and both Ann and Bob interpret it as true in state $\omega_1$ and false otherwise.

The true state of the world is observed through signals. Suppose that the set of signals is $\Sigma = \{p, \neg p\}$. In addition, each player receives $\sigma \in \Sigma$ in a given state if and only if he or she deems $\sigma$ as true in that state. Each row of Table 3.1 indicates the signals received by either player in the corresponding state. We assume that the interpretation of formulas of the form $\mathsf{rec}_i\sigma$ is not ambiguous. For instance, we have $(M, \omega_2) \vDash \mathsf{rec}_A\, p \wedge \mathsf{rec}_B \neg p$ even if $(M, \omega_2, A) \vDash p$ and $(M, \omega_2, B) \vDash \neg p$. In words, Ann thinks at $\omega_2$ that Bob receives the signal "the air temperature is not extreme" while she thinks that the temperature is actually extreme. We use the formulas of the form $\mathsf{rec}_i\sigma$ to obtain the following information partitions:

$$H_A = \{\{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}\}$$
$$H_B = \{\{\omega_1, \omega_3\}, \{\omega_2, \omega_4\}\}.$$

Suppose that the common prior $\mu$ is uniform over $\Omega$. We now want to make a few remarks on how agents form beliefs. We start by noticing that

$$(M, \omega_1, A) \vDash \mathsf{B}_A\, p \ \wedge \ \mathsf{B}_B\, p$$

$$(M, \omega_1, B) \vDash \mathsf{B}_A\, p \ \wedge\ \mathsf{B}_B\, p,$$

or, in compact notation, $(M, \omega_1) \vDash \mathsf{EB}\, p$. That is, everybody believes that $p$ at $\omega_1$. This follows from the fact that

$$\mu\left([[p]]_A | h_A(\omega_1)\right) = \mu\left(\{\omega_1, \omega_2\} | \{\omega_1, \omega_2\}\right) = 1$$
$$\mu\left([[p]]_B | h_B(\omega_1)\right) = \mu\left(\{\omega_1, \omega_3\} | \{\omega_1, \omega_3\}\right) = 1.$$

However, it holds that $(M, \omega_1) \vDash \neg \mathsf{B}_A \mathsf{B}_B\, p$. In words, Ann does not believe that Bob believes that $p$. Indeed we have

$$\mu\left([[\mathsf{B}_B\, p]]_A | h_A(\omega_1)\right) = \mu\left(\{\omega_1, \omega_3\} | \{\omega_1, \omega_2\}\right) = \frac{1}{2}.$$

Therefore, even if Ann and Bob receive the very same signal "the air temperature is extreme" in state $\omega_1$, it is not common belief between them that this is indeed the case. More specifically, the formula $p$ (and the formulas $\mathsf{rec}_A\, p$ and $\mathsf{rec}_B\, p$) is a first-order mutual belief at $\omega_1$, but it is not a second-order mutual belief. *A fortiori*, $p$ is not commonly believed. This shows how ambiguity generates higher-order uncertainty in the interpretation of formulas. Things would be different if the epistemic structure had common interpretation. Suppose that both agents have the same interpretation function as in Ann's column in Table 3.1. It is then immediate that, in state $\omega_1$, the formula $p$ (and the formulas $\mathsf{rec}_A\, p$ and $\mathsf{rec}_B\, p$) is not just first-order mutual belief but also common belief. Formally, $(M, \omega_1) \vDash \mathsf{CB}\, p$.

### 3.2.3 Coordination

We now describe how agents make choices in $G$. We start by assuming the following.

**Assumption 6.** *In any structure $M$, for every $i \in I$ and every $a_i \in A_i$,*

$$M \vDash \mathsf{pl}_i a_i \implies \wedge_{a_i' \neq a_i} \left(\neg \mathsf{pl}_i a_i'\right).$$

The assumption simply says that (it is commonly believed that) everyone does not play more than one action in each state.

Agents have the opportunity to coordinate their choices in $G$ through signals in $\Sigma$. More specifically, they can devise a *coordination strategy* $C$ that tells them how to play $G$ depending on the realizations of signals in $\Sigma$.

**Definition 1** (Coordination strategy)**.** *A coordination strategy $C$ is a finite subset of $\mathcal{L}$ such that:*

1. *for each player $i \in I$ and each signal $\sigma \in \Sigma$, there is a unique action $a_i \in A_i$ such that the formula $\mathsf{rec}_i\,\sigma \implies \mathsf{pl}_i a_i$ belongs to $C$;*

2. *for every $\varphi \in C$, $M \vDash \varphi$.*

A coordination strategy is a finite list of conditional propositions of the following form: "if $i$ receives signal $\sigma$, then $i$ plays action $a_i$", "if $j$ receives signal $\sigma'$, then $j$ plays action $a_j$", and so on. Notice that a strategy associates every signal with one, and only one, action for each player, but different signals may be associated with the same action recommendation. The strategy is public in that every formula contained in it is valid in $M$, hence it is common knowledge among everyone.

A coordination strategy is a set of instructions. Definition 1 ensures that such a set is complete, i.e. it provides everyone with an action recommendation for every signal realization that can possibly be observed. But it says nothing about the rationality, or lack thereof, of these recommended actions. Therefore, we want to restrict our analysis to epistemic structures, and coordination strategies, that meet minimal rationality requirements.

**Assumption 7** (Individual rationality)**.** *In any structure $M$, for every $i \in I$, it holds that $(M, i) \vDash \mathsf{rat}_i$.*

The assumption says that every $i$ chooses an action only if she deems it utility-maximizing. In other words, it is always true, according to player $i$, that $i$'s choices are utility-maximizing. As a consequence, $i$ always believes in her own rationality, and it is commonly believed that it is so. When the underlying epistemic structure has common interpretation, Assumption 7 is tantamount to assuming common belief in rationality, i.e. common belief in the event that everyone is rational.

**Remark 1.** *Let $M$ be a structure satisfying Assumption 7. Then we have:*

1. *$M \vDash \mathsf{CB}\left(\wedge_{i \in I} \mathsf{B}_i\left(\mathsf{rat}_i\right)\right)$;*

2. *If $M$ is a common-interpretation structure, then $M \vDash \mathsf{CB}\left(\wedge_{i \in I} \mathsf{rat}_i\right)$.*

*Proof.* By Assumption 7 and the definition of the belief operator, we have that, for every $i \in I$, $(M, i) \vDash \mathsf{B}_i\left(\mathsf{rat}_i\right)$. Since the interpretation of probability formulas is player-independent, the latter is equivalent to $M \vDash \mathsf{B}_i\left(\mathsf{rat}_i\right)$ for every $i \in I$. Therefore, the formula $\wedge_{i \in I} \mathsf{B}_i\left(\mathsf{rat}_i\right)$ is valid in $M$, and it is always common belief that it is a true formula.

Now suppose that $M$ is a common-interpretation structure. Thus we have that, for every $i, j \in I$, $(M, i) \vDash \mathsf{rat}_i$ if and only if $(M, j) \vDash \mathsf{rat}_i$. But then it is immediate to get $M \vDash \wedge_{i \in I} \mathsf{rat}_i$, from which the result follows. □

In an individually rational structure, any coordination strategy $C$ is self-enforcing in that no one has the incentive to disobey its action recommendations. Formally, for every $\omega$ and every $i$, there exists a unique signal $\sigma$ such that

$$(M, \omega, i) \vDash \mathsf{rec}_i\sigma \wedge \mathsf{pl}_i a_i \wedge (a_i^* = a_i),$$

where $a_i$ is the action prescribed by the formula $\mathsf{rec}_i\sigma \implies \mathsf{pl}_i a_i$ in $C$. To see why this is the case, observe the following. First, Assumptions 4 and 5 assure that there is a unique signal $\sigma = \sigma_{i,\omega}$ such that $(M, \omega, i) \vDash \mathsf{rec}_i\,\sigma$. Second, this signal is associated to a unique action by the coordination strategy $C$: there is a unique action $a_i$ with $\mathsf{rec}_i\,\sigma \implies \mathsf{pl}_i a_i$ in $C$ such that $M \vDash \mathsf{rec}_i\,\sigma \implies \mathsf{pl}_i a_i$. Third, the previous two points together imply $(M, \omega, i) \vDash \mathsf{pl}_i a_i$. Finally, by Assumption 7, we also get $(M, \omega, i) \vDash (a_i^* = a_i)$.

## 3.3 Results

Our goal is to characterize the probability distributions over $A$ that are induced by a given coordination strategy. Formally, every coordination strategy $C$ induces a profile $(\gamma_i)_{i\in I}$ of probability distributions over $A$. For every $i$, we define

$$\gamma_i(a_1, \ldots, a_n) := \mu\left([[\mathsf{pl}_1 a_1 \wedge \cdots \wedge \mathsf{pl}_n a_n]]_i\right) = \mu\left(\{\omega : (M, \omega, i) \vDash \mathsf{pl}_1 a_1 \wedge \cdots \wedge \mathsf{pl}_n a_n\}\right). \quad (3.4)$$

Each $\gamma_i$ is a well-defined probability distribution. First, it is clear from (3.4) that $\gamma_i(a) \geq 0$ for every $a \in A$. Second, by Assumptions 4 and 5, and Definition 1, for every $\omega$ there exists an action profile $a \in A$ such that $(M, \omega, i) \vDash \mathsf{pl}_1 a_1 \wedge \cdots \wedge \mathsf{pl}_n a_n$. Third, by Assumption 6, $a \neq a'$ implies that

$$[[\mathsf{pl}_1 a_1 \wedge \cdots \wedge \mathsf{pl}_n a_n]]_i \cap [[\mathsf{pl}_1 a_1' \wedge \cdots \wedge \mathsf{pl}_n a_n']]_i = \emptyset.$$

Therefore we have that $\sum_{a\in A} \gamma_i(a) = 1$. From now on, when no confusion should arise, we abuse notation and write $\mathsf{pl}a$ instead of $\mathsf{pl}_1 a_1 \wedge \cdots \wedge \mathsf{pl}_n a_n$, and $\mathsf{pl}_{-i} a_{-i}$ instead of $\mathsf{pl}_1 a_1 \wedge \cdots \wedge \mathsf{pl}_{i-1} a_{i-1} \wedge \mathsf{pl}_{i+1} a_{i+1} \wedge \cdots \mathsf{pl}_n a_n$.

### 3.3.1 Common-interpretation structures

Let us consider first the case of common-interpretation structures. It is clear that, under common interpretation, $\gamma_i = \gamma_j$ for every $i, j \in I$. Thus we simplify things by dropping the

subscript $i$. For any $a \in A$ we can write:

$$\gamma(a_1, \ldots, a_n) = \mu\left([[\mathsf{pl}_1 a_1 \wedge \cdots \wedge \mathsf{pl}_n a_n]]\right).$$

Recall that a probability distribution $\gamma \in \Delta(A)$ is a **correlated equilibrium** of $G$ if, for every $i \in I$ and every $a_i \in A_i$,

$$\sum_{a_{-i} \in A_{-i}} \left[u_i(a_i, a_{-i}) - u_i(a_i', a_{-i})\right] \gamma(a_i, a_{-i}) \geq 0 \quad \text{for every } a_i' \in A_i.$$

We can now establish the first result.

**Proposition 4.** *Let $M$ be a common-interpretation epistemic structure satisfying Assumptions 4-7. Then any coordination strategy induces a correlated equilibrium of $G$.*

*Proof.* The argument is standard. Suppose $\sum_{a_{-i} \in A_{-i}} \gamma(a_i, a_{-i}) > 0$. Then we have:

$$\sum_{a_{-i} \in A_{-i}} \left[u_i(a_i, a_{-i}) - u_i(a_i', a_{-i})\right] \gamma(a_i, a_{-i}) \propto \sum_{a_{-i} \in A_{-i}} \left[u_i(a_i, a_{-i}) - u_i(a_i', a_{-i})\right] \gamma(a_{-i} | a_i), \quad (3.5)$$

and the right hand side of (3.5) is equal to

$$\sum_{a_{-i} \in A_{-i}} \left[u_i(a_i, a_{-i}) - u_i(a_i', a_{-i})\right] \mu\left([[\mathsf{pl}_{-i} a_{-i}]] | [[\mathsf{pl}_i a_i]]\right). \quad (3.6)$$

Now we argue that the event $[[\mathsf{pl}_i a_i]]$ in (3.6) is the union of some cells of $H_i$. Assumptions 4 and 5 imply that, for every cell $h_i \in H_i$, there exists a unique signal $\sigma \in \Sigma$ such that $(M, \omega) \vDash \mathsf{rec}_i\, \sigma$ for every $\omega \in h_i$. Combining this with Definition 1 and Assumption 6, we can conclude that, for every $h_i \in H_i$, there exists a unique action $a_i' \in A_i$ such that $(M, \omega) \vDash \mathsf{pl}_i\, a_i'$ for every $\omega \in h_i$.

Since $[[\mathsf{pl}_i a_i]]$ can be written as the union of some cells of $H_i$, and by the law of total probability, we can write

$$\mu\left([[\mathsf{pl}_{-i} a_{-i}]] | [[\mathsf{pl}_i a_i]]\right) = \sum_{\{h_i \in H_i : h_i \subseteq [[\mathsf{pl}_i a_i]]\}} \mu\left([[\mathsf{pl}_{-i} a_{-i}]] | h_i\right) \mu\left(h_i | [[\mathsf{pl}_i a_i]]\right).$$

Substituting in (3.6) and rearranging yields

$$\sum_{\{h_i \in H_i : h_i \subseteq [[\mathsf{pl}_i a_i]]\}} \sum_{a_{-i} \in A_{-i}} \left[u_i(a_i, a_{-i}) - u_i(a_i', a_{-i})\right] \times \left[\mu\left([[\mathsf{pl}_{-i} a_{-i}]] | h_i\right) \mu\left(h_i | [[\mathsf{pl}_i a_i]]\right)\right]. \quad (3.7)$$

By Assumption 7, for every $\omega \in h_i \subseteq [[\mathsf{pl}_i a_i]]$, we have that $(M, \omega) \vDash \mathsf{pl}_i a_i \wedge (a_i^* = a_i)$.

Therefore, (3.7) is non-negative, so proving the claim. □

The result says that, in common-interpretation structures, self-enforcing coordination strategies always lead to an objective correlated equilibrium of the underlying game. The result can be interpreted as a syntactic version of the classical analysis of Aumann (1987). The role of common-interpretation can be described as follows. Even if different agents may receive different signals in the same state, everyone agrees on the profile of actions that is being played at that state. It is never the case that $i$ thinks that $j$ is playing $a_j$ whereas $k$ thinks that $j$ is playing $b_j$ in a given state. Differently put, agents can have different information but they all share the same model or view of the world, so ruling out any form of fundamental disagreement.

The next result is about the opposite direction, namely from correlated equilibria to epistemic structures.

**Proposition 5.** *Let $\gamma$ be a correlated equilibrium of $G$. Then there exist an individually rational, common-interpretation structure $M$, a set of signals $\Sigma$, and a coordination strategy $C$ that induce $\gamma$.*

*Proof.* Suppose $\gamma$ is a correlated equilibrium of $G$. Let $A^* \subseteq A$ be the support of $\gamma$. We define a common-interpretation structure $M = (\Omega, \mu, \pi, \{H_i\}_{i \in I})$ by constructing one state $\omega_a$ for each action profile $a \in A^*$, so that $\Omega = \{\omega_a : a \in A^*\}$. The prior corresponds with the correlated equilibrium: for each state $\omega_a$, we set $\mu(\omega_a) = \gamma(a)$. To define the interpretation function $\pi$ and the information partitions $\{H_i\}_{i \in I}$, we first need to say more about formulas in the language $\mathcal{L}$.

Fix a set $\Sigma$ of signals such that $|\Sigma| = \max_{i \in I} |A_i|$. This allows us to choose, for each player $i \in I$, an injective function $s_i : A_i \longrightarrow \Sigma$ that we use to assign signals to players. Since each $s_i$ is injective, distinct actions correspond to different signals. The interpretation function is a function $\pi : \Omega \times \Phi \longrightarrow \{0, 1\}$ such that, for all $\omega_a \in \Omega$, $i \in I$, $\sigma \in \Sigma$, and $b_i \in A_i$,

$$\pi(\omega_a, \mathsf{rec}_i \sigma) = \begin{cases} 1 & \text{if } \sigma = s_i(a_i), \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \pi(\omega_a, \mathsf{pl}_i b_i) = \begin{cases} 1 & \text{if } b_i = a_i, \\ 0 & \text{otherwise.} \end{cases}$$

This implies that, for each state $\omega_a$ and each player $i$, the formula $\mathsf{rec}_i s_i(a_i) \wedge \mathsf{pl}_i a_i$ is deemed true at $\omega_a$. For each player $i$, the information partition $H_i$ is defined so that, for each state $\omega_a \in \Omega$, the cell $h_i(\omega_a)$ contains all the states where $i$ receives the same signal. By definition of $s_i$, we have

$$h_i(\omega_a) = \{\omega_b \in \Omega : s_i(b_i) = s_i(a_i)\} = \{\omega_b \in \Omega : b_i = a_i\}. \tag{3.8}$$

One can easily verify that the structure $M$ constructed thus far satisfies Assumptions 4-6. To show that $M$ is individually rational, suppose that $(M, \omega_a) \vDash \mathsf{pl}_i a_i$. Player $i$'s expected payoff at $\omega_a$ from playing $a_i$ is

$$\begin{aligned}
U_i(a_i) &= \sum_{a_{-i} \in A_{-i}} u_i(a_i, a_{-i}) \mu \left( [[\mathsf{pl}_{-i} a_{-i}]] | h_i(\omega_a) \right) \\
&\propto \sum_{a_{-i} \in A_{-i}} u_i(a_i, a_{-i}) \mu \left( [[\mathsf{pl}_{-i} a_{-i}]] \cap h_i(\omega_a) \right) \\
&= \sum_{a_{-i} \in A_{-i}} u_i(a_i, a_{-i}) \mu \left( [[\mathsf{pl}_{-i} a_{-i}]] \cap [[\mathsf{pl}_i a_i]] \right) \\
&= \sum_{a_{-i} \in A_{-i}} u_i(a_i, a_{-i}) \mu \left( [[\mathsf{pl}_1 a_1 \wedge \cdots \wedge \mathsf{pl}_n a_n]] \right) \\
&= \sum_{a_{-i} \in A_{-i}} u_i(a_i, a_{-i}) \gamma \left( a_i, a_{-i} \right) .
\end{aligned}$$

Therefore, since $\gamma$ is a correlated equilibrium by assumption, we can conclude that $(M, \omega) \vDash a_i^* = a_i$.

Finally, we need to construct a coordination strategy $C$ that induces $\gamma$. For each $i \in I$ and each $\sigma \in \Sigma$, if $\sigma$ is in the range of $s_i$, then the formula $\mathsf{rec}_i \sigma \implies \mathsf{pl}_i a_i$, with $a_i = s_i^{-1}(\sigma)$, is in $C$. If $\sigma$ is not in the range of $s_i$, then pick an arbitrary $a_i' \in A_i$ and add the formula $\mathsf{rec}_i \sigma \implies \mathsf{pl}_i a_i'$ to $C$. One can easily verify that $C$ is indeed a coordination strategy as per Definition 1, and that it induces the correlated equilibrium $\gamma$. □

In the following example, we illustrate the construction that we have just used in proving Proposition 5.

**Example 2** Consider the base game $G$ in Figure 3.1.

| | $L$ | $C$ | $R$ |
|---|---|---|---|
| $T$ | $0, 0$ | $2, 1$ | $1, 2$ |
| $M$ | $1, 2$ | $0, 0$ | $2, 1$ |
| $B$ | $2, 1$ | $1, 2$ | $0, 0$ |

Figure 3.1: The base game.

This game has a unique Nash equilibrium in which either player randomizes uniformly over her available strategies. Consider the correlated equilibrium $\gamma$ that puts weight $\frac{1}{6}$ on every action profile which gives strictly positive payoffs. We want to find an individually rational, common-interpretation structure that induces such an equilibrium. We start by

noticing that the support of $\gamma$ is the following:

$$A^* = \{(T,C),(T,R),(M,L),(M,R),(B,L),(B,C)\}\,.$$

Let the state space be $\Omega = \{\omega_a : a \in A^*\}$. The common prior over $\Omega$ is uniform. Fix a set of signals $\Sigma = \{\sigma, \sigma', \sigma''\}$. We assign signals to players through functions $s_i : A_i \longrightarrow \Sigma$, with $i = 1, 2$, such that

$$(s_1(T), s_1(M), s_1(B)) = (s_2(L), s_2(C), s_2(R)) = (\sigma, \sigma', \sigma'').$$

The interpretation function is a function $\pi : \Omega \times \Phi \longrightarrow \{0,1\}$ which satisfies the truth assignments contained in the following table.

| $\boldsymbol{\pi}$ | $\omega_{TC}$ | $\omega_{TR}$ | $\omega_{ML}$ | $\omega_{MR}$ | $\omega_{BL}$ | $\omega_{BC}$ |
|---|---|---|---|---|---|---|
| $\mathsf{rec}_1\sigma$ | 1 | 1 | 0 | 0 | 0 | 0 |
| $\mathsf{rec}_1\sigma'$ | 0 | 0 | 1 | 1 | 0 | 0 |
| $\mathsf{rec}_1\sigma''$ | 0 | 0 | 0 | 0 | 1 | 1 |
| $\mathsf{rec}_2\sigma$ | 0 | 0 | 1 | 0 | 1 | 0 |
| $\mathsf{rec}_2\sigma'$ | 1 | 0 | 0 | 0 | 0 | 1 |
| $\mathsf{rec}_2\sigma''$ | 0 | 1 | 0 | 1 | 0 | 0 |
| $\mathsf{pl}_1 T$ | 1 | 1 | 0 | 0 | 0 | 0 |
| $\mathsf{pl}_1 M$ | 0 | 0 | 1 | 1 | 0 | 0 |
| $\mathsf{pl}_1 B$ | 0 | 0 | 0 | 0 | 1 | 1 |
| $\mathsf{pl}_2 L$ | 0 | 0 | 1 | 0 | 1 | 0 |
| $\mathsf{pl}_2 C$ | 1 | 0 | 0 | 0 | 0 | 1 |
| $\mathsf{pl}_2 R$ | 0 | 1 | 0 | 1 | 0 | 0 |

By (3.8), information partitions are defined as follows:

$$H_1 = \{\{\omega_{TC}, \omega_{TR}\}, \{\omega_{ML}, \omega_{MR}\}, \{\omega_{BL}, \omega_{BC}\}\}$$
$$H_2 = \{\{\omega_{TC}, \omega_{BC}\}, \{\omega_{TR}, \omega_{MR}\}, \{\omega_{ML}, \omega_{BL}\}\}\,.$$

It is straightforward to verify that the structure $M$ constructed so far is individually rational.

Finally, the coordination strategy $C$ that induces $\gamma$ can be defined as follows:

$$C = \{\mathsf{rec}_i s_i(a_i) \implies \mathsf{pl}_i a_i : i \in \{1,2\} \text{ and } a_i \in A_i\}\,.$$

Notice that $C$ contains six formulas, and all of them are true in *every* state of the world. This means that it is never the case that, say, player 1 receives signal $\sigma'$ and plays action $T$.

### 3.3.2 Ambiguous structures

We now characterize the equilibrium distributions induced by epistemic structures that are possibly ambiguous. Recall that a profile of probability distributions $(\gamma_1, \ldots, \gamma_n)$ over $A$ is a **subjective correlated equilibrium** of $G$ if, for every $i \in I$ and every $a_i \in A_i$,

$$\sum_{a_{-i} \in A_{-i}} [u_i(a_i, a_{-i}) - u_i(a'_i, a_{-i})] \, \gamma_i(a_i, a_{-i}) \geq 0 \;\text{ for every } a'_i \in A_i.$$

Then we have the following.

**Proposition 6.** *Let $M$ be an epistemic structure satisfying Assumptions 4-7. Then any coordination strategy induces a subjective correlated equilibrium of $G$.*

*Proof.* The argument is the same as in the proof of Proposition 4 with the proviso that, for every $i \in I$, one uses the following decomposition of conditional probabilities:

$$\begin{aligned}\gamma_i(a_{-i}|a_i) &= \mu\left([[\mathsf{pl}_{-i}a_{-i}]]_i | [[\mathsf{pl}_i a_i]]_i\right) \\ &= \sum_{\{h_i \in H_i : h_i \subseteq [[\mathsf{pl}_i a_i]]_i\}} \mu\left([[\mathsf{pl}_{-i}a_{-i}]]_i | h_i\right) \mu\left(h_i | [[\mathsf{pl}_i a_i]]_i\right).\end{aligned}$$

□

The result can be interpreted as follows. Even if players agree on a coordination strategy, and even if they share a common prior, language ambiguity may cause them to ascribe different probabilities to the same event, so leading to inconsistent beliefs. Players may disagree on what action profile is being played in a given state. Contrary to common-interpretation structures, it may well be the case that $i$ thinks that $j$ is playing $a_j$ whereas $k$ thinks that $j$ is playing $b_j$ in a given state. Differently put, agents may have different views of the world stemming from a fundamental disagreement about the interpretation of (some) primitive propositions. We illustrate this point in the next example, where we describe an ambiguous structure whose induced equilibrium distributions are a subjective correlated equilibrium but not an objective correlated equilibrium.

**Example 3** Consider the elementary coordination game in Figure 3.2.

| | $L$ | $R$ |
|---|---|---|
| $U$ | $1,1$ | $0,0$ |
| $D$ | $0,0$ | $1,1$ |

Figure 3.2: The base game.

Suppose that the epistemic structure is the same as in Example 4.3 of Halpern and Kets (2015). The state space is $\Omega = \{\omega, \omega'\}$, and the common prior is uniform. The set of signals is $\Sigma = \{\sigma, \sigma'\}$. Players disagree on the interpretation of signals. Values of their interpretation functions are reported in the following tables.

| $\boldsymbol{\pi_1}$ | $\omega$ | $\omega'$ |
|---|---|---|
| $\mathsf{rec}_1\sigma$ | 1 | 0 |
| $\mathsf{rec}_1\sigma'$ | 0 | 1 |
| $\mathsf{rec}_2\sigma$ | 1 | 0 |
| $\mathsf{rec}_2\sigma'$ | 0 | 1 |
| $\mathsf{pl}_1 U$ | 1 | 0 |
| $\mathsf{pl}_1 D$ | 0 | 1 |
| $\mathsf{pl}_2 L$ | 1 | 0 |
| $\mathsf{pl}_2 R$ | 0 | 1 |

| $\boldsymbol{\pi_2}$ | $\omega$ | $\omega'$ |
|---|---|---|
| $\mathsf{rec}_1\sigma$ | 1 | 1 |
| $\mathsf{rec}_1\sigma'$ | 0 | 0 |
| $\mathsf{rec}_2\sigma$ | 1 | 1 |
| $\mathsf{rec}_2\sigma'$ | 0 | 0 |
| $\mathsf{pl}_1 U$ | 1 | 1 |
| $\mathsf{pl}_1 D$ | 0 | 0 |
| $\mathsf{pl}_2 L$ | 1 | 1 |
| $\mathsf{pl}_2 R$ | 0 | 0 |

Thus we have:

$$[[\mathsf{rec}_1\sigma]]_1 = \{\omega\} \qquad [[\mathsf{rec}_2\sigma]]_1 = \{\omega\}$$
$$[[\mathsf{rec}_1\sigma']]_1 = \{\omega'\} \qquad [[\mathsf{rec}_2\sigma']]_1 = \{\omega'\}$$

for player 1 and

$$[[\mathsf{rec}_1\sigma]]_2 = \{\omega, \omega'\} \qquad [[\mathsf{rec}_2\sigma]]_2 = \{\omega, \omega'\}$$
$$[[\mathsf{rec}_1\sigma']]_2 = \emptyset \qquad [[\mathsf{rec}_2\sigma']]_2 = \emptyset$$

for player 2. In words, each agent thinks that the other always receives the same signal as hers. Information partitions are given by:

$$H_1 = \{\{\omega\}, \{\omega'\}\}$$
$$H_2 = \{\{\omega, \omega'\}\}.$$

Now suppose that the coordination strategy $C$ contains the following four formulas:

$$\begin{aligned}
\mathsf{rec}_1\sigma &\implies \mathsf{pl}_1\, U \\
\mathsf{rec}_2\sigma &\implies \mathsf{pl}_2\, L \\
\mathsf{rec}_1\sigma' &\implies \mathsf{pl}_1\, D \\
\mathsf{rec}_2\sigma' &\implies \mathsf{pl}_2\, R.
\end{aligned}$$

One can verify that $C$ is self-enforcing. The induced subjective probability distributions over $A$ are $\gamma_1(U, L) = \gamma_1(D, R) = \frac{1}{2}$ and $\gamma_2(U, L) = 1$. Finally, we observe that $(M, \omega', 2) \vDash \mathsf{rec}_1\sigma \wedge \mathsf{pl}_1 U \wedge \neg(a_1^* = U)$. In words, 2 thinks that 1 is choosing an action that is not utility-maximizing. The reason is that, as we pointed out in Remark 1, individual rationality in ambiguous structures does not entail common belief in rationality.

The next result is about the opposite direction, namely from subjective correlated equilibria to epistemic structures that induce them.

**Proposition 7.** *Let $(\gamma_1, \ldots, \gamma_n)$ be a subjective correlated equilibrium of $G$. Then there exist an individually rational epistemic structure $M$, a set of signals $\Sigma$, and a coordination strategy $C$ that induce $(\gamma_1, \ldots, \gamma_n)$.*

*Proof.* The argument follows the same logic as in the proof of Proposition 5. Suppose $(\gamma_1, \ldots, \gamma_n)$ is a subjective correlated equilibrium of $G$. For every $i \in I$, let $A_*^i \subseteq A$ be the support of $\gamma_i$. Define $A_* := \times_{i \in I} A_*^i$. Elements $a$ in $A_*$ are profiles of action profiles, and we use the following notation:

$$a = \left(a^1, \ldots, a^n\right) = \left((a_1^1, \ldots, a_n^1), \ldots, (a_1^n, \ldots, a_n^n)\right).$$

We define an epistemic probability structure $M = (\Omega, \mu, \{\pi_i\}_{i \in I}, \{H_i\}_{i \in I})$ as follows. We construct one state $\omega_a$ for each element $a \in A_*$, so that $\Omega = \{\omega_a : a \in A_*\}$. The common prior $\mu$ over $\Omega$ is constructed as a product measure: for every $\omega_a$, we let $\mu(\omega_a) = \prod_{i=1}^n \gamma_i(a^i)$.

Now fix a set $\Sigma$ of signals such that $|\Sigma| = \max_{i \in I} |A_i|$. For each $i \in I$, we can choose an injective function $s_i : A_i \longrightarrow \Sigma$ that we use to assign signals to players. Since each $s_i$ is injective, distinct actions correspond to different signals. For each $i \in I$, the interpretation function is a function $\pi_i : \Omega \times \Phi \longrightarrow \{0, 1\}$ such that, for all $\omega_a \in \Omega$, $j \in I$, $\sigma \in \Sigma$, and $b_j \in A_j$,

$$\pi_i(\omega_a, \mathsf{rec}_j\sigma) = \begin{cases} 1 & \text{if } \sigma = s_j(a_j^i), \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \pi_i(\omega_a, \mathsf{pl}_j b_j) = \begin{cases} 1 & \text{if } b_j = a_j^i, \\ 0 & \text{otherwise.} \end{cases} \tag{3.9}$$

The above definition implies that, for each state $\omega_a$, and each player $i$, the formulas $\left\{\mathsf{rec}_j s_j(a_j^i) \wedge \mathsf{pl}_j a_j^i : j \in I\right\}$ are deemed as true at $\omega_a$ by player $i$. For each $i \in I$, the information partition $H_i$ is defined so that, for each state $\omega_a \in \Omega$, the cell $h_i(\omega_a)$ contains all the states where, according to $i$, $i$ receives the same signal. By definition of $s_i$, we have

$$h_i(\omega_a) = \left\{\omega_b \in \Omega : s_i(b_i^i) = s_i(a_i^i)\right\} = \left\{\omega_b \in \Omega : b_i^i = a_i^i\right\}. \tag{3.10}$$

One can easily verify that the structure $M$ constructed thus far satisfies Assumptions 4-6. To show that $M$ is individually rational, suppose that $(M, \omega_a, i) \vDash \mathsf{pl}_i a_i$. By (3.9), $a_i = a_i^i$. Player $i$'s expected payoff at $\omega_a$ from playing $a_i$ is

$$\begin{aligned}
U_i(a_i) &= \sum_{a_{-i} \in A_{-i}} u_i(a_i, a_{-i}) \mu\left([[\mathsf{pl}_{-i} a_{-i}]]_i | h_i(\omega_a)\right) \\
&\propto \sum_{a_{-i} \in A_{-i}} u_i(a_i, a_{-i}) \mu\left([[\mathsf{pl}_{-i} a_{-i}]]_i \cap h_i(\omega_a)\right) \\
&= \sum_{a_{-i} \in A_{-i}} u_i(a_i, a_{-i}) \mu\left([[\mathsf{pl}_{-i} a_{-i}]]_i \cap [[\mathsf{pl}_i a_i]]_i\right) \\
&= \sum_{a_{-i} \in A_{-i}} u_i(a_i, a_{-i}) \mu\left([[\mathsf{pl}_1 a_1 \wedge \cdots \wedge \mathsf{pl}_n a_n]]_i\right) \\
&= \sum_{a_{-i} \in A_{-i}} u_i(a_i, a_{-i}) \mu\left(\{\omega_b \in \Omega : b^i = (a_i, a_{-i})\}\right).
\end{aligned} \tag{3.11}$$

By definition of the common prior $\mu$, we have that $\mu\left(\{\omega_b \in \Omega : b^i = (a_i, a_{-i})\}\right)$ is equal to

$$\gamma_i(a_i, a_{-i}) \left[ \sum_{(b^1, \ldots, b^{i-1}, b^{i+1}, \ldots, b^n) \in \times_{j \neq i} A_*^j} \gamma_1(b^1) \times \cdots \times \gamma_{i-1}(b^{i-1}) \times \gamma_{i+1}(b^{i+1}) \times \cdots \times \gamma_n(b^n) \right], \tag{3.12}$$

which simplifies to $\gamma_i(a_i, a_{-i})$. Substituting in (3.11), we obtain

$$U_i(a_i) \propto \sum_{a_{-i} \in A_{-i}} u_i(a_i, a_{-i}) \gamma_i\left(a_i, a_{-i}\right).$$

Therefore, since $\gamma_i$ is part of a subjective correlated equilibrium by assumption, we can conclude that $(M, \omega_a, i) \vDash \mathsf{pl}_i a_i \wedge (a_i^* = a_i)$.

It remains to construct a coordination strategy $C$ that induces $(\gamma_1, \ldots, \gamma_n)$. For each $i \in I$ and each $\sigma \in \Sigma$, if $\sigma$ is in the range of the injective function $s_i$, then the formula $\mathsf{rec}_i \sigma \implies \mathsf{pl}_i a_i$, with $a_i = s_i^{-1}(\sigma)$, is in $C$. If $\sigma$ is not in the range of $s_i$, then pick an arbitrary $a_i' \in A_i$ and add the formula $\mathsf{rec}_i \sigma \implies \mathsf{pl}_i a_i'$ to $C$. One can easily verify that $C$

is indeed a coordination strategy as per Definition 1. Finally, to show that it induces the subjective correlated equilibrium $(\gamma_1, \ldots, \gamma_n)$, by using (3.12), we have that, for every $i \in I$ and every $(a_1, \ldots, a_n) \in A$,

$$\mu\left([[\mathsf{pl}_1 a_1 \wedge \cdots \wedge \mathsf{pl}_n a_n]]_i\right) = \mu\left(\{\omega_b \in \Omega : b^i = (a_1, \ldots, a_n)\}\right) = \gamma_i(a_1, \ldots, a_n),$$

so ending the proof. □

## 3.4 Discussion

We use the word ambiguity as it is done in linguistics, where it expresses the fact that the map from sentences to meanings is multi-valued. Our analysis has nothing to do with ambiguity in the decision-theoretic sense of not knowing the "true" probability distribution of a certain event. We take ambiguity as a given and do not model why different players can assign different truth values to the very same formula. Our interpretation is that ambiguity is a structural property of natural language. Differently put, the map from sentences to meanings induced by any natural language is not commonly known. The gist of our analysis is that players can agree on sentences and, at the same time, disagree on meanings. When strategic interaction is conditioned on sentences, it is the uncertainty about their meanings that acts as a correlating device.

Introduced by Aumann (1974), the standard model for correlated equilibrium is set in an event-based epistemic framework. That is, players reason about events which are represented as subsets of a given state space. The language through which agents describe events is not modeled explicitly. Adopting a syntactic approach, our analysis consists in enriching the standard model for correlated equilibrium with a formal language. As a consequence, agents' reasoning about the world and, in particular, the game they are going to play is now expressed through formulas; the state space is a representation of how agents assign meaning to formulas. The standard model can be seen as a reduced form model of the syntactic approach we use. A comparison between the event-based and the syntactic epistemic frameworks, but without game theoretic applications and without language ambiguity, can be found in Halpern (2003).

Players are assumed to be fully rational. Even if they have different interpretation functions, they fully understand that the interpretation of probability formulas, hence beliefs, is not the same for everyone. In addition, their information is always partitional. We refer to Brandenburger et al. (1992) for a construction of correlated equilibrium with boundedly rational players. In their model, players make systematic mistakes in processing information,

so leading to non-partitional information functions. They show that information processing errors are equivalent to introduce "subjectivity" in beliefs. What we show in this paper is that, without any information processing error, ambiguity in natural language provides a justification for heterogeneous beliefs.

At first blush, it might be surprising that players having a common prior over a fixed state space end up having different subjective beliefs about their play in the underlying game. The reason why this is the case can be explained as follows. The state space $\Omega$ can be seen as a collection of $n$ models about the world. When all of these models are exactly the same, then their "projection" over $A$ is obviously the same for everybody. But when the subjective models differ, different players might have different projections over $A$, because the event $(a_1, \ldots, a_n)$ is not the same for everybody (i.e. the set of states where it holds is not the same for everyone). Differently put, language ambiguity induces subjectivity in how players reason about their choices in $G$.

## 3.5 Conclusion

We have examined how players can coordinate their choices when the language through which they communicate is possibly ambiguous. The gist of our results is that, when players publicly agree to condition their play on a set of sentences, the meaning of these sentences is not necessarily commonly known because of language ambiguity. The resulting uncertainty acts as a correlating device, so inducing correlated equilibrium distributions over outcomes. We believe that our analysis also provides a justification for why subjective beliefs about strategic play may not be consistent.

# Chapter 4

# Persuading a committee with privately known preferences

## 4.1 Introduction

It is often the case that decision-makers have to rely on the information provided by self-interested third parties. Juries reach a verdict on the basis of the information provided by a prosecutor; voters are persuaded to vote for a certain candidate through political campaigns; hiring committees review candidates based on the evidence they provide about their skills; and so on. A recent and fruitful approach to model this kind of situations goes under the name of Bayesian Persuasion or Information Design. The basic idea is that the information designer chooses a disclosure policy before observing the realization of some fundamental parameter of interest. Existing research shows that Information Design is a rather powerful tool in that it allows the designer to attain a relatively large set of potential outcomes. However, this hinges upon a number of rather strong assumptions. For instance, it is typically assumed that decision-makers' preferences are commonly known.

In this paper we study the Information Design problem of a persuader who wants to induce a committee to take a certain action. Contrary to existing models, we assume that committee members' preferences about the option to implement are not commonly known. Examples of such situations are not hard to find. A prosecutor may be uncertain about the composition of the jury she is addressing. More specifically, she may not know if a given juror is relatively tough or lenient. Similarly, a job candidate may not be sure about the composition of the hiring committee who is reviewing her job application in that different members might be expert in different domains. The fundamental question we address is: How should the designer choose her persuasion strategy when she is facing this kind of

uncertainty?

We consider a model where a two-member, possibly heterogeneous committee is called upon to choose between two alternatives. The evaluation of either alternative depends on an underlying, unobservable state of nature. Had there been no uncertainty about the state, both committee members would agree on the decision to make. However, under uncertainty they might disagree in that they require different amounts of evidence about the state in order to vote in favor of a given alternative. The information designer provides the committee with evidence about the state. Her preferences are not perfectly aligned with those of the committee because she wants to implement one of the two policies irrespective of the underlying state, whereas both committee members would like to implement one policy in one state and the alternative policy in the other. Furthermore, the designer does not know how much evidence a committee member needs in order to vote for any given option. More specifically, there are two possible types of committee members: those who require a relatively small amount of evidence to vote for the designer's preferred alternative (low types) and those who are harder to persuade (high types).

We characterize the designer's optimal persuasion strategy in a number of contexts. In order to do that, we build on Bergemann and Morris (2019) and adapt their unified framework on Information Design. First, we consider the case in which the designer is allowed to elicit private information. She asks committee members to confidentially report their preference types and then she discloses information about the state of nature based on these reports. This entails that two sets of constraints have to be taken into account, namely obedience and truth-telling. We analyze both the scenario where a unanimous consent is needed to implement the designer's preferred option and the alternative case where a single approval is sufficient. Secondly, we study the designer's problem when she is not allowed to elicit private information. That is, she has a prior distribution about preference types and discloses information just on the basis of it, without asking committee members to transmit any information at all. This implies that only obedience constraints have to be taken into account. Also in this case, both unanimity and single approval are considered.

We show that the optimal persuasion strategy crucially depends on the informativeness of the prior. We identify three possible cases. If the designer is sufficiently confident that no committee member is of the high type, then she tailors her strategy entirely to low types and gives up on persuading high type members. Intuitively, truth-telling constraints require that, in order to persuade high types, low type members have to vote for the designer's preferred option with lower probability. This implies that the expected gain from persuading the former is more than compensated by the loss arising from the latter. In the intermediate case, the designer is sufficiently confident that one committee member is of the high type, but

not both of them. By the same argument as above, she finds it optimal not to persuade the committee to adopt her preferred policy when both members declare to have a high preference parameter. In the remaining case, the designer is confident enough that the committee is composed of high type members. The optimal strategy is now to induce both members to vote for the designer's preferred policy irrespective of the private information they transmit to her. To guarantee that incentive constraints are satisfied, the exact probabilities with which either alternative is recommended are determined by the high preference type.

The above qualitative features of the optimal persuasion strategy are shown to hold irrespective of whether the designer elicits information and of the voting rule adopted. However, significant differences in the designer's gains from the optimal strategy arise. Not surprisingly, both information elicitation and non-unanimous voting rules are beneficial to the designer. Furthermore, uncertainty about committee members' preferences always entails a loss for the designer with respect to the benchmark case with commonly known preference parameters.

The paper proceeds as follows. In the remainder of this section we discuss the related literature. Then in section 4.2 we set up the model and analyze the benchmark case where preference parameters are commonly known. In Section 4.3 we study the case of privately known preferences when the designer is allowed to elicit information from committee members. First we introduce the appropriate solution concept and then we give a full characterization of the designer's optimal persuasion strategy. The case without information elicitation is studied in section 4.4. A discussion of results and modeling choices is contained in section 4.5.

## Related literature

The paper contributes to the literature on Information Design (or Bayesian Persuasion) with multiple receivers. The fundamental question in this literature is how to design an information structure so as to induce one or more agents, who base their decisions on that information, to take a certain action. The seminal contribution is Kamenica and Gentzkow (2011). They study the case with one Sender and one Receiver. Using the technique of concavification, they provide a simple and elegant characterization of the Sender's value function in terms of the Receiver's posterior beliefs about an underlying state of nature. As a consequence, the Sender's preferred optimal signal can easily be identified. Furthermore, they show that, for a large class of environments, Bayesian persuasion is strictly beneficial to the Sender.

The framework of Kamenica and Gentzkow (2011) has been extended along several dimen-

sions, and used in many applications. In the sequel, we confine our attention to contributions that are closest to this paper. Taneva (2019) studies the case with finitely many receivers. She establishes a revelation principle and provides a full characterization of the information structures that can be implemented in a class of binary games with strategic complements or substitutes, showing that public signals are typically optimal for persuading players in games with strategic complements while privately observed signals are optimal for games with strategic substitutes. In a slightly more general environment, Bergemann and Morris (2016a) introduce the solution concept of Bayes Correlated Equilibrium, which is a generalization of the classical Correlated equilibrium to Bayesian games where players are uncertain not only about their opponents' types but also about some underlying state of nature. They show that, for a given game, the set of outcome distributions that the Sender can induce by using any feasible information structure is the same as the set of Bayes Correlated Equilibria of that game. Since the latter set has a more tractable structure than the former, it turns out that the Sender's problem of choosing an information structure can be solved by working exclusively with Bayes Correlated Equilibria. Incidentally, the concavification technique cannot always be used when more than two receivers are considered.

It is important to point out that Bayesian Correlated Equilibrium is a suitable solution concept in Information Design provided that one of the following two assumptions are met: a) the receivers do not have any private information about the underlying state of nature, or b) receivers have private information about the state but the Sender can condition her action recommendations on the actual piece of information that each receiver observes. If neither assumption is satisfied, one has to use different solution concepts in order to characterize the set of implementable outcome distributions. Bergemann and Morris (2016b, 2019) provide these solutions concepts. More specifically, if the Sender is allowed to ask receivers to report the private information they have, then the set of implementable outcomes is identified by a state-contingent version of the Communication equilibrium of Myerson (1982). If the Sender does not have any access to the receivers' private information, then the set of implementable outcomes is identified by a state-contingent version of the Strategic Form Correlated Equilibrium of Forges (1993, 2006). The framework we use to carry out our analysis is that of Bergemann and Morris (2019).

Kolotilin et al. (2017) also consider the problem of designing information with privately informed receivers, and they focus on the case with only one receiver. They compare two kinds of mechanisms. With private mechanisms, the Sender elicits the receiver's private information through communication; with public mechanisms, the Sender designs experiments that are not contingent on reported information. They show that, for a large class of environments, the two types of mechanisms are outcome-equivalent.

Before turning our attention to applications, we point out that all of the contributions mentioned so far rely, more or less explicitly, on the revelation principle. That is, they all assume that the Sender is able to fully commit to the information structure she announces and to select the resulting equilibrium she prefers. A different approach has been proposed by Mathevet et al. (2019). Considering environments where the revelation principle does not necessarily hold, they tackle the Information Design problem by working directly with belief hierarchies over the underlying state. They show that any solution can be represented as the combination of a private and a public component. Interestingly, the latter is obtained by making use of the concavification technique that Kamenica and Gentzkow (2011) use in the case with only one receiver.

As for applications with privately informed receivers, Bergemann et al. (2018) analyze a model with a monopolist (data seller) and a consumer (data buyer). The latter has private information about a decision-relevant variable and can buy additional information from the monopolist. They identify the revenue-maximizing menu of experiments and study their properties. While they do allow for transfers, we rule them out in our analysis and assume that information is disclosed by the Sender at no cost.

The problem of persuading a committee (or group of voters) has been studied in a number of contributions within the Information Design literature. Wang (2013) studies a binary (i.e. with two states) voting game with $n$ players. She makes a comparison between persuasion with public signals and persuasion through private signals, showing that the former is always at least as good for the Sender as the latter. However, this result hinges upon the assumption that private signals are independent of each other and therefore uncorrelated. Relaxing this assumption, Chan et al. (2019) provide a full characterization of the optimal information structure in the same baseline game. They show that the use of correlated signals is strictly beneficial to the Sender when non-unanimous voting rules are adopted. Furthermore, they restrict the set of feasible information structures by requiring that only minimal winning coalitions can implement the Sender's preferred alternative. A related analysis is conducted in Alonso and Câmara (2016). They characterize the optimal persuasion mechanism and study the relationship between voters' welfare and the voting rule adopted. They find that, from the voters' perspective, simple majority rules might be sub-optimal. Their analysis allows for public signals only and does not consider private persuasion. In all the contributions mentioned so far, voters preferences are commonly known, and they do not possess any private information whatsoever. To the best of our knowledge, our paper is the first to study the problem of designing information for a group of receivers under the assumption that their preferences are not commonly known.

Outside of the Information Design literature, Caillaud and Tirole (2007) study how to

persuade a committee and cast their problem into the framework of costly information acquisition in mechanism design. They show that the Sender's optimal policy is to sequentially disclose information to selected subgroups of committee members.

This paper also touches upon the literature on strategic voting under incomplete information. Our basic game is essentially the same as the jury model studied in Austen-Smith and Banks (1996). Furthermore, the Bayes Communication Equilibrium we characterize in the section on persuasion with elicitation is a state-contingent version of the Communication equilibrium in deliberative voting of Gerardi and Yariv (2007). The fundamental difference with both models is that in our paper the information structure is the designer's decision variable whereas in the literature on rational voting the information structure is taken as given.

Finally, our analysis is also related to the literature on persuasion games, for example Glazer and Rubinstein (2004, 2006). However, the key difference is that in persuasion games the information provided to the (single) decision maker is a variable chosen by an informed persuader, whereas in our model the persuader/sender does not have any informational advantage over the voters/receivers. To wit, the sender chooses an information structure and commits to a disclosure policy when she is still uninformed about the true state of nature.

## 4.2 Model

### 4.2.1 Setup

We consider a model with one Sender (she) and a two-member committee of receivers who are called upon to choose between two alternatives. All three agents are expected-utility maximizers. The set of receivers is $I = \{1, 2\}$. The set of alternatives is $X = \{x_0, x_1\}$. We interpret $x_0$ as the status quo and $x_1$ as the non-status quo alternative. The collective decision is made through voting. Every receiver $i \in I$ selects an action from the set $A_i = \{0, 1\}$. We define $A := A_1 \times A_2$. Action 0 stands for "maintain the status quo" and action 1 stands for "implement the alternative $x_1$". Individual decisions are then aggregated via a $k$-voting rule, with $k \in \{1, 2\}$. More specifically, the alternative $x_1$ is implemented if and only if at least $k$ receivers vote for it. Otherwise, the status quo is maintained.

The receivers' evaluation of either alternative depends on an underlying state of nature and on a preference parameter. Two states of nature are possible, and they are identified by the set $\Theta = \{\theta_0, \theta_1\}$. Each receiver $i \in I$ has a preference parameter $t_i$ which belongs to the set $T_i = \{t_\ell, t_h\}$, where $\frac{1}{2} < t_\ell < t_h < 1$. We define $T := T_1 \times T_2$. The payoff function for

each receiver $i \in I$ is a function $u_i : A \times T \times \Theta \longrightarrow \mathbb{R}$ such that:

$$u_i\left((a_1, a_2), (t_1, t_2), \theta\right) = \begin{cases} -t_i & \text{if } a_1 + a_2 \geq k \text{ and } \theta = \theta_0 \\ -(1 - t_i) & \text{if } a_1 + a_2 < k \text{ and } \theta = \theta_1 \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

Both receivers want to implement the alternative $x_1$ if the true state is $\theta_1$, and maintain the status quo if the state is $\theta_0$. The payoff from implementing the correct policy is normalized to zero. If the wrong policy is adopted, then either receiver incurs a utility loss that is measured by the parameter $t_i$. Intuitively, $t_i$ captures how much a receiver is concerned about the implementation of the wrong policy in state $\theta_0$ relative to the implementation of the wrong policy in the other state.

The Sender has preferences over $A \times T \times \Theta$ that are represented by the following payoff function:

$$u_S\left((a_1, a_2), (t_1, t_2), \theta\right) = \begin{cases} 1 & \text{if } a_1 + a_2 \geq k \\ 0 & \text{otherwise.} \end{cases}$$

In words, the Sender wants the alternative $x_1$ to be implemented irrespective of the true state of nature. This entails a conflict of interests between the Sender and the receivers.

Receivers cannot observe the true state of nature. All agents share a common prior probability distribution $\psi$ over $\Theta$. Without loss of generality, we assume $\psi(\theta_0) = \psi(\theta_1) = 1/2$. Throughout the paper, we assume that the Sender's utility function $u_S$ is always commonly known. As for receivers, we consider both the benchmark case where their preference parameters are commonly known and the case where their preferences are privately observed. In the latter case, all agents share a common prior distribution $\pi$ over $T$. We assume types to be independently and identically distributed across receivers. In addition, we define $p := \mathsf{Prob}(t_1 = t_\ell) = \mathsf{Prob}(t_2 = t_\ell)$ and assume $p \in (0, 1)$.

The parameter $t_i$ can also be interpreted in terms of beliefs about the underlying state of nature. In the corresponding single-receiver decision problem, which is examined in Kamenica and Gentzkow (2011), a receiver $i$ would find it optimal to vote for the alternative $x_1$ if and only if the posterior probability he ascribes to $\theta_1$ is at least $t_i$. The assumption that $t_i > 1/2$ implies that, if $i$ were to make a decision solely on the basis of the prior information he has, he would vote for $x_0$.

Finally, notice that, when the committee is heterogeneous, i.e. $t_1 \neq t_2$, it could happen that receivers have diverging opinions about the alternative to be implemented even if they share a common belief about the prevailing state of nature. This might be the case even if, under complete information about $\theta$, they would both agree on what policy to choose.

### 4.2.2 Benchmark with commonly known preferences

We begin our analysis by studying the benchmark case where preference parameters $t_i$'s are commonly known. That is, fix a profile $(t_i, t_j) \in T$ and assume it is common knowledge. The fundamental object is a *persuasion mechanism*. Formally, let $S_i$ be a finite set of messages or signal realizations for receiver $i$. A persuasion mechanism $\sigma$ is a map $\sigma : \Theta \longrightarrow \Delta(S_1 \times S_2)$. With a slight abuse of notation, we denote $\sigma(s_1, s_2|\theta)$ as the probability that the mechanism $\sigma$ ascribes to receiver 1 receiving signal realization $s_1$ and receiver 2 receiving signal $s_2$, conditional on the true state being $\theta$.

The problem we are interested in is to choose a persuasion mechanism so as to maximize the Sender's objective function, i.e. maximize the probability of implementing the alternative $x_1$. The interaction between Sender and receivers unfolds as follows:

1. Nature selects the true state $\theta \in \Theta$;
2. The Sender chooses and publicly announces a persuasion mechanism $\sigma : \Theta \longrightarrow \Delta(S)$, where $S = S_1 \times S_2$;
3. The Sender observes the true state $\theta$ and sends signals to receivers according to the announced mechanism $\sigma$;
4. Each receiver privately observes his signal realization $s_i$ and then chooses an action from the set $A_i$;
5. Finally, the chosen social alternative is implemented and payoffs realize.

For any receiver $i \in I$, behavior at stage 4 is described by a strategy $\alpha_i : S_i \longrightarrow \Delta(A_i)$. We assume that receivers play strategies $(\alpha_1, \alpha_2)$ that constitute a Bayesian Nash Equilibrium of the extensive game induced by the chosen persuasion mechanism. Furthermore, we assume that the Sender is able to fully and credibly commit to the mechanism she announces, and to select the equilibrium that receivers are going to play. By standard arguments[1], it follows that it is without loss of generality to confine our attention to persuasion mechanisms that are both *direct* and *obedient*. More specifically, direct mechanisms are such that $S_i = A_i$ for every $i \in I$. Obedience requires that, for every $i \in I$, and $a_i \in A_i$, we have

$$\sum_{\theta \in \Theta, a_j \in A_j} \psi(\theta)\sigma(a_i, a_j|\theta)u_i((a_i, a_j), (t_i, t_j), \theta) \geq \sum_{\theta \in \Theta, a_j \in A_j} \psi(\theta)\sigma(a_i, a_j|\theta)u_i((a_i', a_j), (t_i, t_j), \theta) \tag{4.2}$$

[1]For a formal proof, see Taneva (2019) and Bergemann and Morris (2016a).

for all $a_i' \in A_i$. The obedience condition (4.2) identifies the solution concept of Bayes Correlated Equilibrium introduced by Bergemann and Morris (2016a), and it is a special case of their Definition 1. More specifically, in the framework of Bergemann and Morris (2016a), one can obtain (4.2) by assuming the following. First, each receiver $i$ has a utility function $\tilde{u}_i : A \times \Theta \longrightarrow \mathbb{R}$ such that, for all $a \in A$ and $\theta \in \Theta$, $\tilde{u}_i(a_1, a_2, \theta) = u_i((a_1, a_2), t, \theta)$. Second, all receivers have the *null* information structure, i.e. they have "no information over and above the common prior $\psi$" (Bergemann and Morris, 2016a, p. 492).

We now introduce a restriction on the set of feasible persuasion mechanisms. Following Chan et al. (2019), we require any mechanism to be *pivotal*.

**Assumption 8** (Pivotality)**.** *If $k = 1$, then $\sigma(1, 1|\theta) = 0$ for every $\theta \in \Theta$.*

The assumption says that, when a non-unanimous voting rule is adopted, the Sender cannot use action recommendation profiles where both receivers are told to vote for the alternative $x_1$. In other words, the Sender can only use minimal winning coalitions of receivers in order to implement her preferred policy $x_1$. The reason for making this assumption is to rule out those Bayesian Nash Equilibria in weakly dominated strategies of the underlying voting game where more than $k$ receivers vote for the alternative $x_1$. Without this assumption, it is easy to show that, when $k = 1$, the Sender can implement her preferred alternative with probability 1 in both states. Nonetheless, either receiver would find it optimal to obey the Sender's recommendations. Indeed, if receiver $i$ is obedient, he always votes 1, so determining which policy is going to be implemented. But then $j$ can never profit from disobeying because his deviation would never affect the voting outcome.

In order to formally state the Sender's optimization problem, we partition the set $A = A_1 \times A_2$ into two subsets $A^{x_1}$ and $A^{x_0}$, where $A^{x_1} = \{a \in A : a_1 + a_2 \geq k\}$ is the set of action profiles where the alternative $x_1$ is implemented and $A^{x_0} = \{a \in A : a_1 + a_2 < k\}$ is the set of action profiles where the status quo is maintained. Thus the Sender's problem is

$$\max_{\sigma:\Theta \longrightarrow \Delta(A)} \quad U_S(\sigma) = \frac{1}{2}\left[\sum_{a \in A^{x_1}} \sigma(a|\theta_0) + \sum_{a \in A^{x_1}} \sigma(a|\theta_1)\right] \tag{P-0}$$

such that $\sigma$ is pivotal and satisfies obedience constraints (4.2). Using payoff functions (4.1), the optimization problem when $k = 1$ can be written as follows:

$$\max_{\sigma:\Theta \longrightarrow \Delta(A)} \quad \frac{1}{2}\left[\sigma(1, 0|\theta_0) + \sigma(0, 1|\theta_0) + \sigma(1, 1|\theta_0) + \sigma(1, 0|\theta_1) + \sigma(0, 1|\theta_1) + \sigma(1, 1|\theta_1)\right]$$

such that

$$\sigma(1,1|\theta_0) = \sigma(1,1|\theta_1) = 0 \tag{4.3}$$
$$t_1\sigma(1,0|\theta_0) \leq (1-t_1)\sigma(1,0|\theta_1) \tag{4.4}$$
$$t_2\sigma(0,1|\theta_0) \leq (1-t_2)\sigma(0,1|\theta_1) \tag{4.5}$$
$$(1-t_1)\sigma(0,0|\theta_1) \leq t_1\sigma(0,0|\theta_0) \tag{4.6}$$
$$(1-t_2)\sigma(0,0|\theta_1) \leq t_2\sigma(0,0|\theta_0), \tag{4.7}$$

where (4.3) is the pivotality restriction, (4.4) and (4.5) are the obedience constraints for action recommendation $a_i = 1$ for player 1 and player 2, and (4.6) and (4.7) are the obedience constraints for action recommendation $a_i = 0$. Notice that receiver $i$'s decision whether to obey or not the Sender's action recommendation depends only on action recommendation profiles where $i$ himself is pivotal.

When $k = 2$, we have:

$$\max_{\sigma:\Theta\longrightarrow\Delta(A)} \quad \frac{1}{2}\left[\sigma(1,1|\theta_0) + \sigma(1,1|\theta_1)\right]$$

such that

$$t_1\sigma(1,1|\theta_0) \leq (1-t_1)\sigma(1,1|\theta_1) \tag{4.8}$$
$$t_2\sigma(1,1|\theta_0) \leq (1-t_2)\sigma(1,1|\theta_1) \tag{4.9}$$
$$(1-t_1)\sigma(0,1|\theta_1) \leq t_1\sigma(0,1|\theta_0) \tag{4.10}$$
$$(1-t_2)\sigma(1,0|\theta_1) \leq t_2\sigma(1,0|\theta_0), \tag{4.11}$$

where (4.8) and (4.9) are the obedience constraints for action recommendation $a_i = 1$ for player 1 and player 2, and (4.10) and (4.11) are the obedience constraints for action recommendation $a_i = 0$.

We can now fully characterize the optimal persuasion mechanism when receivers' preferences are commonly known. The general case with $n$ players and complete information about preferences is studied in Chan et al. (2019).

**Proposition 8.** *(1) Suppose $k = 1$. Then the mechanism $\sigma$ that solves (P-0) is such that:*

- *if $t_1 = t_2$, then*
  - $\sigma(1,0|\theta_1) + \sigma(0,1|\theta_1) = 1$,
  - $\sigma(1,0|\theta_0) = \frac{1-t_i}{t_i}\,\sigma(1,0|\theta_1)$,

– $\sigma(0,1|\theta_0) = \frac{1-t_i}{t_i}\,\sigma(0,1|\theta_1)$;

- *if* $t_1 < t_2$, *then* $\sigma(1,0|\theta_1) = 1$ *and* $\sigma(1,0|\theta_0) = \frac{1-t_1}{t_1}$;
- *if* $t_1 > t_2$, *then* $\sigma(0,1|\theta_1) = 1$ *and* $\sigma(0,1|\theta_0) = \frac{1-t_2}{t_2}$.

*(2) Suppose* $k = 2$. *Then the mechanism* $\sigma$ *that solves* (P-0) *is such that* $\sigma(1,1|\theta_1) = 1$. *Furthermore,*

- *if* $t_1 = t_2$, *then* $\sigma(1,1|\theta_0) = \frac{1-t_i}{t_i}$;
- *if* $t_1 < t_2$, *then* $\sigma(1,1|\theta_0) = \frac{1-t_2}{t_2}$;
- *if* $t_1 > t_2$, *then* $\sigma(1,1|\theta_0) = \frac{1-t_1}{t_1}$.

*Proof.* (1) Let $k = 1$. First, we argue that $\sigma(0,0|\theta_1) = 0$. By pivotality, this is equivalent to $\sigma(1,0|\theta_1) + \sigma(0,1|\theta_1) = 1$. Suppose $\sigma(0,0|\theta_1) > 0$. Then it would always be feasible to shift probability away from $\sigma(0,0|\theta_1)$ to $\sigma(1,0|\theta_1)$ or $\sigma(0,1|\theta_1)$ without tightening any obedience constraint. This shift would entail a strict increase in the objective function, so leading to a contradiction. Second, we argue that both (4.4) and (4.5) are binding. By way of contradiction, suppose this is not the case. Since $t_i > (1 - t_i)$ by assumption, by (4.4) and (4.5) it follows that $\sigma(1,0|\theta_0) + \sigma(0,1|\theta_0) < 1$. Furthermore, we already proved that $\sigma(0,0|\theta_1) = 0$. But then it is feasible to shift probability away from $\sigma(0,0|\theta_0)$ to either $\sigma(1,0|\theta_1)$ or $\sigma(0,1|\theta_1)$ so as to make constraint (4.4) or (4.5), respectively, binding. This would entail a strict increase in the objective function, so leading to a contradiction.

Therefore, from the binding constraints (4.4) and (4.5) we obtain

$$\sigma(1,0|\theta_0) = \frac{1-t_1}{t_1}\sigma(1,0|\theta_1)$$
$$\sigma(0,1|\theta_0) = \frac{1-t_2}{t_2}\sigma(0,1|\theta_1),$$

from which the result easily follows.

(2) Let $k = 2$. By the same argument as above, $\sigma(1,1|\theta_1) = 1$. The result follows from the fact that, if $t_1 = t_2$, then both (4.8) and (4.9) are binding. And that if $t_1 \neq t_2$, then the obedience constraint for $a_i = 1$ must be binding for the receiver with the largest $t_i$ only.

□

The interpretation of Proposition 8 is straightforward. First, the alternative $x_1$ is implemented with probability 1 if the true state is $\theta_1$ and with probability strictly less than 1 in

the other state, irrespective of the voting rule adopted or preference heterogeneity. We will see that this is no longer the case when preferences are privately observed. Now suppose that the voting rule is unanimity, so that the Sender can use only one winning coalition to implement $x_1$. The probability of recommending to vote for $x_1$ conditional on state $\theta_0$ is determined by the receiver with the largest parameter $t_i$. This can also be seen from the fact that the obedience constraint for action recommendation $a_i = 1$ is always binding for the receiver with the highest $t_i$ while it is slack for the receiver with the lowest $t_i$ if the committee is heterogeneous. The probability $\frac{1-t_i}{t_i}$ of recommending action 1 conditional on $\theta_0$ is such that the resulting posterior probability ascribed to state $\theta_1$ is exactly equal to $t_i$. Clearly, this implies that, if a receiver $i$ with $t_i = t_h$ obeys the Sender's recommendation to vote 1, so does the other receiver with a lower preference parameter.

Consider the case with $k = 1$. Now the Sender can use two different winning coalitions to implement $x_1$. Since the receiver with the lowest $t_i$ is easier to persuade, it is clear that the Sender wants to target him rather than the other voter. If the committee is heterogeneous, the receiver $i$ with $t_i = t_\ell$ is told to vote 1 with probability 1 in state $\theta_1$ and with probability $\frac{1-t_i}{t_i}$ in the other state, so that his posterior about $\theta_1$ is exactly equal to his preference parameter. The other voter is told instead to vote for 0 irrespective of the true state.

We emphasize that the optimal mechanism in Proposition 8 inherits some features of the optimal information structure for the one-receiver case in Kamenica and Gentzkow (2011). Even with two receivers, it is optimal for the Sender to partially obfuscate the true realization of $\theta$. More specifically, the Sender's optimal persuasion mechanism can be interpreted as the conflation of two signals: one that, conditional on the state being $\theta_1$, fully discloses the state to either one or both receivers; and another signal that, contingent on $\theta_0$, reveals imperfectly the true state. Such a property is clearly driven by the fact that Sender and receivers agree on the policy to adopt in state $\theta_1$ while they disagree in the other state.

The Sender's expected payoff from the optimal mechanism is clearly affected by both the voting rule and preference parameters. If the committee is homogeneous, then the Sender is indifferent between the two voting rules. But if the committee is heterogeneous, then the Sender has a strict preference for the non-unanimous rule. To see this, notice that, conditional on the true state being $\theta_0$, the probability of implementing $x_1$ is $\mathsf{P}(x_1|\theta_0) = \frac{1-t_\ell}{t_\ell}$ when $k = 1$ and $\mathsf{P}(x_1|\theta_0) = \frac{1-t_h}{t_h}$ when $k = 2$. The former probability is strictly greater than the latter due to assumption that $t_\ell < t_h$.

We conclude this section by pointing out that the use of private recommendations is not necessary when the composition of the committee is common knowledge. It is easy to see that the equilibrium outcomes induced by Proposition 8 can also be obtained through signals whose realizations are publicly observed. However, this is not necessarily true when $n > 2$

and non-unanimous voting rules are adopted. In addition, public signals are not without loss of generality when uncertainty about receivers' preferences is introduced.

## 4.3 Persuasion with information elicitation

In this section we give a characterization of the optimal persuasion mechanism under two main assumptions: 1) each receiver $i$ privately observes his preference parameter $t_i$; and 2) the Sender is allowed to elicit receivers' private information by asking them to report their types.

### 4.3.1 The solution concept

The possibility of eliciting information implies that a persuasion mechanism is a menu of experiments contingent on reports. More specifically, let $R_i$ be a finite set of reports available to receiver $i = 1, 2$. As in the benchmark case with complete information, $S_i$ denotes the set of signal realizations for receiver $i$. Let $R = R_1 \times R_2$ and $S = S_1 \times S_2$. Furthermore, let $\mu$ be a function $\mu : R \times \Theta \longrightarrow \Delta(S)$. We call $((R_i, S_i)_{i \in I}, \mu)$ a persuasion mechanism. Abusing notation, we denote $\mu(s_1, s_2 | r_1, r_2, \theta)$ as the probability that $\mu$ ascribes to receiver 1 receiving signal realization $s_1$ and receiver 2 receiving signal $s_2$, conditional on the true state being $\theta$ and on receivers 1 and 2 submitting reports $r_1$ and $r_2$, respectively.

Any mechanism $\mu$ induces an extensive game with imperfect information. More specifically, the sequence of events in the game induced by $\mu$ is the following.

1. Nature selects the true state $\theta \in \Theta$ and preference parameters $(t_1, t_2) \in T_1 \times T_2$;
2. Each receiver $i$ privately observes his own preference type $t_i$;
3. The Sender chooses and publicly announces a persuasion mechanism $((R_i, S_i)_{i \in I}, \mu)$;
4. Each receiver $i$ sends a report $r_i$ to the Sender. The Sender observes the true state $\theta$, collects reports $r = (r_1, r_2)$, and then sends signals to receivers according to the announced $\mu$;
5. Each receiver privately observes his signal realization $s_i$ and then chooses an action from the set $A_i$;
6. Finally, the chosen social alternative is implemented and payoffs realize.

Receivers' behavior in the game above is captured by a pair of functions. More specifically, either receiver transmits reports to the Sender according to a function $\rho_i : T_i \longrightarrow R_i$. In addition, he makes his final decision according to the function $\delta_i : T_i \times S_i \longrightarrow A_i$. We assume that receivers play strategy profiles $(\rho, \delta) = ((\rho_1, \delta_1), (\rho_2, \delta_2))$ that constitute a Bayes Nash Equilibrium (BNE) of the game induced by the announced mechanism. A strategy profile $(\rho, \delta)$ is a BNE of such a game if, for every $i \in I$, and every $t_i \in T_i$, we have

$$\begin{aligned}\sum_{(s_i,s_j)\in S, t_j\in T_j, \theta\in\Theta} \psi(\theta)\pi(t_i,t_j)\mu\left(s_i,s_j|\rho_i(t_i),\rho_j(t_j),\theta\right)u_i((\delta_i(t_i,s_i),\delta_j(t_j,s_j)),(t_i,t_j),\theta) \geq \\ \sum_{(s_i,s_j)\in S, t_j\in T_j, \theta\in\Theta} \psi(\theta)\pi(t_i,t_j)\mu\left(s_i,s_j|\hat{\rho}_i,\rho_j(t_j),\theta\right)u_i((\hat{\delta}_i(s_i),\delta_j(t_j,s_j)),(t_i,t_j),\theta),\end{aligned} \tag{4.12}$$

for all $\hat{\rho}_i \in R_i$ and all functions $\hat{\delta}_i : S_i \longrightarrow A_i$.

We assume that the Sender can credibly and fully commit to the persuasion mechanism she announces. In addition, she is able to select any BNE of the game induced by the mechanism she has committed to. As a consequence, we can invoke the *revelation principle* and establish the following.

**Proposition 9.** *Fix any mechanism $((R_i, S_i)_{i\in I}, \mu)$. For any BNE induced by such a mechanism, there exists an outcome-equivalent persuasion mechanism $((T_i, A_i)_{i\in I}, \sigma)$, with $\sigma : T \times \Theta \longrightarrow \Delta(A)$, such that, for every $i \in I$, and every $t_i \in T_i$, we have*

$$\begin{aligned}\sum_{(a_i,a_j)\in A, t_j\in T_j, \theta\in\Theta} \psi(\theta)\pi(t_i,t_j)\sigma(a_i,a_j|t_i,t_j,\theta)u_i((a_i,a_j),(t_i,t_j),\theta) \geq \\ \sum_{(a_i,a_j)\in A, t_j\in T_j, \theta\in\Theta} \psi(\theta)\pi(t_i,t_j)\sigma(a_i,a_j|t'_i,t_j,\theta)u_i((\delta_i(a_i),a_j),(t_i,t_j),\theta)\end{aligned} \tag{4.13}$$

*for all $t'_i \in T_i$ and all functions $\delta_i : A_i \longrightarrow A_i$.*

When $t_i = t'_i$ in (4.13), we call the corresponding inequalities *obedience* constraints, whereas when $t_i \neq t'_i$ we indicate them as *truth-telling* constraints.

*Proof.* The argument can be easily adapted from the proof of Proposition 2 in Myerson (1982). Let $((R_i, S_i)_{i\in I}, \mu)$ be a persuasion mechanism, and let $(\rho, \delta)$ be the BNE induced by it. For any $(a, t) \in A \times T$, define

$$(a,t)^{-1} := \{s \in S : \delta_i(t_i, s_i) = a_i \text{ for all } i \in I\}.$$

The corresponding direct mechanism is $((T_i, A_i)_{i\in I}, \sigma)$, where $\sigma$ is such that, for every $a \in A$,

$t \in T$, and $\theta \in \Theta$,

$$\sigma(a|t,\theta) = \sum_{s \in (a,t)^{-1}} \mu\left(s|((\rho_1(t_1), \rho_2(t_2)), \theta\right).$$

It is then straightforward to verify that $\sigma$ is incentive compatible and that it induces the same outcome distribution as $\mu$. □

Proposition 9 can easily be extended to any finite game. Its content is standard. It asserts that, in order to identify the set of implementable outcomes, we can confine our attention to persuasion mechanisms that are *direct* and *incentive compatible*. Direct mechanisms are such that, for every receiver, the set of reports is equal to his type set; and the set of signal realizations coincides with the set of action profiles. Incentive compatibility is defined by the inequality constraints (4.13). A mechanism is incentive compatible if any receiver finds it optimal to truthfully reveal his type *and* to obey the action recommendation he gets from the Sender.

The solution concept identified by (4.13) is nothing other than a state-contingent version of the Communication Equilibrium (CE) of Myerson (1982). To emphasize the fact that the Sender can condition her action recommendations on $\theta$, we call any mechanism $\sigma$ that satisfies (4.13) a **Bayes Communication Equilibrium** (BCE). It is immediate to see that, for any given game, a CE is also a BCE, whereas the converse is not necessarily true. The two solution concepts are necessarily equivalent either when $|\Theta| = 1$ or when $\sigma$ is required to be independent of $\theta$.

It is important to discuss the connection between the solution concept we are adopting here and those proposed by Bergemann and Morris (2019) for the case of persuasion with information elicitation. The incentive compatibility condition (4.13) is a particular case of the incentive compatibility notion in (Bergemann and Morris, 2019, Definition 2). More specifically, one can obtain (4.13) in their framework as follows. First, the basic game has a state space $\widetilde{\Theta} = T \times \Theta$, and the corresponding common prior is defined as $\widetilde{\psi}(t,\theta) = \pi(t) \times \psi(\theta)$, for every $t \in T$ and $\theta \in \Theta$. The payoff function $\tilde{u}_i : A \times \widetilde{\Theta}$ is the obvious rewriting of $u_i$. Second, the information structure is $\left((\widetilde{T}_i)_{i \in I}, \widetilde{\pi}\right)$, where $\widetilde{T}_1 = \widetilde{T}_2 = \{t_\ell, t_h\}$ and $\widetilde{\pi} : \widetilde{\Theta} \longrightarrow \Delta(\widetilde{T})$ is such that, for every $\tilde{t} \in \widetilde{T}$,

$$\widetilde{\pi}(\tilde{t}|(\tilde{t},\theta_0)) = \widetilde{\pi}(\tilde{t}|(\tilde{t},\theta_1)) = 1.$$

This captures the assumption that receivers privately observe their true preference types. Finally, the Sender is constrained to use decision rules that are *partially* join feasible. As per (Bergemann and Morris, 2019, Definition 6), a decision rule is join feasible if it is independent of the true state $\theta$. We require that $\widetilde{\sigma} : \widetilde{T} \times \widetilde{\Theta} \longrightarrow \Delta(A)$ be partially so. That is, for every

$a \in A$, $\theta \in \Theta$, $\tilde{t} \in \widetilde{T}$, and every $t, t' \in T$, we have

$$\widetilde{\sigma}(a|\tilde{t}, (t, \theta)) = \widetilde{\sigma}(a|\tilde{t}, (t', \theta)).$$

It is then straightforward to verify that, with the three caveats that we have just mentioned, the notion of incentive compatibility in (Bergemann and Morris, 2019, Definition 2) reduces to (4.13).

Before stating the Sender's problem, we introduce two restrictions on the set of feasible persuasion mechanisms.

**Assumption 9** (Pivotality)**.** *If $k = 1$, then $\sigma(1, 1|t, \theta) = 0$ for any $t \in T$, and $\theta \in \Theta$.*

Assumption 9 is the incomplete information version of Assumption 8.

**Assumption 10** (Symmetry)**.** *For any $\theta \in \Theta$, $\sigma(a_1, a_2|t_1, t_2, \theta) = \sigma(a_2, a_1|t_2, t_1, \theta)$ for any $a \in A$ and $t \in T$.*

Assumption 10 states that we consider persuasion mechanisms that are symmetric across receivers. Since the Sender's objective function is symmetric across receivers, the assumption is without loss of generality.

Finally, given a prior distribution $\pi$ over preference types, the Sender's choice problem is:

$$\max_{\sigma} \quad U_S(\sigma) = \frac{1}{2}\left[\sum_{t \in T} \pi(t) \left( \sum_{a \in A^{x_1}} \sigma(a|t, \theta_0) + \sum_{a \in A^{x_1}} \sigma(a|t, \theta_1) \right)\right] \tag{P-1}$$

such that $\sigma$ is:

- a BCE of the underlying voting game;
- symmetric;
- pivotal.

We give a full characterization of the solution in the subsequent section.

### 4.3.2 Unanimity

We now have all the ingredients to solve the Sender's choice problem (P-1) in the case in which unanimity is required to implement the social alternative $x_1$.

We use the following abbreviations to save on notation. First, we denote the action profile $(1, 1)$ simply as $(1)$. In addition, we express any type profile $(t_1, t_2) \in T$ simply as $t_{12}$. For

example, we use $\sigma(1|t_{\ell h},\theta)$ to denote $\sigma(1,1|t_\ell,t_h,\theta)$. Finally, we exploit symmetry and use $\sigma(1|t_{\ell h},\theta)$ or $\sigma(1|t_{h\ell},\theta)$ interchangeably for any $\theta\in\Theta$.

Given a prior $p$, the Sender's expected utility from a mechanism $\sigma$ when $k=2$ is

$$U_S(\sigma)=\frac{1}{2}\sum_{\theta\in\Theta}\left[p^2\sigma(1|t_{\ell\ell},\theta)+2p(1-p)\sigma(1|t_{\ell h},\theta)+(1-p)^2\sigma(1|t_{hh},\theta)\right]. \tag{4.14}$$

It is clear that, for every type profile and in each state, the Sender earns a positive expected payoff only from the action profile where both receivers vote for the alternative $x_1$.

The set of incentive constraints defining a BCE of the underlying voting game includes 7 inequalities for each type of either receiver. However, using symmetry and the fact that the alternative $x_1$ can be implemented with only one action profile for each type profile, one can show[2] that the set of relevant constraints reduces to the following:

$$t_h\left[p\sigma(1|t_{\ell h},\theta_0)+(1-p)\sigma(1|t_{hh},\theta_0)\right]\leq(1-t_h)\left[p\sigma(1|t_{\ell h},\theta_1)+(1-p)\sigma(1|t_{hh},\theta_1)\right] \tag{4.15}$$

$$\begin{aligned}&t_\ell\left[p\sigma(1|t_{\ell\ell},\theta_0)+(1-p)\sigma(1|t_{\ell h},\theta_0)\right]+(1-t_\ell)\left[p\sigma(1|t_{\ell h},\theta_1)+(1-p)\sigma(1|t_{hh},\theta_1)\right]\leq\\&\quad t_\ell\left[p\sigma(1|t_{\ell h},\theta_0)+(1-p)\sigma(1|t_{hh},\theta_0)\right]+(1-t_\ell)\left[p\sigma(1|t_{\ell\ell},\theta_1)+(1-p)\sigma(1|t_{\ell h},\theta_1)\right]\end{aligned} \tag{4.16}$$

$$\begin{aligned}&t_h\left[p\sigma(1|t_{\ell h},\theta_0)+(1-p)\sigma(1|t_{hh},\theta_0)\right]+(1-t_h)\left[p\sigma(1|t_{\ell\ell},\theta_1)+(1-p)\sigma(1|t_{\ell h},\theta_1)\right]\leq\\&\quad t_h\left[p\sigma(1|t_{\ell\ell},\theta_0)+(1-p)\sigma(1|t_{\ell h},\theta_0)\right]+(1-t_h)\left[p\sigma(1|t_{\ell h},\theta_1)+(1-p)\sigma(1|t_{hh},\theta_1)\right],\end{aligned} \tag{4.17}$$

where (4.15) the obedience constraint against deviations to $\delta_i(a_i)=0$ for any $a_i\in A_i$ for type $t_h$, while (4.16) and (4.17) are truth-telling constraints for type $t_\ell$ and $t_h$, respectively. For either type, all incentive constraints crucially depend on the probability of being pivotal and receiving an action recommendation 1 conditional on being of that type. More specifically, constraint (4.15) is driven by the quantity $p\sigma(1|t_{\ell h},\theta)+(1-p)\sigma(1|t_{hh},\theta)$, which is the probability of receiving recommendation 1 and being pivotal conditional on being a high-type receiver and on the state $\theta$. Intuitively, this probability in state $\theta_0$ must be sufficiently low in order for the high type to obey the Sender's recommendation. As for truth-telling, constraints (4.16) and (4.17) simply say that each type compares the expected loss from being sincere with the loss from lying. Notice that obedience and truth-telling are sufficient

[2]See Appendix A.1.

to ensure that double deviations, i.e. misreporting and disobeying at the same time, are not profitable. This stems from the fact that the Sender has only one action profile at her disposal to implement the alternative she prefers. We will see later on that, in the case of single approval, obedience and truth-telling alone are not sufficient to prevent double deviations.

We now proceed to characterize the optimal persuasion mechanism(s) as a function of the prior probability $p$. In order to do that, we first illustrate three subsets of feasible mechanisms and then describe the solution set of problem (P-1) in terms of these subsets. Formally, let $\Sigma$ be the set of symmetric BCEs of the fundamental voting game. Then we need to consider three subsets of $\Sigma$, namely $\Sigma^A$, $\Sigma^B$, and $\Sigma^C$.

**Mechanisms in $\Sigma^A$:** The first subset contains any mechanism $\sigma^A$ that satisfies the following properties:

- $\sigma^A(1|t_{\ell\ell},\theta_1)=1$,
- $\sigma^A(1|t_{\ell\ell},\theta_0)=\frac{1-t_\ell}{t_\ell}$,
- $\sigma^A(1|t,\theta)=0$ for any $t\in T\backslash\{t_{\ell\ell}\}$ and any $\theta\in\Theta$.

In words, the action profile where receivers vote 1 is recommended with positive probability only when both of them are of the low type.

**Mechanisms in $\Sigma^B$:** Any mechanism $\sigma^B$ in the second subset has the following properties:

- $\sigma^B(1|t_{\ell\ell},\theta_1)=\sigma^B(1|t_{\ell h},\theta_1)=1$,
- $\sigma^B(1|t_{hh},\theta_1)=0$,
- $\sigma^B(1|t_{\ell h},\theta_0)\in(0,1)\cap\left(\frac{1-t_h}{t_h}-\frac{1-p}{p},\frac{1-t_h}{t_h}\right)\cap\left(\frac{1-t_\ell}{t_\ell}+\frac{(1-t_h)p}{t_h(1-p)}-\frac{p}{1-p},\frac{1-t_\ell}{t_\ell}+\frac{(1-t_h)p}{t_h(1-p)}\right)$,
- $\sigma^B(1|t_{hh},\theta_0)=\frac{p}{1-p}\left[\frac{1-t_h}{t_h}-\sigma^B(1|t_{\ell h},\theta_0)\right]$,
- $\sigma^B(1|t_{\ell\ell},\theta_0)=\frac{1-t_h}{t_h}+\frac{1-p}{p}\left[\frac{1-t_\ell}{t_\ell}-\sigma^B(1|t_{\ell h},\theta_0)\right]$.

Conditional on the state being $\theta_1$, the action profile that implements $x_1$ is recommended with probability 1 if and only if at least one of the receivers is of the low type. In the other state, the probability of implementing $x_1$ is chosen so as to guarantee that the truth-telling constraint for $t_\ell$ and the obedience constraint for $t_h$ are both binding.

**Mechanisms in $\Sigma^C$:** Any mechanism $\sigma^C$ in the third subset has the following properties:

- $\sigma^C(1|t,\theta_1)=1$ for any $t\in T$,
- $\sigma^C(1|t_{\ell\ell},\theta_0)\in(0,1)\cap\left(\frac{1-t_h}{pt_h}-\frac{1-p}{p},\frac{1-t_h}{pt_h}\right)\cap\left(-\frac{(1-t_h)(1-2p)}{p^2t_h},\frac{(1-p)^2}{p^2}-\frac{(1-t_h)(1-2p)}{p^2t_h}\right)$,
- $\sigma^C(1|t_{\ell h},\theta_0)=\frac{1}{1-p}\left[\frac{1-t_h}{t_h}-p\sigma^C(1|t_{\ell\ell},\theta_0)\right]$,
- $\sigma^C(1|t_{hh},\theta_0)=\frac{1}{(1-p)^2}\left[\frac{(1-t_h)(1-2p)}{t_h}+p^2\sigma^C(1|t_{\ell\ell},\theta_0)\right]$.

In words, the alternative $x_1$ is implemented with probability 1 in state $\theta_1$, irrespective of the underlying type profile. Conditional on state $\theta_0$, the probability of recommending 1 is chosen so as to make sure that all incentive constraints (4.15)-(4.17) are binding.

We are now ready to state the following.

**Proposition 10.** *Let $k=2$. The following is true:*

1. *if $p\geq\frac{t_\ell+t_h}{2t_h}$, then any $\sigma\in\Sigma^A$ solves (P-1);*
2. *if $p\in\left[\frac{t_\ell}{t_h},\frac{t_\ell+t_h}{2t_h}\right]$, then any $\sigma\in\Sigma^B$ solves (P-1);*
3. *if $p\leq\frac{t_\ell}{t_h}$, then any $\sigma\in\Sigma^C$ solves (P-1);*
4. *If $p=\frac{t_\ell+t_h}{2t_h}$, then any $\sigma\in\mathsf{conv}\left(\Sigma^A\cup\Sigma^B\right)$ solves (P-1);*
5. *If $p=\frac{t_\ell}{t_h}$, then any $\sigma\in\mathsf{conv}\left(\Sigma^B\cup\Sigma^C\right)$ solves (P-1),*

*where $\mathsf{conv}(Y)$ denotes the convex hull of set $Y$.*

*Proof.* See Appendix A.2. □

The result can be interpreted as follows. Suppose $p\geq\frac{t_\ell+t_h}{2t_h}$, so that the Sender is confident enough that the true type profile is $t_{\ell\ell}$. Conditional on the profile of reports $t_{\ell\ell}$, the Sender finds it optimal to recommend actions according to the optimal mechanism of Proposition 8 under complete information. That is, she recommends to vote 1 with probability 1 in state $\theta_1$ and with probability $\frac{1-t_\ell}{t_\ell}$ in the other state, so that the resulting posterior about $\theta_1$ is equal to $t_\ell$ for either receiver. For any other profile of reports, the best persuasion strategy is to maintain the status quo $x_0$ in either state with probability 1. To see why this is the case, suppose the Sender implements the alternative $x_1$ even when at least a high type is reported. In order to do that, she must increase the probability of the high type voting for 1. At the same time, she must decrease the probability of the low type voting for 1 to make sure that

the low type does not find it profitable to misreport his type. Since the prior $p$ is sufficiently high, the expected gain from persuading committees with at least one receiver of the high type is more than compensated by the loss from reducing the probability of implementing $x_1$ contingent on the type profile $t_{\ell\ell}$. Intuitively, the Sender tailors her persuasion strategy entirely on a committee of low types, and she gives up on using possible committees with at least one receiver of the high type to implement her preferred strategy.

Now consider the intermediate case where $p \in \left[\frac{t_\ell}{t_h}, \frac{t_\ell+t_h}{2t_h}\right]$. Intuitively, the prior is not high enough to conclude that the committee is likely to be made of low types only as in the previous case. However, the prior $p$ is still sufficiently large to conclude that the committee is very unlikely to include two receivers of the high type. Consequently, the best persuasion strategy for the Sender is to implement $x_1$ in $\theta_1$ with probability 1 if at least one receiver reports a preference parameter $t_\ell$, and with probability zero otherwise. The probability of implementing $x_1$ in the other state is chosen so as to guarantee that both the obedience constraint for the high type and the truth-telling constraint for the low type are binding. The decision of not implementing $x_1$ in state $\theta_1$ contingent on reports $t_{hh}$ is driven by the same logic illustrated in the previous case with $p \geq \frac{t_\ell+t_h}{2t_h}$. As a consequence, notice that truth-telling for the low type implies $\sigma^B(1|t_{\ell\ell}, \theta_0) < \sigma^A(1|t_{\ell\ell}, \theta_0) = \frac{1-t_\ell}{t_\ell}$.

If $p \leq \frac{t_\ell}{t_h}$, the Sender believes that both receivers are most likely to have a preference parameter $t_h$. The resulting optimal persuasion strategy is such that $x_1$ is implemented with probability 1 in state $\theta_1$, irrespective of reported types. Conditional on state $\theta_0$, the Sender implements $x_1$ with strictly positive probability for any profile of reports. More specifically, the ex-ante probability of implementing $x_1$ in $\theta_0$ is equal to $\frac{1-t_h}{t_h}$. Intuitively, the Sender tailors her persuasion strategy entirely on a committee of high types, which is clearly the harder–or costlier–to persuade. Again, the cost lies in the fact that the probability of recommending 1 contingent on state $\theta_0$ must be sufficiently low in order to guarantee that high types obey that recommendation and low types do not want to misreport their types.

Finally, $p = \frac{t_\ell+t_h}{2t_h}$ and $p = \frac{t_\ell}{t_h}$ are boundary cases. More specifically, when $p = \frac{t_\ell+t_h}{2t_h}$, the Sender gets the same expected utility from persuasion mechanisms in $\Sigma^A$ and $\Sigma^B$. Since the set of constraints defining a BCE is convex, and since the Sender's objective function is linear, it follows that any convex combination of mechanisms in $\Sigma^A$ and $\Sigma^B$ is a solution to the Sender's problem (P-1). An analogous reasoning applies when $p = \frac{t_\ell}{t_h}$.

By substituting in the Sender's utility function (4.14), we can easily find the value function of problem (P-1) for each optimal mechanism. We have:

$$U_S(\sigma^A) = \frac{p^2}{2t_\ell},$$

$$U_S(\sigma^B) = \frac{1}{2}\left[\frac{p(1-t_h)}{t_h} + \frac{p(1-p)(1-t_\ell)}{t_\ell} + p(2-p)\right],$$
$$U_S(\sigma^C) = \frac{1}{2t_h}.$$

It is clear that $U_S(\sigma^A)$ is decreasing with respect to $t_\ell$, $U_S(\sigma^C)$ is decreasing with respect to $t_h$, and $U_S(\sigma^C)$ is decreasing with respect to both $t_\ell$ and $t_h$. Intuitively, the higher preference parameters are, the more difficult it gets to persuade receivers to vote in favor of the alternative $x_1$. As far as the prior $p$ is concerned, $U_S(\sigma^A)$ increases with respect to it at an increasing rate, while $U_S(\sigma^C)$ is independent of it. Finally, $U_S(\sigma^B)$ is increasing with respect to $p$ as long as $p \leq \frac{t_\ell + t_h}{2t_h}$, and it does so at a decreasing rate. The interpretation is that, as $p$ gets larger, the committee is more likely to be composed of receivers of the low type. This means that it is easier to convince them to vote for $x_1$.

### 4.3.3 Single approval

We now proceed to characterize the solution set of problem (P-1) when a single approval is sufficient to implement the alternative $x_1$.

In addition to the abbreviations introduced in the previous subsection, we express the action profiles $(1,0)$ and $(0,1)$ simply as $(10)$ and $(01)$, respectively. For instance, $\sigma(10|t_{\ell h}, \theta)$ stands for $\sigma(1,0|t_\ell, t_h, \theta)$. Notice that, by symmetry, we have $\sigma(10|t_{\ell h}, \theta) = \sigma(01|t_{h\ell}, \theta)$ for any $\theta \in \Theta$.

Given a prior $p$, the Sender's expected utility from a persuasion mechanism $\sigma$ is

$$U_S(\sigma) = \sum_{\theta\in\Theta} \left[p^2\sigma(10|t_{\ell\ell}, \theta) + p(1-p)\left(\sigma(10|t_{\ell h}, \theta) + \sigma(01|t_{\ell h}, \theta)\right) + (1-p)^2\sigma(10|t_{hh}, \theta)\right]. \quad (4.18)$$

Notice that, for every type profile and in each state, the Sender can implement the alternative $x_1$ by using two action recommendation profiles, namely $(1,0)$ and $(0,1)$. Due to the assumption of pivotality, the action profile $(1,1)$ is never recommended.

To guarantee that $\sigma$ is a BCE, we have to check 7 incentive constraints for each type of each receiver. Like in the case with $k = 2$, symmetry and pivotality allow us to reduce the set of constraints. In particular, it is sufficient[3] to satisfy the following 5 inequalities to identify a BCE for the voting game with $k = 1$:

[3] See Appendix A.3.

$$t_\ell \left[p\sigma(10|t_{\ell\ell},\theta_0) + (1-p)\sigma(10|t_{\ell h},\theta_0)\right] \leq (1-t_\ell)\left[p\sigma(10|t_{\ell\ell},\theta_1) + (1-p)\sigma(10|t_{\ell h},\theta_1)\right] \tag{4.19}$$

$$\begin{aligned} t_\ell \left[2p\sigma(10|t_{\ell\ell},\theta_0) + (1-p)\left(\sigma(10|t_{\ell h},\theta_0) + \sigma(01|t_{\ell h},\theta_0)\right)\right] + \\ (1-t_\ell)\left[p\left(\sigma(10|t_{\ell h},\theta_1) + \sigma(01|t_{\ell h},\theta_1)\right) + 2(1-p)\sigma(10|t_{hh},\theta_1)\right] \leq \\ t_\ell \left[p\left(\sigma(10|t_{\ell h},\theta_0) + \sigma(01|t_{\ell h},\theta_0)\right) + 2(1-p)\sigma(10|t_{hh},\theta_0)\right] + \\ (1-t_\ell)\left[2p\sigma(10|t_{\ell\ell},\theta_1) + (1-p)\left(\sigma(10|t_{\ell h},\theta_1) + \sigma(01|t_{\ell h},\theta_1)\right)\right] \end{aligned} \tag{4.20}$$

$$t_h \left[p\sigma(01|t_{\ell h},\theta_0) + (1-p)\sigma(10|t_{hh},\theta_0)\right] \leq (1-t_h)\left[p\sigma(01|t_{\ell h},\theta_1) + (1-p)\sigma(10|t_{hh},\theta_1)\right] \tag{4.21}$$

$$\begin{aligned} t_h \left[p\left(\sigma(10|t_{\ell h},\theta_0) + \sigma(01|t_{\ell h},\theta_0)\right) + 2(1-p)\sigma(10|t_{hh},\theta_0)\right] + \\ (1-t_h)\left[2p\sigma(10|t_{\ell\ell},\theta_1) + (1-p)\left(\sigma(10|t_{\ell h},\theta_1) + \sigma(01|t_{\ell h},\theta_1)\right)\right] \leq \\ t_h \left[2p\sigma(10|t_{\ell\ell},\theta_0) + (1-p)\left(\sigma(10|t_{\ell h},\theta_0) + \sigma(01|t_{\ell h},\theta_0)\right)\right] + \\ (1-t_h)\left[p\left(\sigma(10|t_{\ell h},\theta_1) + \sigma(01|t_{\ell h},\theta_1)\right) + 2(1-p)\sigma(10|t_{hh},\theta_1)\right] \end{aligned} \tag{4.22}$$

$$\begin{aligned} t_h \left[p\left(\sigma(10|t_{\ell h},\theta_0) + \sigma(01|t_{\ell h},\theta_0)\right) + 2(1-p)\sigma(10|t_{hh},\theta_0)\right] + \\ (1-t_h)\left[p\sigma(10|t_{\ell\ell},\theta_1) + (1-p)\sigma(01|t_{\ell h},\theta_1)\right] \leq \\ t_h \left[p\sigma(10|t_{\ell\ell},\theta_0) + (1-p)\sigma(01|t_{\ell h},\theta_0)\right] + \\ (1-t_h)\left[p\left(\sigma(10|t_{\ell h},\theta_1) + \sigma(01|t_{\ell h},\theta_1)\right) + 2(1-p)\sigma(10|t_{hh},\theta_1)\right]. \end{aligned} \tag{4.23}$$

The first two inequalities refer to the low type: (4.19) is the obedience constraint against deviations to $\delta_i(a_i) = 0$ for any $a_i \in A_i$ and (4.20) is the truth-telling constraint. The remaining inequalities are for type $t_h$: (4.21) is the obedience constraint against deviations to $\delta_i(a_i) = 0$ for any $a_i \in A_i$, (4.22) is the truth-telling constraint, and (4.23) is the constraint that prevents double deviations, i.e. misreporting and deviating to $\delta_i(a_i) = 0$ for any $a_i \in A_i$. Contrary to the case of unanimity, we now have that obedience and truth-telling alone are no longer sufficient to prevent double deviations. This follows directly from the fact that,

when $k = 1$, the Sender can use two action profiles to persuade receivers to vote for $x_1$, so opening the way for a larger set of possible deviations.

As we did in the previous subsection, we characterize the solution set of (P-1) as a function of the prior probability $p$ by first illustrating subsets $\Sigma^D$, $\Sigma^E$, and $\Sigma^F$ of feasible persuasion mechanisms.

**Mechanisms in $\Sigma^D$:** The first set contains any mechanism $\sigma^D$ that satisfies the following properties:

- $\sigma^D(10|t_{\ell\ell}, \theta_1) = \sigma^D(10|t_{\ell h}, \theta_1) = \frac{1}{2}$,
- $\sigma^D(01|t_{\ell h}, \theta) = \sigma^D(10|t_{hh}, \theta) = 0$ for any $\theta \in \Theta$,
- $\sigma^D(10|t_{\ell\ell}, \theta_0) = \sigma^D(10|t_{\ell h}, \theta_0) = \frac{1-t_\ell}{2t_\ell}$.

In words, only receivers of the low type are recommended to vote for 1 with positive probability, irrespective of the state $\theta$. More specifically, the low type receives a recommendation to vote 1 with probability equal to $\frac{1}{2}$ in state $\theta_1$ and probability $\frac{1-t_\ell}{2t_\ell}$ in the other state. Consequently, when both receivers are of the low type, the alternative $x_1$ is implemented with probability 1 in state $\theta_1$ and probability $\frac{1-t_\ell}{t_\ell}$ in the other state. It is immediate to check that obedience constraints for both types are binding, and so are the truth-telling constraint for the low type and the double deviation constraint for the high type.

**Mechanisms in $\Sigma^E$:** Any mechanism $\sigma^E$ in the second set has the following properties:

- $\sigma^E(10|t_{\ell\ell}, \theta_1) = \frac{1}{2}$,
- $\sigma^E(10|t_{\ell h}, \theta_1) = 1 - \sigma^E(01|t_{\ell h}, \theta_1)$,
- $\sigma^E(10|t_{hh}, \theta_1) = 0$,
- $\sigma^E(10|t_{\ell\ell}, \theta_0) = \frac{(1-p)(1-t_h)}{pt_h}\sigma^E(01|t_{\ell h}, \theta_1) - \frac{1-p}{p}\sigma^E(01|t_{\ell h}, \theta_0) + \frac{(2p-1)(1-t_h)}{2t_h} + \frac{(1-p)(1-t_\ell)}{t_\ell}$,
- $\sigma^E(10|t_{\ell h}, \theta_0) = -\frac{2(1-t_h)}{t_h}\sigma^E(01|t_{\ell h}, \theta_1) + \sigma^E(01|t_{\ell h}, \theta_0) + \frac{p(1-t_h)}{t_h} + \frac{(1-p)(1-t_\ell)}{t_\ell}$,
- $\sigma^E(10|t_{hh}, \theta_0) = \frac{p(1-t_h)}{(1-p)t_h}\sigma^E(01|t_{\ell h}, \theta_1) - \frac{p}{1-p}\sigma^E(01|t_{\ell h}, \theta_0)$,
- $\sigma^E(01|t_{\ell h}, \theta_0)$ and $\sigma^E(01|t_{\ell h}, \theta_1)$ are such that:

$$\sigma^E(01|t_{\ell h}, \theta_0), \sigma^E(01|t_{\ell h}, \theta_1) \in (0, 1) \tag{4.24}$$

$$\sigma^E(01|t_{\ell h}, \theta_1) < \frac{p}{2(1-p)} \tag{4.25}$$

$$\sigma^E(01|t_{\ell h},\theta_1) > \frac{t_h}{1-t_h}\sigma^E(01|t_{\ell h},\theta_0) \tag{4.26}$$

$$\sigma^E(01|t_{\ell h},\theta_1) < \frac{t_h}{1-t_h}\sigma^E(01|t_{\ell h},\theta_0) + \frac{(1-p)t_h}{2p(1-t_h)} \tag{4.27}$$

$$\sigma^E(01|t_{\ell h},\theta_1) < \frac{t_h}{1-t_h}\sigma^E(01|t_{\ell h},\theta_0) + \frac{pt_h}{2(1-p)(1-t_h)} - \frac{p(2p-1)}{2(1-p)} - \frac{pt_h(1-t_\ell)}{t_\ell(1-t_h)} \tag{4.28}$$

$$\sigma^E(01|t_{\ell h},\theta_1) < \frac{t_h}{2(1-t_h)}\sigma^E(01|t_{\ell h},\theta_0) + \frac{p}{2} + \frac{(1-p)(1-t_\ell)t_h}{2(1-t_h)t_\ell}. \tag{4.29}$$

Here the high type is persuaded to vote 1 in state $\theta_1$ only if the other receiver is of the low type. Furthermore, the alternative $x_1$ is implemented with probability 1 in state $\theta_1$ if there is at least a receiver of the low type. The remaining action recommendations are such that the truth-telling constraint for the low type is binding, and so are the obedience and double deviations constraints for the high type.

**Mechanisms in $\Sigma^F$:** Any mechanism $\sigma^F$ in the third subset has the following properties:

- $\sigma^F(10|t_{\ell\ell},\theta_1) = \sigma^F(10|t_{hh},\theta_1) = \frac{1}{2}$,
- $\sigma^F(10|t_{\ell h},\theta_1) = \sigma^F(01|t_{\ell h},\theta_1) = \frac{1}{2}$,
- $\sigma^F(10|t_{\ell\ell},\theta_0) = \sigma^F(10|t_{\ell h},\theta_0) = \sigma^F(01|t_{\ell h},\theta_0) = \sigma^F(10|t_{hh},\theta_0) = \frac{1-t_h}{2t_h}$.

In the last subset, the Sender implements $x_1$ with probability 1 in state $\theta_1$, and with probability $\frac{1-t_h}{t_h}$ in the other state. Furthermore, the obedience constraint for the low type is slack while all the remaining constraints are binding.

We can now establish the following.

**Proposition 11.** *Let $k=1$ and assume $\frac{1-t_h}{t_h} - \frac{4(1-t_\ell)}{t_\ell} \geq -3$. The following is true:*

1. *if $p \geq \frac{t_h}{2t_h-t_\ell}$, then any $\sigma \in \Sigma^D$ solves (P-1);*
2. *if $p \in \left[\frac{t_\ell}{2t_h-t_\ell}, \frac{t_h}{2t_h-t_\ell}\right]$, then any $\sigma \in \Sigma^E$ solves (P-1);*
3. *if $p \leq \frac{t_\ell}{2t_h-t_\ell}$, then any $\sigma \in \Sigma^F$ solves (P-1);*
4. *if $p = \frac{t_h}{2t_h-t_\ell}$, then any $\sigma \in \mathsf{conv}\left(\Sigma^D \cup \Sigma^E\right)$ solves (P-1);*
5. *f $p = \frac{t_\ell}{2t_h-t_\ell}$, then any $\sigma \in \mathsf{conv}\left(\Sigma^E \cup \Sigma^F\right)$ solves (P-1).*

*Proof.* See Appendix A.4. □

The assumption $\frac{1-t_h}{t_h} - \frac{4(1-t_\ell)}{t_\ell} \geq -3$ ensures that parameters $t_\ell$ and $t_h$ are not too far apart form each other. The interpretation of Proposition 11 is analogous to the case with unanimity. We can partition the range of $p$ into three regions, and the choice of the optimal persuasion mechanism crucially depends on the region where the prior lies. If $p$ is high enough, the Sender tailors her persuasion strategy toward the low type. More specifically, she recommends to vote 1 with strictly positive probability to any receiver that submits a low report, and with probability zero if the report is $t_h$. Contrary to the case with unanimity, this recommendation strategy enables the Sender to implement her preferred alternative with positive probability in either state even when the profile of reports is $t_{\ell h}$ or $t_{h\ell}$, and not just $t_{\ell\ell}$. This clearly depends on the fact that, under single approval, two action recommendation profiles can be used to implement $x_1$, giving the Sender more leeway than under unanimity. The reason why high types are never told to vote 1 is essentially the same as in the case with $k = 2$.

In case 2., the Sender is sufficiently confident that both committee members cannot be of the high type. As a consequence, when both receivers submit a report $t_h$, they are told not to vote for $x_1$ contingent on state $\theta_1$. The remaining action recommendations are such that the truth-telling constraint for the low type, and the obedience and double-deviations constraints for the high type are all binding. In the remaining case, the prior $p$ is so low that both receivers are likely to have preference parameter $t_h$. Therefore, $x_1$ is implemented with probability 1 in state $\theta_1$ and probability $\frac{1-t_h}{t_h}$ in state $\theta_0$, irrespective of the profile of reports. Finally, notice that, when $p \leq \frac{t_\ell}{2t_h - t_\ell}$, the overall probability of implementing $x_1$ is the same for both voting rules. In the complementary case, it is not surprising that such a probability is strictly greater when a single approval is needed.

Finally, substituting into (4.18), we obtain the value function for each optimal mechanism:

$$\begin{aligned} U_S(\sigma^D) &= \frac{p}{2t_\ell}, \\ U_S(\sigma^E) &= \frac{p^2}{2t_h} + \frac{p(1-p)}{t_\ell}, \\ U_S(\sigma^F) &= \frac{1}{2t_h}. \end{aligned}$$

It is clear that $U_S(\sigma^D)$ is decreasing with respect to $t_\ell$, $U_S(\sigma^F)$ is decreasing with respect to $t_h$, and $U_S(\sigma^E)$ is decreasing with respect to both $t_\ell$ and $t_h$. As far as the prior $p$ is concerned, $U_S(\sigma^D)$ increases linearly with respect to it, while $U_S(\sigma^F)$ is independent of it. Finally, $U_S(\sigma^E)$ is increasing with respect to $p$ as long as $p \leq \frac{t_h}{2t_h - t_\ell}$, and it does so at a

decreasing rate. The interpretation is the same as when $k = 2$.

## 4.4 Persuasion without information elicitation

In this section we characterize the persuasion mechanism that solves the Sender's choice problem under two main assumptions: 1) each receiver $i$ privately observes his preference parameter $t_i$; and 2) the Sender is *not* allowed to elicit receivers' private information by asking them to report their types.

Contrary to section 4.3, now the Sender does not have access to the private information that receivers possess about their preferences. This can be justified in several ways. For instance, one may assume that receivers have a taste for privacy: they just prefer not to disclose any information at all about themselves to external parties. Alternatively, one may argue that the Sender has limited commitment power. As we have seen in the previous section, if information elicitation is allowed, then the Sender collects reports in order to choose the profile of action recommendations she prefers. If receivers believe that the Sender's promise to treat reports confidentially is not credible, then one can model the strategic interaction between Sender and receivers by dispensing with any type of information reporting.

### 4.4.1 The solution concept

The interaction between Sender and receivers takes place according to the sequence of events we outlined in the previous Section, with the important difference that now receivers do not transmit any information at all to the Sender. In other words, we now have a scenario of unilateral communication (from Sender to receivers) whereas communication is bilateral in Section 4.3. The fundamental object we are interested in is still a (direct) persuasion mechanism $\sigma : T \times \Theta \longrightarrow \Delta(A)$. However, due to the absence of information elicitation, we have to restrict the set of feasible mechanisms. Following Bergemann and Morris (2019), we require $\sigma$ to be *publicly feasible*. Before giving a formal definition, we need to introduce some additional objects.

For any $i \in I$, let $B_i$ be the set of pure strategies in the fundamental voting game, with typical element $b_i : T_i \longrightarrow A_i$. With a slight abuse of notation, we denote $b_i(t_i)$ as the action $a_i \in A_i$ that strategy $b_i$ prescribes to type $t_i$ of receiver $i$. In addition, $B = B_1 \times B_2$. Now let $\phi : \Theta \longrightarrow \Delta(B)$ be a *strategy recommendation*. In words, $\phi$ is a rule that, conditional on state $\theta$, selects the profile $b \in B$ of pure strategies with probability $\phi(b|\theta)$. We say that $\phi$

induces the persuasion mechanism $\sigma$ if, for any $a \in A$, $t \in T$, and $\theta \in \Theta$, we have

$$\sigma(a|t,\theta) = \sum_{\{b \in B : b(t)=a\}} \phi(b|\theta). \tag{4.30}$$

Furthermore, the strategy recommendation $\phi : \Theta \longrightarrow \Delta(B)$ is *obedient* if, for every $i \in I$, $t_i \in T_i$, and $b_i \in B_i$,

$$\begin{aligned} \sum_{b_j \in B_j, t_j \in T_j, \theta \in \Theta} & \psi(\theta)\pi(t_i,t_j)\phi(b_i,b_j|\theta)u_i((b_i(t_i),b_j(t_j)),(t_i,t_j),\theta) \geq \\ & \sum_{b_j \in B_j, t_j \in T_j, \theta \in \Theta} \psi(\theta)\pi(t_i,t_j)\phi(b_i,b_j|\theta)u_i((a_i',b_j(t_j)),(t_i,t_j),\theta) \end{aligned} \tag{4.31}$$

for all $a_i' \in A_i$.

**Definition 2** (Publicly feasible obedience.)**.** *A persuasion mechanism $\sigma : T \times \Theta \longrightarrow \Delta(A)$ is publicly feasible obedient if there exists a strategy recommendation $\phi : \Theta \longrightarrow \Delta(B)$ that induces $\sigma$ according to (4.30) and that is obedient according to (4.31).*

The restriction to publicly feasible mechanisms can be interpreted as follows. Since the Sender does not have access to receivers' private information, she sends recommendations according to a rule $\phi$ that does not depend on reports. Consequently, the Sender no longer recommends a single action to each receiver. Rather, she recommends a profile of actions, one for each possible receiver type. This captures the fact that the Sender does not know the true preference parameter of the receiver she is facing, nor can she rely on reports. Notice that, even though $\phi$ is independent from $T$, the persuasion mechanism $\sigma$ induced by $\phi$ is not necessarily so. Furthermore, it is straightforward to see that any publicly feasible obedient mechanism $\sigma$ is also a BCE, while a BCE is not necessarily publicly feasible.

Given a prior distribution $\pi$ over preference types, the Sender's choice problem without information elicitation is:

$$\max_{\sigma} \quad U_S(\sigma) = \frac{1}{2}\left[\sum_{t \in T}\pi(t)\left(\sum_{a \in A^{x_1}}\sigma(a|t,\theta_0) + \sum_{a \in A^{x_1}}\sigma(a|t,\theta_1)\right)\right] \tag{P-2}$$

such that $\sigma$ is:

- *publicly feasible obedient*;
- symmetric;
- pivotal.

Notice that we require the persuasion mechanism $\sigma$ to be symmetric, but we do not impose such a requirement on the strategy recommendation $\phi$ that induces it. That is, we do not require that $\phi(b, b'|\theta) = \phi(b', b|\theta)$ in every state and for every pair of pure strategies $b, b'$. In general, imposing symmetry on $\phi$ implies that the corresponding mechanism $\sigma$ is symmetric too, while the converse is not necessarily true. However, as we show in Appendices A.5 and A.7, every strategy recommendation $\phi$ that we illustrate in this section without information elicitation turns out to be symmetric too.

### 4.4.2 Unanimity

We proceed to give a full characterization of the solution to problem (P-2) when $k = 2$. Rather than optimizing directly with respect to mechanisms $\sigma : T \times \Theta \longrightarrow \Delta(A)$, it is convenient to use strategy recommendations $\phi : \Theta \longrightarrow \Delta(B)$ as control variables. In order to do that, notice that in our binary environment, for each $i \in I$, the set $B_i$ contains four pure strategies. More specifically, we have $B_1 = B_2 = \{b_{00}, b_{01}, b_{10}, b_{11}\}$, where the first subscript denotes the action recommended to the low type, and the second subscript is the action recommended to the high type. One can show[4] that, given $\phi$, the Sender's objective function in (P-2) reduces to

$$U_S(\phi) = \frac{1}{2} \sum_{\theta \in \Theta} \left[ \phi(b_{11}, b_{11}|\theta) + p^2 \phi(b_{10}, b_{10}|\theta) + 2p\phi(b_{10}, b_{11}|\theta) \right]. \tag{4.32}$$

In addition, the set of relevant obedience constraints is:

$$t_\ell \left[ \phi(b_{10}, b_{11}|\theta_0) + p\phi(b_{10}, b_{10}|\theta_0) \right] \leq (1 - t_\ell) \left[ \phi(b_{10}, b_{11}|\theta_1) + p\phi(b_{10}, b_{10}|\theta_1) \right], \tag{4.33}$$

$$(1 - t_h) \left[ \phi(b_{10}, b_{11}|\theta_1) + p\phi(b_{10}, b_{10}|\theta_1) \right] \leq t_h \left[ \phi(b_{10}, b_{11}|\theta_0) + p\phi(b_{10}, b_{10}|\theta_0) \right], \tag{4.34}$$

$$t_h \left[ \phi(b_{11}, b_{11}|\theta_0) + p\phi(b_{11}, b_{10}|\theta_0) \right] \leq (1 - t_h) \left[ \phi(b_{11}, b_{11}|\theta_1) + p\phi(b_{11}, b_{10}|\theta_1) \right], \tag{4.35}$$

where the first inequality guarantees that type $t_\ell$ obeys strategy $b_{10}$. Constraints (4.34) and (4.35) make sure that type $t_h$ finds it optimal to obey strategies $b_{10}$ and $b_{11}$, respectively.

The fact that only three incentive constraints are sufficient can be explained as follows. A strategy recommendation can prescribe either the same action to each type or different actions to different types. Bearing in mind that $t_\ell$ is easier to persuade than $t_h$ to vote for 1, if both types are told to vote 0, then it suffices to consider the obedience constraint for the

[4] See Appendix A.5.

low type. Clearly, it is sufficient to consider the obedience constraint for the high type if the recommendation is 1. By the same token, it cannot be incentive compatible to recommend 0 to $t_\ell$ and 1 to $t_h$. In the remaining case, one has to consider both incentive constraints. Therefore, under unanimity, the Sender can induce both receivers to adopt option $x_1$ by choosing strategy recommendation profiles where only $b_{10}$ and $b_{11}$ are used.

Like we did in section 4.3, we characterize the optimal mechanism as a function of $p$ by first illustrating three subsets of strategy recommendations that induce publicly feasible mechanisms. More specifically, we denote $\Phi$ as the set of feasible strategy recommendations, and focus on subsets $\Phi^A$, $\Phi^B$, and $\Phi^C$.

**Strategy recommendations in $\Phi^A$:** The first subset contains any strategy recommendation $\phi^A$ such that:

- $\phi^A(b_{10}, b_{10}|\theta_1) = 1$,
- $\phi^A(b_{10}, b_{10}|\theta_0) = \frac{1-t_\ell}{t_\ell}$,
- $\phi^A(b_{10}, b_{11}|\theta) = \phi^A(b_{11}, b_{11}|\theta) = 0$ for any $\theta \in \Theta$.

In words, the Sender recommends strategy $b_{10}$ to either player with probability 1 contingent on state $\theta_1$, and with probability determined by the low type (and less than 1) in the other state. Furthermore, notice that the persuasion mechanism $\sigma$ induced by $\phi^A$ via (4.30) is such that the action profile $(1, 1)$ is implemented with positive probability only when both receivers are of the low type. More precisely, $\sigma(1|t_{\ell\ell}, \theta_1) = 1$, $\sigma(1|t_{\ell\ell}, \theta_0) = \frac{1-t_\ell}{t_\ell}$, and $\sigma(1|t, \theta) = 0$ for any $t \neq t_{\ell\ell}$ and any $\theta$. This shows that mechanisms in $\Sigma^A$ are publicly feasible.

**Strategy recommendations in $\Phi^B$:** The second subset contains any strategy recommendation $\phi^B$ that has the following properties:

- $\phi^B(b_{10}, b_{11}|\theta_1) = \frac{1}{2}$,
- $\phi^B(b_{10}, b_{10}|\theta_1) = \phi^B(b_{11}, b_{11}|\theta_1) = 0$,
- $\phi^B(b_{10}, b_{11}|\theta_0) \in \left[0, \frac{1-t_h}{2t_h}\right]$,
- $\phi^B(b_{10}, b_{10}|\theta_0) = \frac{1}{p}\left(\frac{1-t_\ell}{2t_\ell} - \phi^B(b_{10}, b_{11}|\theta_0)\right)$,
- $\phi^B(b_{11}, b_{11}|\theta_0) = p\left(\frac{1-t_h}{2t_h} - \phi^B(b_{10}, b_{11}|\theta_0)\right)$.

The mechanism $\sigma$ induced by $\phi^B$ is such that $\sigma(1|t_{\ell\ell},\theta_1)=1$, $\sigma(1|t_{\ell h},\theta_1)=\sigma(1|t_{h\ell},\theta_1)=\frac{1}{2}$, and $\sigma(1|t_{hh},\theta_1)=0$. The strategy recommendations contingent on $\theta_0$ are such that incentive constraints (4.33) and (4.35) are both binding. Notice that the persuasion mechanism $\sigma$ induced by $\phi^B$ does not belong to $\Sigma^B$.

**Strategy recommendations in $\Phi^C$:** Finally, the third subset contains any strategy recommendation $\phi^C$ such that:

- $\phi^C(b_{11},b_{11}|\theta_1)=1$,
- $\phi^C(b_{11},b_{11}|\theta_0)=\frac{1-t_h}{t_h}$,
- $\phi^C(b_{10},b_{10}|\theta)=\phi^C(b_{10},b_{11}|\theta)=0$ for any $\theta\in\Theta$.

In words, the Sender recommends strategy $b_{11}$ to either player with probability 1 contingent on state $\theta_1$, and with probability determined by the high type (and less than 1) in the other state. Furthermore, the persuasion mechanism $\sigma$ induced by $\phi^C$ is such that the action profile $(1,1)$ is implemented with positive probability for every type profile in either state. More precisely, $\sigma(1|t,\theta_1)=1$ and $\sigma(1|t,\theta_0)=\frac{1-t_h}{t_h}$ for any $t\in T$. Notice that the induced mechanism $\sigma$ does belong to the set $\Sigma^C$.

We can now establish the following.

**Proposition 12.** *Let $k=2$ and let $\phi\in\Phi$. Then we have:*

1. *If $p>\frac{t_\ell+t_h}{2t_h}$, then $\phi$ solves (P-2) if and only if $\phi\in\Phi^A$;*
2. *If $p\in\left(\frac{2t_\ell}{t_\ell+t_h},\frac{t_\ell+t_h}{2t_h}\right)$, then $\phi$ solves (P-2) if and only if $\phi\in\Phi^B$;*
3. *If $p<\frac{2t_\ell}{t_\ell+t_h}$, then $\phi$ solves (P-2) if and only if $\phi\in\Phi^C$;*
4. *If $p=\frac{t_\ell+t_h}{2t_h}$, then $\phi$ solves (P-2) if and only if $\phi\in\mathsf{conv}(\Phi^A\cup\Phi^B)$;*
5. *If $p=\frac{2t_\ell}{t_\ell+t_h}$, then $\phi$ solves (P-2) if and only if $\phi\in\mathsf{conv}(\Phi^B\cup\Phi^C)$.*

*Proof.* See Appendix A.6. □

The interpretation is analogous to the case with information elicitation. As we already hinted, mechanisms in $\Sigma^A$ are publicly feasible. This implies that mechanisms in $\Sigma^A$ and strategy recommendations in $\Phi^A$ are payoff-equivalent. Similarly, mechanisms in $\Sigma^C$ and strategy recommendations in $\Phi^C$ are payoff-equivalent, even though the former are not necessarily publicly feasible. Mechanisms in $\Sigma^B$ are not publicly feasible, and they give the Sender

a strictly larger expected utility than recommendations in $\Phi^B$. Therefore, public feasibility is a substantial restriction to the Sender's persuasion strategy only when $p \in \left(\frac{t_\ell}{t_h}, \frac{t_\ell + t_h}{2t_h}\right)$.

Finally, substituting into (4.32), we obtain the value function for each optimal mechanism:

$$\begin{aligned} U_S(\phi^A) &= \frac{p^2}{2t_\ell}, \\ U_S(\phi^B) &= \frac{p(t_\ell + t_h)}{4t_\ell t_h}, \\ U_S(\phi^C) &= \frac{1}{2t_h}. \end{aligned}$$

It is easy to see that $U_S(\phi^A)$ decreases with respect to $t_\ell$, that $U_S(\phi^B)$ is decreasing with respect to both $t_\ell$ and $t_h$, and $U_S(\phi^C)$ decreases with respect to $t_h$. As far as $p$ is concerned, $U_S(\phi^A)$ increases with respect to it at a quadratic rate, $U_S(\phi^B)$ grows linearly with it, and $U_S(\phi^C)$ is independent ot if.

### 4.4.3 Single approval

We now characterize the solution to (P-2) when $k = 1$. Even in this case we optimize directly with respect to strategy recommendations $\phi$ rather than mechanisms $\sigma$. One can show[5] that the Sender's objective function reduces to

$$U_S(\phi) = \sum_{\theta \in \Theta} \left[\phi(b_{11}, b_{00}|\theta) + p\phi(b_{10}, b_{00}|\theta)\right]. \tag{4.36}$$

In addition, the set of relevant incentive constraints is

$$(1 - t_\ell)\left[\phi(b_{00}, b_{00}|\theta_1) + (1-p)\phi(b_{10}, b_{00}|\theta_1)\right] \leq t_\ell\left[\phi(b_{00}, b_{00}|\theta_0) + (1-p)\phi(b_{10}, b_{00}|\theta_0)\right], \tag{4.37}$$

$$t_\ell\left[\phi(b_{10}, b_{00}|\theta_0)\right] \leq (1 - t_\ell)\left[\phi(b_{10}, b_{00}|\theta_1)\right], \tag{4.38}$$

$$(1 - t_h)\left[\phi(b_{10}, b_{00}|\theta_1)\right] \leq t_h\left[\phi(b_{10}, b_{00}|\theta_0)\right], \tag{4.39}$$

$$t_h\left[\phi(b_{11}, b_{00}|\theta_0)\right] \leq (1 - t_h)\left[\phi(b_{11}, b_{00}|\theta_1)\right], \tag{4.40}$$

[5] See Appendix A.7.

where the first constraint guarantees that type $t_\ell$ obeys strategy $b_{00}$, constraints (4.38) and (4.39) make sure that both types obey strategy $b_{10}$, and (4.40) is the obedience constraint for type $t_h$ and strategy $b_{11}$.

Contrary to the case with unanimity, the assumption of pivotality is now a substantial restriction. More specifically, it implies that only strategy recommendations where at least one of the receivers is told to adopt $b_{00}$ can be used. As a consequence, we can characterize the optimal strategy recommendation by making use of just two subsets of feasible strategies.

**Strategy recommendations in $\Phi^D$:** The first subset contains any strategy recommendation $\phi^D$ that has the following properties:

- $\phi^D(b_{10}, b_{00}|\theta_1) = \frac{1}{2}$,
- $\phi^D(b_{10}, b_{00}|\theta_0) = \frac{1-t_\ell}{2t_\ell}$,
- $\phi^D(b_{11}, b_{00}|\theta) = 0$ for any $\theta \in \Theta$.

In words, only strategies with $t_h$ voting for 0 are recommended with positive probability in either state. More specifically, the persuasion mechanism $\sigma$ induced by $\phi^D$ via (4.30) is such that $\sigma(10|t_{\ell\ell}, \theta_1) + \sigma(01|t_{\ell\ell}, \theta_1) = 1$, $\sigma(10|t_{\ell h}, \theta_1) = \sigma(01|t_{h\ell}, \theta_1) = \frac{1}{2}$, and $\sigma(10|t_{hh}, \theta_1) = \sigma(01|t_{hh}, \theta_1) = 0$. In addition, $\sigma(10|t_{\ell\ell}, \theta_0) + \sigma(01|t_{\ell\ell}, \theta_0) = \frac{1-t_\ell}{t_\ell}$, $\sigma(10|t_{\ell h}, \theta_0) = \sigma(01|t_{h\ell}, \theta_0) = \frac{1-t_\ell}{2t_\ell}$, and $\sigma(10|t_{hh}, \theta_0) = \sigma(01|t_{hh}, \theta_0) = 0$.

**Strategy recommendations in $\Phi^E$:** The second subset contains any strategy recommendation $\phi^E$ that has the following properties:

- $\phi^E(b_{11}, b_{00}|\theta_1) = \frac{1}{2}$,
- $\phi^E(b_{11}, b_{00}|\theta_0) = \frac{1-t_h}{2t_h}$,
- $\phi^E(b_{10}, b_{00}|\theta) = 0$ for any $\theta \in \Theta$.

In this case, the mechanism induced by $\phi^E$ is such that $\sigma(10|t, \theta_1) + \sigma(01|t, \theta_1) = 1$ for every $t \in T$ and $\sigma(10|t, \theta_0) + \sigma(01|t, \theta_0) = \frac{1-t_h}{t_h}$. Therefore, the alternative $x_1$ is implemented with probability 1 conditional on state $\theta_1$, and with probability $\frac{1-t_h}{t_h}$ conditional on the other state, irrespective of the true preference profile.

We can now establish the following.

**Proposition 13.** *Let $k = 1$ and let $\phi \in \Phi$. Then we have:*

1. *If $p > \frac{t_\ell}{t_h}$, then $\phi$ solves (P-2) if and only if $\phi \in \Phi^D$;*

2. *If* $p < \frac{t_\ell}{t_h}$, *then* $\phi$ *solves* (P-2) *if and only if* $\phi \in \Phi^E$;

3. *If* $p = \frac{t_\ell}{t_h}$, *then* $\phi$ *solves* (P-2) *if and only if* $\phi \in \mathsf{conv}\left(\Phi^D \cup \Phi^E\right)$.

*Proof.* See Appendix A.8. □

Strategy recommendations in $\Phi^D$ are payoff-equivalent to mechanisms in $\Sigma^D$, and those in $\Phi^E$ are payoff-equivalent to mechanisms in $\Sigma^F$. Therefore, public feasibility turns out to be a relevant restriction to the Sender's persuasion strategy only when $p \in \left(\frac{t_\ell}{2t_h - t_\ell}, \frac{t_h}{2t_h - t_\ell}\right)$.

Finally, substituting into (4.36), we obtain the value function for each optimal mechanism:

$$U_S(\phi^D) = \frac{p}{2t_\ell},$$
$$U_S(\phi^E) = \frac{1}{2t_h}.$$

It is immediate to see how either function varies with respect to parameters $p$, $t_\ell$, and $t_h$.

## 4.5 Discussion

1. When preferences are commonly known, the Sender manages to implement her preferred alternative with probability 1 contingent on state $\theta_1$. Under incomplete information, this is not necessarily the case. More specifically, the Sender may find it optimal to implement $x_1$ in $\theta_1$ with probability 1 only for a strict subset of type profiles. Intuitively, to increase the overall probability of implementing $x_1$ in state $\theta_1$, the Sender has to decrease the overall probability of implementing $x_1$ in the other state. If she does not, receivers may not be willing to obey the action recommendation to vote for 1 and, even if they are, receivers of the low type have an incentive to misreport their true preference parameter $t_\ell$. Depending on the prior $p$, the expected gain from implementing $x_1$ in state $\theta_1$ with larger probability could be more than compensated by the loss incurred in the other state $\theta_0$.

2. With just one receiver, the qualitative features of the optimal persuasion strategy are essentially the same. More precisely, if $p < \frac{t_\ell}{t_h}$ the Sender's optimal strategy is to recommend action 1 to either type with probability 1 conditional on state $\theta_1$, and with probability $\frac{1-t_h}{t_h}$ conditional on the other state. If $p > \frac{t_\ell}{t_h}$, then she always recommends action 0 to the high type, whereas type $t_\ell$ is told to vote 1 with probability 1 in state $\theta_1$ and with probability $\frac{1-t_\ell}{t_\ell}$ in state $\theta_0$. When $p = \frac{t_\ell}{t_h}$, then any mixture between the two mechanisms is optimal. Notice that the corresponding Sender's value function is exactly the same as in the case of publicly feasible persuasion when $k = 1$. Furthermore,

the Sender never has a strict incentive to use information elicitation when only one receiver is present. Bergemann and Morris (2019) show that this equivalence between information elicitation and public feasibility holds for every game with one receiver and a binary set of states $\Theta$.

3. We have assumed that prior distributions $\psi \in \Delta(\Theta)$ and $\pi \in \Delta(T)$ are statistically independent. This means that, when the Sender observes $\theta$, she cannot learn anything more about preference types than what is already contained in the prior. Similarly, receivers do not learn additional information about $\theta$ from learning their preference parameters. The assumption is made in order to isolate the effect of uncertainty about preference parameters. In the limit case of perfect correlation between the two variables, we have a scenario where both Sender and receivers are able to learn the true profile $(t, \theta)$ of payoff-relevant parameters, so that any persuasion strategy becomes fruitless.

## 4.6 Conclusion

We have studied the information design problem of a Sender who wants to persuade a committee whose members have privately known preferences, i.e. thresholds of reasonable doubt. We have seen how the optimal persuasion strategy depends on the informativeness of the prior distribution over preference types. The Sender finds it optimal to target her strategy towards the profile of preference types that is deemed more likely to occur. Both "separation" and "pooling" may occur. When both receivers are more likely to be of the low type, every report of being of the high type receives a recommendation to not vote for the policy that the Sender prefers. On the other hand, when the committee is more likely to be composed of high types, the Sender implements her preferred policy with the same probability independently of reported types. The possibility of relying on private reports is beneficial to the Sender. When the persuasion strategy is required to be publicly feasible, the Sender turns out to be worse off for intermediate values of the prior. In particular, when the committee is more likely to be heterogeneous, She can discriminate between type profiles that are homogeneous, but not between profiles where both high and low types are present. Showing that uncertainty is always costly for the Sender, our results offer a theoretical insight for why firms or political parties are willing to invest in order to obtain information about their potential voters or customers that is as detailed as possible, e.g. microtargeting. Finally, we conjecture that most of the qualitative features of the optimal persuasion strategy carry over to the case with more than two players, either with unanimous

or non unanimous voting rules. A full analysis of this case is left for future research.

# Appendix A

# Proofs and additional computation for Chapter 4

## A.1 Incentive constraints for the case with information elicitation and $k = 2$

The whole set of BCE incentive compatibility constraints for type $t_\ell$ is:

$$t_\ell \left[p\sigma(1|t_{\ell\ell},\theta_0) + (1-p)\sigma(1|t_{\ell h},\theta_0)\right] \leq (1-t_\ell)\left[p\sigma(1|t_{\ell\ell},\theta_1) + (1-p)\sigma(1|t_{\ell h},\theta_1)\right] \quad \text{(A.1)}$$

$$(1-t_\ell)\left[p\sigma(10|t_{\ell\ell},\theta_1) + (1-p)\sigma(01|t_{\ell h},\theta_1)\right] \leq t_\ell \left[p\sigma(10|t_{\ell\ell},\theta_0) + (1-p)\sigma(01|t_{\ell h},\theta_0)\right] \quad \text{(A.2)}$$

$$\begin{aligned} &t_\ell \left[p\sigma(1|t_{\ell\ell},\theta_0) + (1-p)\sigma(1|t_{\ell h},\theta_0)\right] + (1-t_\ell)\left[p\sigma(10|t_{\ell\ell},\theta_1) + (1-p)\sigma(01|t_{\ell h},\theta_1)\right] \leq \\ &\quad (1-t_\ell)\left[p\sigma(1|t_{\ell\ell},\theta_1) + (1-p)\sigma(1|t_{\ell h},\theta_1)\right] + t_\ell \left[p\sigma(10|t_{\ell\ell},\theta_0) + (1-p)\sigma(01|t_{\ell h},\theta_0)\right] \end{aligned} \quad \text{(A.3)}$$

$$\begin{aligned} &t_\ell \left[p\sigma(1|t_{\ell\ell},\theta_0) + (1-p)\sigma(1|t_{\ell h},\theta_0)\right] + (1-t_\ell)\left[p\sigma(1|t_{\ell h},\theta_1) + (1-p)\sigma(1|t_{hh},\theta_1)\right] \leq \\ &\quad t_\ell \left[p\sigma(1|t_{\ell h},\theta_0) + (1-p)\sigma(1|t_{hh},\theta_0)\right] + (1-t_\ell)\left[p\sigma(1|t_{\ell\ell},\theta_1) + (1-p)\sigma(1|t_{\ell h},\theta_1)\right] \end{aligned} \quad \text{(A.4)}$$

$$t_\ell \left[p\sigma(1|t_{\ell\ell},\theta_0) + (1-p)\sigma(1|t_{\ell h},\theta_0)\right] + (1-t_\ell)\left[p\sigma(1|t_{\ell h},\theta_1) + (1-p)\sigma(1|t_{hh},\theta_1)\right] + (1-t_\ell)\left[p\sigma(10|t_{\ell h},\theta_1) + (1-p)\sigma(10|t_{hh},\theta_1)\right] \leq t_\ell \left[p\sigma(1|t_{\ell h},\theta_0) + (1-p)\sigma(1|t_{hh},\theta_0)\right] + (1-t_\ell)\left[p\sigma(1|t_{\ell\ell},\theta_1) + (1-p)\sigma(1|t_{\ell h},\theta_1)\right] + t_\ell \left[p\sigma(10|t_{\ell h},\theta_0) + (1-p)\sigma(01|t_{hh},\theta_0)\right] \quad \text{(A.5)}$$

$$t_\ell \left[p\sigma(1|t_{\ell\ell},\theta_0) + (1-p)\sigma(1|t_{\ell h},\theta_0)\right] + (1-t_\ell)\left[p\sigma(10|t_{\ell h},\theta_1) + (1-p)\sigma(10|t_{hh},\theta_1)\right] \leq t_\ell \left[p\sigma(10|t_{\ell h},\theta_0) + (1-p)\sigma(01|t_{hh},\theta_0)\right] + (1-t_\ell)\left[p\sigma(1|t_{\ell\ell},\theta_1) + (1-p)\sigma(1|t_{\ell h},\theta_1)\right], \quad \text{(A.6)}$$

where (A.1) is the obedience constraint against deviations to $\delta_i(a_i) = 0$ for any $a_i \in A_i$, (A.2) is the obedience constraint against deviations to $\delta_i(a_i) = 1$ for any $a_i \in A_i$, (A.3) is the obedience constraint against deviations to $\delta_i(1) = 0$ and $\delta_i(0) = 1$, (A.4) is the truth-telling constraint, (A.5) is the truth-telling constraint against deviations to $\delta_i(a_i) = 1$ for any $a_i \in A_i$, and (A.6) is the truth-telling constraint against deviations to $\delta_i(1) = 0$ and $\delta_i(0) = 1$. The truth-telling constraint against deviations to $\delta_i(a_i) = 0$ for any $a_i \in A_i$ coincides with (A.1).

Similarly, the set of BCE constraints for type $t_h$ is:

$$t_h \left[p\sigma(1|t_{\ell h},\theta_0) + (1-p)\sigma(1|t_{hh},\theta_0)\right] \leq (1-t_h)\left[p\sigma(1|t_{\ell h},\theta_1) + (1-p)\sigma(1|t_{hh},\theta_1)\right] \quad \text{(A.7)}$$

$$(1-t_h)\left[p\sigma(10|t_{\ell h},\theta_1) + (1-p)\sigma(01|t_{hh},\theta_1)\right] \leq t_h \left[p\sigma(10|t_{\ell h},\theta_0) + (1-p)\sigma(01|t_{hh},\theta_0)\right] \quad \text{(A.8)}$$

$$t_h \left[p\sigma(1|t_{\ell h},\theta_0) + (1-p)\sigma(1|t_{hh},\theta_0)\right] + (1-t_h)\left[p\sigma(10|t_{\ell h},\theta_1) + (1-p)\sigma(01|t_{hh},\theta_1)\right] \leq (1-t_h)\left[p\sigma(1|t_{\ell h},\theta_1) + (1-p)\sigma(1|t_{hh},\theta_1)\right] + t_h \left[p\sigma(10|t_{\ell h},\theta_0) + (1-p)\sigma(01|t_{hh},\theta_0)\right] \quad \text{(A.9)}$$

$$t_h \left[p\sigma(1|t_{\ell h},\theta_0) + (1-p)\sigma(1|t_{hh},\theta_0)\right] + (1-t_h)\left[p\sigma(1|t_{\ell\ell},\theta_1) + (1-p)\sigma(1|t_{\ell h},\theta_1)\right] \leq t_h \left[p\sigma(1|t_{\ell\ell},\theta_0) + (1-p)\sigma(1|t_{\ell h},\theta_0)\right] + (1-t_h)\left[p\sigma(1|t_{\ell h},\theta_1) + (1-p)\sigma(1|t_{hh},\theta_1)\right] \quad \text{(A.10)}$$

$$t_h\left[p\sigma(1|t_{\ell h},\theta_0)+(1-p)\sigma(1|t_{hh},\theta_0)\right]+(1-t_h)\left[p\sigma(1|t_{\ell\ell},\theta_1)+(1-p)\sigma(1|t_{\ell h},\theta_1)\right]+$$
$$(1-t_h)\left[p\sigma(10|t_{\ell\ell},\theta_1)+(1-p)\sigma(01|t_{\ell h},\theta_1)\right]\leq$$
$$t_h\left[p\sigma(1|t_{\ell\ell},\theta_0)+(1-p)\sigma(1|t_{\ell h},\theta_0)\right]+(1-t_h)\left[p\sigma(1|t_{\ell h},\theta_1)+(1-p)\sigma(1|t_{hh},\theta_1)\right]+$$
$$t_h\left[p\sigma(10|t_{\ell\ell},\theta_0)+(1-p)\sigma(01|t_{\ell h},\theta_0)\right] \quad \text{(A.11)}$$

$$t_h\left[p\sigma(1|t_{\ell h},\theta_0)+(1-p)\sigma(1|t_{hh},\theta_0)\right]+(1-t_h)\left[p\sigma(10|t_{\ell\ell},\theta_1)+(1-p)\sigma(01|t_{\ell h},\theta_1)\right]\leq$$
$$t_h\left[p\sigma(01|t_{\ell\ell},\theta_0)+(1-p)\sigma(01|t_{\ell h},\theta_0)\right]+(1-t_h)\left[p\sigma(1|t_{\ell h},\theta_1)+(1-p)\sigma(1|t_{hh},\theta_1)\right], \quad \text{(A.12)}$$

where (A.7) is the obedience constraint against deviations to $\delta_i(a_i)=0$ for any $a_i\in A_i$, (A.8) is the obedience constraint against deviations to $\delta_i(a_i)=1$ for any $a_i\in A_i$, (A.9) is the obedience constraint against deviations to $\delta_i(1)=0$ and $\delta_i(0)=1$, (A.10) is the truth-telling constraint, (A.11) is the truth-telling constraint against deviations to $\delta_i(a_i)=1$ for any $a_i\in A_i$, and (A.12) is the truth-telling constraint against deviations to $\delta_i(1)=0$ and $\delta_i(0)=1$. The truth-telling constraint against deviations to $\delta_i(a_i)=0$ for any $a_i\in A_i$ coincides with (A.7).

The full problem is to maximize (4.14) subject to the whole set of constraints (A.1)-(A.12). It is immediate that (A.3) is implied by (A.2) and (A.1); and that (A.9) is implied by (A.8) and (A.7). Consider now a reduced problem where we maximize the same function over the same set of constraints, but with the restriction that $\sigma(10|t,\theta_1)=\sigma(01|t,\theta_1)=0$ for every $t\in T$. It is then straightforward to verify that, for any solution $\sigma$ of the original problem, there exists an outcome-equivalent solution $\sigma'$ of the reduced problem, obtained from $\sigma$ by shifting for each $t$ the probability mass on $\sigma(10|t,\theta_1)$ and $\sigma(01|t,\theta_1)$ to $\sigma(00|t,\theta_1)$. Conversely, if a mechanism solves the simplified problem, then it is also a solution of the original problem. Thus it is without loss of generality to confine our attention to the reduced problem. This implies that constraints (A.2), (A.5), (A.6), (A.8), (A.11), and (A.12) are redundant. Finally, to show that also (A.1) is redundant, notice that using the assumption $t_\ell<t_h$ in constraint (A.7) yields

$$t_\ell\left[p\sigma(1|t_{\ell h},\theta_0)+(1-p)\sigma(1|t_{hh},\theta_0)\right]\leq(1-t_\ell)\left[p\sigma(1|t_{\ell h},\theta_1)+(1-p)\sigma(1|t_{hh},\theta_1)\right].$$

Using the latter in (A.4), one obtains (A.1). Therefore, we are left with constraints (A.4),

(A.7), and (A.10), which are the three constraints mentioned in the main text.

## A.2 Proof of Proposition 10

We start by observing that Problem (P-1) has a linear objective function and a set of linear constraints, irrespective of whether $k = 1$ or $k = 2$. Therefore, Karush-Kuhn-Tucker conditions are both necessary and sufficient. The same applies to Problem (P-2).

The set of probability constraints for Problem (P-1) when $k = 2$ is

$$\begin{aligned} (\mu_1) &\rightarrow \ \sigma(1|t_{\ell\ell}, \theta_0) \leq 1; & (\mu_4) &\rightarrow \sigma(1|t_{\ell\ell}, \theta_1) \leq 1; \\ (\mu_2) &\rightarrow \ \sigma(1|t_{\ell h}, \theta_0) \leq 1; & (\mu_5) &\rightarrow \sigma(1|t_{\ell h}, \theta_1) \leq 1; \\ (\mu_3) &\rightarrow \ \sigma(1|t_{hh}, \theta_0) \leq 1; & (\mu_6) &\rightarrow \sigma(1|t_{hh}, \theta_1) \leq 1, \end{aligned}$$

where $\mu_i$'s are the associated Lagrange multipliers.

Let $\lambda_1$, $\lambda_2$, and $\lambda_3$ be the Lagrange multipliers corresponding to incentive constraints (4.16), (4.15) and (4.17), respectively. The set of first-order conditions with respect to the six control variables is:

$$\frac{p^2}{2} - \mu_1 - pt_\ell\lambda_1 + pt_h\lambda_3 \leq 0, \tag{A.13}$$

$$p(1-p) - \mu_2 + (2p-1)t_\ell\lambda_1 - pt_h\lambda_2 + (1-2p)t_h\lambda_3 \leq 0, \tag{A.14}$$

$$\frac{(1-p)^2}{2} - \mu_3 + (1-p)t_\ell\lambda_1 - (1-p)t_h\lambda_2 - (1-p)t_h\lambda_3 \leq 0, \tag{A.15}$$

$$\frac{p^2}{2} - \mu_4 + p(1-t_\ell)\lambda_1 - p(1-t_h)\lambda_3 \leq 0, \tag{A.16}$$

$$p(1-p) - \mu_5 + (1-2p)(1-t_\ell)\lambda_1 + p(1-t_h)\lambda_2 + (2p-1)(1-t_h)\lambda_3 \leq 0, \tag{A.17}$$

$$\frac{(1-p)^2}{2} - \mu_6 - (1-p)(1-t_\ell)\lambda_1 + (1-p)(1-t_h)\lambda_2 + (1-p)(1-t_h)\lambda_3 \leq 0, \tag{A.18}$$

where each condition holds with equality if the corresponding variable is strictly positive. We now show what are the conditions under which mechanisms in $\Sigma^A$, $\Sigma^B$, and $\Sigma^C$ are solutions to problem (P-1).

- **$\Sigma^A$ is a solution.** Suppose the optimal mechanism belongs to $\Sigma^A$. It follows by substitution that (4.15) and (4.16) are binding. Furthermore, $t_\ell < t_h$ implies that (4.17) is slack. By complementary slackness, we have $\mu_1 = \mu_2 = \mu_3 = \mu_5 = \mu_6 = 0$, and $\lambda_3 = 0$. From first-order conditions (A.13)-(A.16) we easily get $\lambda_1 = \frac{p}{2t_\ell} > 0$,

$\lambda_2 \geq \frac{1}{2t_h} > 0$ and $\mu_4 = \frac{p^2}{2t_\ell} > 0$. FOC (A.17) is equivalent to

$$\lambda_2 \leq \frac{2p - 1 - t_\ell}{2t_\ell(1 - t_h)}. \tag{A.19}$$

The right-hand side of (A.19) is greater than $\frac{1}{2t_h}$ if and only if $p \geq \frac{t_\ell + t_h}{2t_h}$. Finally, FOC (A.18) holds if and only if

$$\lambda_2 \leq \frac{p - t_\ell}{2t_\ell(1 - t_h)}. \tag{A.20}$$

Since it is always the case that $\frac{2p-1-t_\ell}{2t_\ell(1-t_h)} \leq \frac{p-t_\ell}{2t_\ell(1-t_h)}$, we can conclude that any mechanism in $\Sigma^A$ is a solution provided that $p \geq \frac{t_\ell + t_h}{2t_h}$.

- $\Sigma^B$ **is a solution.** Suppose the optimal mechanism is in $\Sigma^B$. First, the intersection

$$(0,1) \cap \left( \frac{1 - t_h}{t_h} - \frac{1 - p}{p}, \frac{1 - t_h}{t_h} \right) \cap \left( \frac{1 - t_\ell}{t_\ell} + \frac{(1 - t_h)p}{t_h(1 - p)} - \frac{p}{1 - p}, \frac{1 - t_\ell}{t_\ell} + \frac{(1 - t_h)p}{t_h(1 - p)} \right)$$

is nonempty if and only if

$$p > \frac{t_h - t_\ell}{t_h + 2t_\ell t_h - 2t_\ell}. \tag{A.21}$$

Suppose that $p$ satisfies (A.21). Then $\sigma^B(1|t_{\ell h}, \theta_0)$, $\sigma^B(1|t_{hh}, \theta_0)$, and $\sigma^B(1|t_{\ell\ell}, \theta_0)$ are all strictly between 0 and 1. Thus we have that $\mu_1 = \mu_2 = \mu_3 = \mu_6 = 0$. Condition (A.13) and $\mu_1 = 0$ imply $\lambda_1 > 0$. Summing side by side FOC's (A.13)-(A.15) one gets $\lambda_2 > 0$. In addition, combining the binding constraints (4.16) and (4.15), it follows that (4.17) is slack, so implying $\lambda_3 = 0$. From FOC (A.13) we obtain $\lambda_1 = \frac{p}{2t_\ell}$, while FOC (A.14) implies $\lambda_2 = \frac{1}{2t_h}$. From (A.16) we obtain $\mu_4 = \frac{p^2}{2t_\ell} > 0$. From (A.17),

$$\mu_5 = \frac{p(1 - 2p + t_\ell)}{2t_\ell} + \frac{p(1 - t_h)}{2t_h},$$

which is non-negative if and only if $p \leq \frac{t_h + t_\ell}{2t_h}$. Finally, using all of the above, we have that (A.18) is satisfied if and only if $p \geq \frac{t_\ell}{t_h}$. Since $t_h > t_\ell$ by assumption, $\frac{t_\ell}{t_h} < \frac{t_h + t_\ell}{2t_h}$ is always true. Finally, notice that

$$\frac{t_\ell}{t_h} > \frac{t_h - t_\ell}{t_h + 2t_\ell t_h - 2t_\ell}$$

if and only if $2t_\ell(t_h - t_\ell) > t_h(t_h - 2t_\ell^2)$, which is always true since $\frac{1}{2} < t_\ell < t_h$ by assumption. Therefore, we can conclude that every mechanism in $\Sigma^B$ is a solution if

$$\frac{t_\ell}{t_h} \leq p \leq \frac{t_h + t_\ell}{2t_h}.$$

- $\Sigma^C$ **is a solution.** Now suppose that the optimal mechanism belongs to $\Sigma^C$. First, the intersection

$$(0,1) \cap \left( \frac{1-t_h}{pt_h} - \frac{1-p}{p}, \frac{1-t_h}{pt_h} \right) \cap \left( -\frac{(1-t_h)(1-2p)}{p^2 t_h}, \frac{(1-p)^2}{p^2} - \frac{(1-t_h)(1-2p)}{p^2 t_h} \right)$$

is always non-empty. This implies that $\sigma^C(1|t_{\ell\ell}, \theta_0)$, $\sigma^C(1|t_{\ell h}, \theta_0)$, and $\sigma^C(1|t_{hh}, \theta_0)$ are all strictly between 0 and 1. It follows that $\mu_1 = \mu_2 = \mu_3 = 0$, and that all three incentive constraints are binding. From (A.13)-(A.15), it is immediate to obtain $\lambda_1 = \lambda_3 \frac{t_h}{t_\ell} + \frac{p}{2t_\ell} > 0$ and $\lambda_2 = \frac{1}{2t_h} > 0$. From (A.18), $\mu_6 \geq 0$ if and only if $\lambda_3 \leq \frac{t_\ell - pt_h}{2t_h(1-t_\ell)}$. To make sure that $\lambda_3 \geq 0$, it must be the case that $p \leq \frac{t_\ell}{t_h}$. It remains to check FOC (A.18). If $p \leq \frac{1}{2}$, (A.18) is satisfied for any $\lambda_3 \geq 0$. If $p \in \left(\frac{1}{2}, \frac{t_\ell}{t_h}\right]$, then it must be the case that

$$\lambda_3 \leq \frac{pt_h + pt_\ell - 2p^2 t_h}{4pt_h^2 - 2t_h^2 + 2t_\ell t_h - 4pt_\ell t_h}. \tag{A.22}$$

The right hand side of of (A.22) is non-negative due to the fact that $p \in \left(\frac{1}{2}, \frac{t_\ell}{t_h}\right]$. Therefore, $\lambda_3$ satisfies

$$0 \leq \lambda_3 \leq \min \left\{ \frac{t_\ell - pt_h}{2t_h(1-t_\ell)}, \frac{pt_h + pt_\ell - 2p^2 t_h}{4pt_h^2 - 2t_h^2 + 2t_\ell t_h - 4pt_\ell t_h} \right\}.$$

Thus any mechanism in $\Sigma^C$ is a solution provided that $p \leq \frac{t_\ell}{t_h}$.

- Finally, suppose $p = \frac{t_\ell + t_h}{2t_h}$. We have established that mechanisms in $\Sigma^A$ and $\Sigma^B$ belong to the solution set. Now consider any convex combination $\sigma = x\sigma^A + (1-x)\sigma^B$, where $x \in [0,1]$. Since the feasible set is convex, $\sigma$ is feasible. Furthermore, it follows immediately from the linearity of the objective function that $U_S(\sigma) = U_S(\sigma^A) = U_S(\sigma^B)$. This shows that $\sigma$ is a solution for every $x \in [0,1]$. The same line of reasoning applies when $p = \frac{t_\ell}{t_h}$, so ending the proof.

## A.3 Incentive constraints for the case with information elicitation and $k = 1$

The whole set of BCE incentive compatibility constraints for type $t_\ell$ is:

$$t_\ell \left[ p\sigma(10|t_{\ell\ell}, \theta_0) + (1-p)\sigma(10|t_{\ell h}, \theta_0) \right] \leq (1-t_\ell) \left[ p\sigma(10|t_{\ell\ell}, \theta_1) + (1-p)\sigma(10|t_{\ell h}, \theta_1) \right] \tag{A.23}$$

$$(1-t_\ell)+t_\ell\left[2p\sigma(10|t_{\ell\ell},\theta_0)+(1-p)\left(\sigma(10|t_{\ell h},\theta_0)+\sigma(01|t_{\ell h},\theta_0)\right)\right]\leq$$
$$t_\ell+(1-t_\ell)\left[2p\sigma(10|t_{\ell\ell},\theta_1)+(1-p)\left(\sigma(10|t_{\ell h},\theta_1)+\sigma(01|t_{\ell h},\theta_1)\right)\right] \tag{A.24}$$

$$\begin{aligned}&t_\ell\left[2p\sigma(10|t_{\ell\ell},\theta_0)+(1-p)\left(\sigma(10|t_{\ell h},\theta_0)+\sigma(01|t_{\ell h},\theta_0)\right)\right]+\\&(1-t_\ell)\left[p\left(\sigma(10|t_{\ell h},\theta_1)+\sigma(01|t_{\ell h},\theta_1)\right)+2(1-p)\sigma(10|t_{hh},\theta_1)\right]\leq\\&t_\ell\left[p\left(\sigma(10|t_{\ell h},\theta_0)+\sigma(01|t_{\ell h},\theta_0)\right)+2(1-p)\sigma(10|t_{hh},\theta_0)\right]+\\&(1-t_\ell)\left[2p\sigma(10|t_{\ell\ell},\theta_1)+(1-p)\left(\sigma(10|t_{\ell h},\theta_1)+\sigma(01|t_{\ell h},\theta_1)\right)\right]\end{aligned} \tag{A.25}$$

$$\begin{aligned}&t_\ell\left[2p\sigma(10|t_{\ell\ell},\theta_0)+(1-p)\left(\sigma(10|t_{\ell h},\theta_0)+\sigma(01|t_{\ell h},\theta_0)\right)\right]+\\&(1-t_\ell)\left[p\sigma(10|t_{\ell h},\theta_1)+(1-p)\sigma(10|t_{hh},\theta_1)\right]\leq\\&t_\ell\left[p\sigma(10|t_{\ell h},\theta_0)+(1-p)\sigma(10|t_{hh},\theta_0)\right]+\\&(1-t_\ell)\left[2p\sigma(10|t_{\ell\ell},\theta_1)+(1-p)\left(\sigma(10|t_{\ell h},\theta_1)+\sigma(01|t_{\ell h},\theta_1)\right)\right]\end{aligned} \tag{A.26}$$

$$\begin{aligned}&(1-t_\ell)+t_\ell\left[2p\sigma(10|t_{\ell\ell},\theta_0)+(1-p)\left(\sigma(10|t_{\ell h},\theta_0)+\sigma(01|t_{\ell h},\theta_0)\right)\right]+\\&t_\ell\left[p\sigma(01|t_{\ell h},\theta_0)+(1-p)\sigma(10|t_{hh},\theta_0)\right]\leq\\&t_\ell+(1-t_\ell)\left[p\sigma(01|t_{\ell h},\theta_1)+(1-p)\sigma(10|t_{hh},\theta_1)\right]+\\&(1-t_\ell)\left[2p\sigma(10|t_{\ell\ell},\theta_1)+(1-p)\left(\sigma(10|t_{\ell h},\theta_1)+\sigma(01|t_{\ell h},\theta_1)\right)\right],\end{aligned} \tag{A.27}$$

where (A.23) is the obedience constraint against deviations to $\delta_i(a_i)=0$ for any $a_i\in A_i$, (A.24) is the obedience constraint against deviations to $\delta_i(a_i)=1$ for any $a_i\in A_i$, (A.25) is the truth-telling constraint, (A.26) is the truth-telling constraint against deviations to $\delta_i(a_i)=0$ for any $a_i\in A_i$, and (A.27) is the truth-telling constraint against deviations to $\delta_i(1)=0$ and $\delta_i(0)=1$. The truth-telling constraint against deviations to $\delta_i(a_i)=1$ for any $a_i\in A_i$ coincides with (A.24). The obedience constraint against deviations to $\delta_i(1)=0$ and $\delta_i(0)=1$ is implied by (A.23) and (A.24).

The set of incentive constraints for type $t_h$ is:

$$t_h\left[p\sigma(01|t_{\ell h},\theta_0)+(1-p)\sigma(10|t_{hh},\theta_0)\right]\leq(1-t_h)\left[p\sigma(01|t_{\ell h},\theta_1)+(1-p)\sigma(10|t_{hh},\theta_1)\right] \tag{A.28}$$

$$\begin{aligned}(1-t_h)+t_h\left[p\left(\sigma(10|t_{\ell h},\theta_0)+\sigma(01|t_{\ell h},\theta_0)\right)+2(1-p)\sigma(10|t_{hh},\theta_0)\right]\leq\\ t_h+(1-t_h)\left[p\left(\sigma(10|t_{\ell h},\theta_1)+\sigma(01|t_{\ell h},\theta_1)\right)+2(1-p)\sigma(10|t_{hh},\theta_1)\right]\end{aligned} \tag{A.29}$$

$$\begin{aligned}t_h\left[p\left(\sigma(10|t_{\ell h},\theta_0)+\sigma(01|t_{\ell h},\theta_0)\right)+2(1-p)\sigma(10|t_{hh},\theta_0)\right]+\\ (1-t_h)\left[2p\sigma(10|t_{\ell\ell},\theta_1)+(1-p)\left(\sigma(10|t_{\ell h},\theta_1)+\sigma(01|t_{\ell h},\theta_1)\right)\right]\leq\\ t_h\left[2p\sigma(10|t_{\ell\ell},\theta_0)+(1-p)\left(\sigma(10|t_{\ell h},\theta_0)+\sigma(01|t_{\ell h},\theta_0)\right)\right]+\\ (1-t_h)\left[p\left(\sigma(10|t_{\ell h},\theta_1)+\sigma(01|t_{\ell h},\theta_1)\right)+2(1-p)\sigma(10|t_{hh},\theta_1)\right]\end{aligned} \tag{A.30}$$

$$\begin{aligned}t_h\left[p\left(\sigma(10|t_{\ell h},\theta_0)+\sigma(01|t_{\ell h},\theta_0)\right)+2(1-p)\sigma(10|t_{hh},\theta_0)\right]+\\ (1-t_h)\left[p\sigma(10|t_{\ell\ell},\theta_1)+(1-p)\sigma(01|t_{\ell h},\theta_1)\right]\leq\\ t_h\left[p\sigma(10|t_{\ell\ell},\theta_0)+(1-p)\sigma(01|t_{\ell h},\theta_0)\right]+\\ (1-t_h)\left[p\left(\sigma(10|t_{\ell h},\theta_1)+\sigma(01|t_{\ell h},\theta_1)\right)+2(1-p)\sigma(10|t_{hh},\theta_1)\right]\end{aligned} \tag{A.31}$$

$$\begin{aligned}(1-t_h)+t_h\left[p\left(\sigma(10|t_{\ell h},\theta_0)+\sigma(01|t_{\ell h},\theta_0)\right)+2(1-p)\sigma(10|t_{hh},\theta_0)\right]+\\ t_h\left[p\sigma(10|t_{\ell\ell},\theta_0)+(1-p)\sigma(10|t_{\ell h},\theta_0)\right]\leq\\ t_h+(1-t_h)\left[p\sigma(10|t_{\ell\ell},\theta_1)+(1-p)\sigma(10|t_{\ell h},\theta_1)\right]+\\ (1-t_h)\left[p\left(\sigma(10|t_{\ell h},\theta_1)+\sigma(01|t_{\ell h},\theta_1)\right)+2(1-p)\sigma(10|t_{hh},\theta_1)\right],\end{aligned} \tag{A.32}$$

where (A.28) is the obedience constraint against deviations to $\delta_i(a_i)=0$ for any $a_i\in A_i$, (A.29) is the obedience constraint against deviations to $\delta_i(a_i)=1$ for any $a_i\in A_i$, (A.30) is the truth-telling constraint, (A.31) is the truth-telling constraint against deviations to $\delta_i(a_i)=0$ for any $a_i\in A_i$, and (A.32) is the truth-telling constraint against deviations to $\delta_i(1)=0$ and $\delta_i(0)=1$. Similarly to the low type, the truth-telling constraint against deviations to $\delta_i(a_i)=1$ for any $a_i\in A_i$ coincides with (A.29), and the obedience constraint

against deviations to $\delta_i(1) = 0$ and $\delta_i(0) = 1$ is implied by (A.28) and (A.29).

Now we argue that constraints (A.26) and (A.27) are redundant. From constraint (A.28) and the assumption $t_\ell < t_h$ we easily obtain

$$t_\ell \left[p\sigma(01|t_{\ell h}, \theta_0) + (1-p)\sigma(10|t_{hh}, \theta_0)\right] \leq (1-t_\ell)\left[p\sigma(01|t_{\ell h}, \theta_1) + (1-p)\sigma(10|t_{hh}, \theta_1)\right]. \tag{A.33}$$

Combining (A.33) and (A.25) gives (A.26); combining (A.33) and (A.24) gives (A.27).

To find the Sender's optimal persuasion mechanism, we solve a relaxed problem where we ignore constraints (A.24), (A.29), and (A.32). It is tedious but straightforward to verify that, at every solution we are going to find in Proposition 11, constraints (A.24), (A.29), and (A.32) are all satisfied.

## A.4 Proof of Proposition 11

The set of probability constraints for Problem (P-1) when $k = 1$ is

$$\begin{aligned}
(\mu_1) \rightarrow\ & \sigma(10|t_{\ell\ell}, \theta_0) \leq \frac{1}{2}; & (\mu_2) \rightarrow\ & \sigma(10|t_{\ell\ell}, \theta_1) \leq \frac{1}{2};\\
(\mu_3) \rightarrow\ & \sigma(10|t_{\ell h}, \theta_0) + \sigma(01|t_{\ell h}, \theta_0) \leq 1; & (\mu_4) \rightarrow\ & \sigma(10|t_{\ell h}, \theta_1) + \sigma(01|t_{\ell h}, \theta_1) \leq 1;\\
(\mu_5) \rightarrow\ & \sigma(10|t_{hh}, \theta_0) \leq \frac{1}{2}; & (\mu_6) \rightarrow\ & \sigma(10|t_{hh}, \theta_1) \leq \frac{1}{2},
\end{aligned}$$

where $\mu_i$'s are the associated Lagrange multipliers.

Let $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$, and $\lambda_5$ be the Lagrange multipliers corresponding to constraints (4.19), (4.20), (4.21), (4.22), and (4.23), respectively. The set of first-order conditions with respect to the eight control variables is:

$$p^2 - \mu_1 - pt_\ell\lambda_1 - 2pt_\ell\lambda_2 + 2pt_h\lambda_4 + pt_h\lambda_5 \leq 0, \tag{A.34}$$

$$p(1-p) - \mu_3 + (2p-1)t_\ell\lambda_2 - pt_h\lambda_3 + (1-2p)t_h(\lambda_4 + \lambda_5) \leq 0, \tag{A.35}$$

$$p(1-p) - \mu_3 - (1-p)t_\ell\lambda_1 + (2p-1)t_\ell\lambda_2 + (1-2p)t_h\lambda_4 - pt_h\lambda_5 \leq 0, \tag{A.36}$$

$$(1-p)^2 - \mu_5 + 2(1-p)t_\ell\lambda_2 - (1-p)t_h\lambda_3 - 2(1-p)t_h(\lambda_4 + \lambda_5) \leq 0, \tag{A.37}$$

$$p^2 - \mu_2 + p(1-t_\ell)\lambda_1 + 2p(1-t_\ell)\lambda_2 - 2p(1-t_h)\lambda_4 - p(1-t_h)\lambda_5 \leq 0, \tag{A.38}$$

$$p(1-p) - \mu_4 + (1-2p)(1-t_\ell)\lambda_2 + p(1-t_h)\lambda_3 + (2p-1)(1-t_h)(\lambda_4 + \lambda_5) \leq 0, \tag{A.39}$$

$$p(1-p) - \mu_4 + (1-p)(1-t_\ell)\lambda_1 + (1-2p)(1-t_\ell)\lambda_2 + p(1-t_h)\lambda_5 + (2p-1)(1-t_h)\lambda_4 \leq 0, \tag{A.40}$$

$$(1-p)^2 - \mu_6 - 2(1-p)(1-t_\ell)\lambda_2 + (1-p)(1-t_h)\lambda_3 + 2(1-p)(1-t_h)(\lambda_4+\lambda_5) \leq 0, \quad \text{(A.41)}$$

where each condition holds with equality if the corresponding variable is strictly positive. We now show what are the conditions under which mechanisms in $\Sigma^D$, $\Sigma^E$, and $\Sigma^F$ are solutions to problem (P-1).

- **$\Sigma^D$ is a solution.** Suppose the optimal mechanism belongs to $\Sigma^D$. It follows immediately that $\mu_1 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = 0$. Furthermore, constraints (4.19), (4.20), (4.21) and (4.23) are binding, whereas (4.22) is slack, so implying $\lambda_4 = 0$. From FOC (A.34), (A.36) and (A.40), one gets

$$\lambda_1 = \frac{1}{t_\ell} - \frac{1-p}{(t_h - t_\ell)(2p-1)} \quad \text{(A.42)}$$

and

$$\lambda_2 = \frac{p(1-p)}{(t_h - t_\ell)(2p-1)}. \quad \text{(A.43)}$$

Both $\lambda_1$ and $\lambda_2$ are non-negative if and only if $p \geq \frac{t_h}{2t_h - t_\ell}$. In addition, $\lambda_5 = \frac{1-p}{t_h - t_\ell} > 0$. Using the above in (A.38), $\mu_2 = \frac{1}{t_\ell} > 0$. FOC (A.35) is satisfied as long as $\lambda_3 \geq \frac{(1-p)^2}{p(t_h - t_\ell)} > 0$. The fact that $p \geq \frac{t_h}{2t_h - t_\ell}$ implies that also FOC (A.37) is satisfied. Finally, combining the above with FOC (A.39) and (A.41) one gets $\lambda_3 = \frac{(1-p)^2}{p(t_h - t_\ell)}$. Thus any mechanism in $\Sigma^D$ is solution provided that $p \geq \frac{t_h}{2t_h - t_\ell}$.

- **$\Sigma^E$ is a solution.** To show that $\Sigma^E$ is non-empty, notice that (4.26) and (4.28) can hold simultaneously if and only if

$$\frac{pt_h}{2(1-p)(1-t_h)} - \frac{p(2p-1)}{2(1-p)} - \frac{pt_h(1-t_\ell)}{t_\ell(1-t_h)} > 0,$$

which is equivalent to

$$1 - \frac{(2p-1)(1-t_h)}{t_h} - \frac{2(1-p)(1-t_\ell)}{t_\ell} > 0. \quad \text{(A.44)}$$

Now, the left-hand side of (A.44) is increasing with respect to $p$. Since we are looking for solutions when $p \geq \frac{t_\ell}{2t_h - t_\ell}$, and since $\frac{t_\ell}{2t_h - t_\ell} > \frac{1}{3}$, we can write

$$1 - \frac{(2p-1)(1-t_h)}{t_h} - \frac{2(1-p)(1-t_\ell)}{t_\ell} > 1 + \frac{(1-t_h)}{3t_h} - \frac{4(1-t_\ell)}{3t_\ell} \geq 0, \quad \text{(A.45)}$$

where the last inequality follows from the assumption $\frac{1-t_h}{t_h} - \frac{4(1-t_\ell)}{t_\ell} \geq -3$. It is then straightforward to verify that all remaining inequalities in the Definition of $\Sigma^E$ can

hold simultaneously. In addition, one can verify that any mechanism in $\Sigma^E$ is such that every control variable, except for $\sigma^E(10|t_{hh},\theta_1)$, is strictly positive, and that the three probability constraints for state $\theta_0$ are all slack. Therefore, if an optimal mechanism is in $\Sigma^E$, it follows immediately that $\mu_1 = \mu_3 = \mu_5 = \mu_6 = 0$. Furthermore, incentive constraints (4.20), (4.21) and (4.23) are all binding while (4.19) and (4.22) are slack, so implying that $\lambda_1 = \lambda_4 = 0$. Furthermore, from FOC (A.34)-(A.36) we easily obtain $\lambda_2 = \frac{p}{t_\ell} > 0$, $\lambda_3 = \frac{1-p}{t_h} > 0$, and $\lambda_5 = \frac{p}{t_h} > 0$. FOC (A.37) always holds with equality. From FOC (A.38) we can easily obtain $\mu_2$, which is always strictly positive. From (A.40) we can easily solve for $\mu_4$, which is non-negative if and only if $p \leq \frac{t_h}{2t_h - t_\ell}$. In addition, FOC (A.39) always holds with equality. Finally, (A.41) is satisfied as long as $p \geq \frac{t_\ell}{2t_h - t_\ell}$.

- $\Sigma^F$ **is a solution.** Suppose the optimal mechanism belongs to $\Sigma^F$. It follows immediately that $\mu_1 = \mu_3 = \mu_5 = 0$. Furthermore, incentive constraints (4.20), (4.21), (4.22), and (4.23) are binding, whereas (4.19) is slack, so implying $\lambda_1 = 0$. From FOC (A.34)-(A.36) we easily obtain $\lambda_2 = \frac{p}{t_\ell} + \frac{t_h}{t_\ell}\lambda_4 > 0$, $\lambda_3 = \frac{1-p}{t_h} > 0$, and $\lambda_5 = \frac{p}{t_h} > 0$. FOC (A.37) always holds with equality. From (A.38) we can easily obtain $\mu_2$, which is non-negative for any $\lambda_4 \geq 0$. FOCs (A.39) and (A.40) turn out to be equivalent. From either of them we can easily obtain the expression for $\mu_4$. If $p \leq \frac{1}{2}$, then $\mu_4 > 0$ for every $\lambda_4 \leq 0$. If $p > \frac{1}{2}$, then $\mu_4$ is non-negative if and only if

$$\lambda_4 \leq \frac{t_\ell}{(2p-1)(t_h - t_\ell)} \left[ p(1-p) + \frac{p(1-2p)(1-t_\ell)}{t_\ell} + \frac{p^2(1-t_h)}{t_h} \right]. \tag{A.46}$$

  The right-hand side of (A.46) is non-negative if and only if $p \leq \frac{t_h}{2t_h - t_\ell}$. Notice that $\frac{t_h}{2t_h - t_\ell} > \frac{1}{2}$. Finally, from FOC (A.41) we can easily obtain $\mu_6$, which is non-negative if and only if

$$\lambda_4 \leq \frac{t_\ell}{2(1-p)(t_h - t_\ell)} \left[ (1-p)^2 - \frac{2p(1-p)(1-t_\ell)}{t_\ell} + \frac{(1-p^2)(1-t_h)}{t_h} \right]. \tag{A.47}$$

  The right-hand side of (A.47) is non-negative if and only if $p \leq \frac{t_\ell}{2t_h - t_\ell}$. Therefore, we can conclude that every mechanism is $\Sigma^F$ is a solution provided that $p \leq \frac{t_\ell}{2t_h - t_\ell}$.

- When $p = \frac{t_\ell}{2t_h - t_\ell}$ or $p = \frac{t_h}{2t_h - t_\ell}$, the same argument as in the previous proposition applies, so ending the proof.

## A.5 Incentive constraints for the case without information elicitation and $k = 2$

The full set of obedience constraints for player 1 is the following:

$$\begin{aligned}(1-t_\ell)\left[\phi(b_{00},b_{11}|\theta_1)+p\phi(b_{00},b_{10}|\theta_1)+(1-p)\phi(b_{00},b_{01}|\theta_1)\right]\leq\\ t_\ell\left[\phi(b_{00},b_{11}|\theta_0)+p\phi(b_{00},b_{10}|\theta_0)+(1-p)\phi(b_{00},b_{01}|\theta_0)\right]\end{aligned}\qquad(t_\ell,b_{00})$$

$$\begin{aligned}(1-t_h)\left[\phi(b_{00},b_{11}|\theta_1)+p\phi(b_{00},b_{10}|\theta_1)+(1-p)\phi(b_{00},b_{01}|\theta_1)\right]\leq\\ t_h\left[\phi(b_{00},b_{11}|\theta_0)+p\phi(b_{00},b_{10}|\theta_0)+(1-p)\phi(b_{00},b_{01}|\theta_0)\right]\end{aligned}\qquad(t_h,b_{00})$$

$$\begin{aligned}(1-t_\ell)\left[\phi(b_{01},b_{11}|\theta_1)+p\phi(b_{01},b_{10}|\theta_1)+(1-p)\phi(b_{01},b_{01}|\theta_1)\right]\leq\\ t_\ell\left[\phi(b_{01},b_{11}|\theta_0)+p\phi(b_{01},b_{10}|\theta_0)+(1-p)\phi(b_{01},b_{01}|\theta_0)\right]\end{aligned}\qquad(t_\ell,b_{01})$$

$$\begin{aligned}t_h\left[\phi(b_{01},b_{11}|\theta_0)+p\phi(b_{01},b_{10}|\theta_0)+(1-p)\phi(b_{01},b_{01}|\theta_0)\right]\leq\\ (1-t_h)\left[\phi(b_{01},b_{11}|\theta_1)+p\phi(b_{01},b_{10}|\theta_1)+(1-p)\phi(b_{01},b_{01}|\theta_1)\right]\end{aligned}\qquad(t_h,b_{01})$$

$$\begin{aligned}t_\ell\left[\phi(b_{10},b_{11}|\theta_0)+p\phi(b_{10},b_{10}|\theta_0)+(1-p)\phi(b_{10},b_{01}|\theta_0)\right]\leq\\ (1-t_\ell)\left[\phi(b_{10},b_{11}|\theta_1)+p\phi(b_{10},b_{10}|\theta_1)+(1-p)\phi(b_{10},b_{01}|\theta_1)\right]\end{aligned}\qquad(t_\ell,b_{10})$$

$$\begin{aligned}(1-t_h)\left[\phi(b_{10},b_{11}|\theta_1)+p\phi(b_{10},b_{10}|\theta_1)+(1-p)\phi(b_{10},b_{01}|\theta_1)\right]\leq\\ t_h\left[\phi(b_{10},b_{11}|\theta_0)+p\phi(b_{10},b_{10}|\theta_0)+(1-p)\phi(b_{10},b_{01}|\theta_0)\right]\end{aligned}\qquad(t_h,b_{10})$$

$$\begin{aligned}t_\ell\left[\phi(b_{11},b_{11}|\theta_0)+p\phi(b_{11},b_{10}|\theta_0)+(1-p)\phi(b_{11},b_{01}|\theta_0)\right]\leq\\ (1-t_\ell)\left[\phi(b_{11},b_{11}|\theta_1)+p\phi(b_{11},b_{10}|\theta_1)+(1-p)\phi(b_{11},b_{01}|\theta_1)\right]\end{aligned}\qquad(t_\ell,b_{11})$$

$$t_h\left[\phi(b_{11},b_{11}|\theta_0)+p\phi(b_{11},b_{10}|\theta_0)+(1-p)\phi(b_{11},b_{01}|\theta_0)\right]\leq$$
$$(1-t_h)\left[\phi(b_{11},b_{11}|\theta_1)+p\phi(b_{11},b_{10}|\theta_1)+(1-p)\phi(b_{11},b_{01}|\theta_1)\right], \qquad (t_h,b_{11})$$

where the tag $(t,b)$ indicates that the corresponding constraint is for type $t$ to obey her action recommendation contained in $b$.

Now, since $t_\ell < t_h$ by assumption, it is immediate that constraint $(t_\ell, b_{11})$ is implied by $(t_h, b_{11})$, and that $(t_h, b_{00})$ is implied by $(t_\ell, b_{00})$.

Now define

$$\begin{aligned} x :=&\quad \phi(b_{01},b_{11}|\theta_1)+p\phi(b_{01},b_{10}|\theta_1)+(1-p)\phi(b_{01},b_{01}|\theta_1)\\ y :=&\quad \phi(b_{01},b_{11}|\theta_0)+p\phi(b_{01},b_{10}|\theta_0)+(1-p)\phi(b_{01},b_{01}|\theta_0), \end{aligned}$$

so that we can rewrite constraints $(t_\ell, b_{01})$ and $(t_h, b_{01})$ as

$$\begin{aligned} (1-t_\ell)x &\leq t_\ell y\\ t_h y &\leq (1-t_h)x, \end{aligned}$$

respectively. We argue that $x = y = 0$. To show this, let us consider two cases. First, suppose $y = 0$. It is immediate that $x = 0$ by $(t_\ell, b_{01})$. Consequently, both $(t_\ell, b_{01})$ and $(t_h, b_{01})$ hold trivially with equality. Alternatively, suppose $y > 0$. By constraint $(t_h, b_{01})$, we must have $x > 0$ as well. But then we obtain

$$\frac{y}{x} \leq \frac{1-t_h}{t_h}$$

from $(t_h, b_{01})$ and

$$\frac{1-t_\ell}{t_\ell} \leq \frac{y}{x}$$

from $(t_\ell, b_{01})$. These two inequalities cannot hold simultaneously because $t_\ell < t_h$ by assumption. Therefore, we must have $x = y = 0$, which is equivalent to

$$\begin{aligned} &\phi(b_{01},b_{11}|\theta_1)=\phi(b_{01},b_{10}|\theta_1)=\phi(b_{01},b_{01}|\theta_1)=0\\ &\phi(b_{01},b_{11}|\theta_0)=\phi(b_{01},b_{10}|\theta_0)=\phi(b_{01},b_{01}|\theta_0)=0. \end{aligned}$$

Notice that the obedience constraints $(t_\ell, b_{01})$ and $(t_h, b_{01})$ for player 2 read as

$$\begin{aligned}(1-t_\ell)\left[\phi(b_{11},b_{01}|\theta_1)+p\phi(b_{10},b_{01}|\theta_1)+(1-p)\phi(b_{01},b_{01}|\theta_1)\right] \leq \\ t_\ell\left[\phi(b_{11},b_{01}|\theta_0)+p\phi(b_{10},b_{01}|\theta_0)+(1-p)\phi(b_{01},b_{01}|\theta_0)\right] \qquad (t_\ell, b_{01}, \text{ player 2})\end{aligned}$$

and

$$\begin{aligned}t_h\left[\phi(b_{11},b_{01}|\theta_0)+p\phi(b_{10},b_{01}|\theta_0)+(1-p)\phi(b_{01},b_{01}|\theta_0)\right] \leq \\ (1-t_h)\left[\phi(b_{11},b_{01}|\theta_1)+p\phi(b_{10},b_{01}|\theta_1)+(1-p)\phi(b_{01},b_{01}|\theta_1)\right]. \qquad (t_h, b_{01}, \text{ player 2})\end{aligned}$$

By the same argument that we have just used for player 1, we can conclude that

$$\begin{aligned}\phi(b_{11},b_{01}|\theta_1)=\phi(b_{10},b_{01}|\theta_1)=\phi(b_{01},b_{01}|\theta_1)=0 \\ \phi(b_{11},b_{01}|\theta_0)=\phi(b_{10},b_{01}|\theta_0)=\phi(b_{01},b_{01}|\theta_0)=0.\end{aligned}$$

Now we argue that it is without loss of generality to set, for each $\theta \in \Theta$,

$$\phi(b_{00},b_{01}|\theta)=\phi(b_{01},b_{00}|\theta)=0 \tag{A.48}$$

$$\phi(b_{00},b_{10}|\theta)=\phi(b_{10},b_{00}|\theta)=0 \tag{A.49}$$

$$\phi(b_{00},b_{11}|\theta)=\phi(b_{11},b_{00}|\theta)=0. \tag{A.50}$$

To see why this is true, suppose that we are at a solution where some variables in (A.48)-(A.50) are positive. Then it is always feasible to shift all the probability mass from these positive variables to $\phi(b_{00}, b_{00}|\theta)$ in the corresponding state. This shift does not affect the Sender's objective function. In addition, constraint $(t_\ell, b_{00})$ will hold trivially with equality for either player, whereas other incentive constraints will not be affected.

To sum up, the variables that are not necessarily equal to zero are

$$\phi(b_{00},b_{00}|\theta),\ \phi(b_{10},b_{10}|\theta),\ \phi(b_{10},b_{11}|\theta),\ \phi(b_{11},b_{10}|\theta),\ \phi(b_{11},b_{11}|\theta).$$

Since the induced $\sigma$ is symmetric by assumption, we must also have

$$\sigma(1,1|t_{\ell h},\theta)=\sigma(1,1|t_{h\ell},\theta),$$

which simplifies to

$$\phi(b_{10},b_{11}|\theta)+\phi(b_{11},b_{11}|\theta)=\phi(b_{11},b_{10}|\theta)+\phi(b_{11},b_{11}|\theta).$$

Thus we obtain that $\phi(b_{10}, b_{11}|\theta) = \phi(b_{11}, b_{10}|\theta)$.

Using all the above, the incentive constraints for both players are equivalent to each other and simplify to the three inequalities (4.33)-(4.35) in the main text.

Using all these restrictions on $\phi$'s, the Sender's objective function simplifies to:

$$\begin{aligned}
U_S(\phi) = &\frac{1}{2}\Big\{p^2\left[\phi(b_{10}, b_{10}|\theta_0) + 2\phi(b_{10}, b_{11}|\theta_0) + \phi(b_{11}, b_{11}|\theta_0)\right] \\
&+ 2p(1-p)\left[\phi(b_{10}, b_{11}|\theta_0) + \phi(b_{11}, b_{11}|\theta_0)\right] \\
&+ (1-p)^2\left[\phi(b_{11}, b_{11}|\theta_0)\right]\Big\} \\
&+ \frac{1}{2}\Big\{p^2\left[\phi(b_{10}, b_{10}|\theta_1) + 2\phi(b_{10}, b_{11}|\theta_1) + \phi(b_{11}, b_{11}|\theta_1)\right] \\
&+ 2p(1-p)\left[\phi(b_{10}, b_{11}|\theta_1) + \phi(b_{11}, b_{11}|\theta_1)\right] \\
&+ (1-p)^2\left[\phi(b_{11}, b_{11}|\theta_1)\right]\Big\} \\
= &\frac{1}{2}\sum_{\theta\in\Theta}\left[\phi(b_{11}, b_{11}|\theta) + p^2\phi(b_{10}, b_{10}|\theta) + 2p\phi(b_{10}, b_{11}|\theta)\right],
\end{aligned}$$

which is expression (4.32) in the main text.

## A.6 Proof of Proposition 12

The control variables are $\phi(b_{10}, b_{10}|\theta)$, $\phi(b_{10}, b_{11}|\theta)$, and $\phi(b_{11}, b_{11}|\theta)$. The probability constraints are

$$\begin{aligned}
\phi(b_{10}, b_{10}|\theta_0) + \phi(b_{11}, b_{11}|\theta_0) + 2\phi(b_{10}, b_{11}|\theta_0) \leq 1 \\
\phi(b_{10}, b_{10}|\theta_1) + \phi(b_{11}, b_{11}|\theta_1) + 2\phi(b_{10}, b_{11}|\theta_1) \leq 1.
\end{aligned}$$

At any solution, the following is true.

- Constraint (4.33) is binding. By way of contradiction, suppose this is not the case. Then we must have either $\phi(b_{10}, b_{10}|\theta_1) > 0$ or $\phi(b_{10}, b_{11}|\theta_1) > 0$ (or both). Suppose $\phi(b_{10}, b_{10}|\theta_1) > 0$. Then it is feasible to decrease $\phi(b_{10}, b_{10}|\theta_1)$ by a sufficiently small $\epsilon > 0$ and increase $\phi(b_{11}, b_{11}|\theta_1)$ by the same amount, leading to an increase in the objective function equal to $\epsilon(1-p^2) > 0$ without violating any incentive or probability constraint. An analogous reasoning applies to the complementary case with $\phi(b_{10}, b_{11}|\theta_1) > 0$.

- Constraint (4.34) is redundant. This follows from the fact that (4.33) binds and the assumption $t_\ell < t_h$.

- The probability constraint for state $\theta_1$ is binding. Suppose not. Then we must have $\phi(b_{11}, b_{11}|\theta_1) < 1$. Consequently, it is always feasible to increase $\phi(b_{11}, b_{11}|\theta_1)$ so as to relax constraint (4.35) and increase the Sender's objective function.

Using the results above, we solve a relaxed problem where we ignore the probability constraint in state $\theta_0$. One can verify that, at any solution that we are going to find, the probability constraint in $\theta_0$ is always satisfied. Notice that in this relaxed problem the constraint (4.35) is always binding. If not, then it would always be feasible to raise the objective function by increasing $\phi(b_{11}, b_{11}|\theta_0)$. From the binding constraints (4.33) and (4.35) it is immediate to obtain

$$\phi(b_{10}, b_{10}|\theta_0) = \frac{1 - t_\ell}{pt_\ell} \left[\phi(b_{10}, b_{11}|\theta_1) + p\phi(b_{10}, b_{10}|\theta_1)\right] - \frac{1}{p}\phi(b_{10}, b_{11}|\theta_0) \tag{A.51}$$

and

$$\phi(b_{11}, b_{11}|\theta_0) = \frac{1 - t_h}{t_h} \left[\phi(b_{11}, b_{11}|\theta_1) + p\phi(b_{10}, b_{11}|\theta_1)\right] - p\phi(b_{10}, b_{11}|\theta_0). \tag{A.52}$$

Substituting (A.51) and (A.52) into the objective function, the Sender's problem reduces to maximize

$$U_S(\phi) = \frac{1}{2t_h}\phi(b_{11}, b_{11}|\theta_1) + \frac{p^2}{2t_\ell}\phi(b_{10}, b_{10}|\theta_1) + \frac{p(t_\ell + t_h)}{2t_h t_\ell}\phi(b_{10}, b_{11}|\theta_1)$$

with respect to $\phi(b_{10}, b_{11}|\theta_0)$, $\phi(b_{10}, b_{10}|\theta_1)$, $\phi(b_{10}, b_{11}|\theta_1)$, and $\phi(b_{11}, b_{11}|\theta_1)$, subject to non-negativity constraints and the probability constraint

$$\phi(b_{10}, b_{10}|\theta_1) + \phi(b_{11}, b_{11}|\theta_1) + 2\phi(b_{10}, b_{11}|\theta_1) = 1.$$

The first-order conditions for optimality with respect to $\phi(b_{10}, b_{10}|\theta_1)$, $\phi(b_{10}, b_{11}|\theta_1)$, and $\phi(b_{11}, b_{11}|\theta_1)$, are respectively

$$\frac{p^2}{2t_\ell} - \lambda \leq 0 \tag{A.53}$$

$$\frac{p(t_\ell + t_h)}{2t_\ell t_h} - 2\lambda \leq 0 \tag{A.54}$$

$$\frac{1}{2t_h} - \lambda \leq 0, \tag{A.55}$$

where $\lambda$ is the Lagrange multiplier for the probability constraint. The first-order condition with respect to $\phi(b_{10}, b_{11}|\theta_0)$ is trivial. Now, it is immediate that at least one of the control variables in $\theta_1$ must be strictly positive. Suppose $\phi(b_{10}, b_{10}|\theta_1) > 0$. Then $\lambda = \frac{p^2}{2t_\ell}$ from

(A.53). Using this, the other two first-order conditions hold simultaneously if and only if $p \geq \frac{t_\ell + t_h}{2t_h}$. If the latter holds with equality, then both (A.53) and (A.54) hold with equality, and (A.55) holds with strict inequality, so that we must have $\phi(b_{11}, b_{11}|\theta_1) = 0$. If $p > \frac{t_\ell + t_h}{2t_h}$, then (A.53) holds with equality, and (A.54) and (A.55) hold with strict inequality, so that we have $\phi(b_{10}, b_{10}|\theta_1) = 1$ and $\phi(b_{10}, b_{11}|\theta_1) = \phi(b_{11}, b_{11}|\theta_1) = 0$.

Now suppose that $\phi(b_{10}, b_{11}|\theta_1) > 0$. From (A.54) we get $\lambda = \frac{p(t_\ell + t_h)}{4t_\ell t_h}$. Using this, the other two first-order conditions hold simultaneously if and only if $\frac{2t_\ell}{t_\ell + t_h} \leq p \leq \frac{t_\ell + t_h}{2t_h}$. If $p = \frac{2t_\ell}{t_\ell + t_h}$, then (A.54) and (A.55) hold with equality, and (A.53) holds with strict inequality, so that $\phi(b_{10}, b_{10}|\theta_1) = 0$. If $\frac{2t_\ell}{t_\ell + t_h} < p < \frac{t_\ell + t_h}{2t_h}$, then (A.54) holds with equality, and (A.53) and (A.55) hold with strict inequality, so that we have $\phi(b_{10}, b_{11}|\theta_1) = \frac{1}{2}$ and $\phi(b_{10}, b_{10}|\theta_1) = \phi(b_{11}, b_{11}|\theta_1) = 0$.

Finally, suppose $\phi(b_{11}, b_{11}|\theta_1) > 0$. We get $\lambda = \frac{1}{2t_h}$ from (A.55). Using this, the other two first-order conditions hold simultaneously if and only if $p \leq \frac{2t_\ell}{t_\ell + t_h}$. If the latter holds with equality, then both (A.54) and (A.55) hold with equality, and (A.53) holds with strict inequality, so that we must have $\phi(b_{10}, b_{10}|\theta_1) = 0$. If $p < \frac{2t_\ell}{t_\ell + t_h}$, then (A.55) holds with equality, and (A.53) and (A.54) hold with strict inequality, so that we have $\phi(b_{11}, b_{11}|\theta_1) = 1$ and $\phi(b_{10}, b_{10}|\theta_1) = \phi(b_{10}, b_{11}|\theta_1) = 0$. In all cases, the values for control variables in state $\theta_0$ are found by using (A.51) and (A.52) and the corresponding non-negativity constraints.

## A.7 Incentive constraints for the case without information elicitation and $k = 1$

By pivotality, $\sigma(1, 1|t, \theta) = 0$ for every $t \in T$ and every $\theta \in \Theta$. This implies that, for every $\theta \in \Theta$, we have the following:

$$\begin{aligned}
&\phi(b_{01}, b_{01}|\theta) = \phi(b_{01}, b_{10}|\theta) = \phi(b_{01}, b_{11}|\theta) = 0 \\
&\phi(b_{10}, b_{01}|\theta) = \phi(b_{10}, b_{10}|\theta) = \phi(b_{10}, b_{11}|\theta) = 0 \\
&\phi(b_{11}, b_{01}|\theta) = \phi(b_{11}, b_{10}|\theta) = \phi(b_{11}, b_{11}|\theta) = 0.
\end{aligned}$$

It follows that the full set of obedience constraints for player 1 is the following:

$$\begin{aligned}
(1 - t_\ell)\left[\phi(b_{00}, b_{00}|\theta_1) + p\phi(b_{00}, b_{01}|\theta_1) + (1 - p)\phi(b_{00}, b_{10}|\theta_1)\right] \leq \\
t_\ell\left[\phi(b_{00}, b_{00}|\theta_0) + p\phi(b_{00}, b_{01}|\theta_0) + (1 - p)\phi(b_{00}, b_{10}|\theta_0)\right] \qquad (t_\ell, b_{00})
\end{aligned}$$

$$(1-t_h)\left[\phi(b_{00},b_{00}|\theta_1)+p\phi(b_{00},b_{01}|\theta_1)+(1-p)\phi(b_{00},b_{10}|\theta_1)\right]\leq$$
$$t_h\left[\phi(b_{00},b_{00}|\theta_0)+p\phi(b_{00},b_{01}|\theta_0)+(1-p)\phi(b_{00},b_{10}|\theta_0)\right] \qquad (t_h,b_{00})$$

$$(1-t_\ell)\phi(b_{01},b_{00}|\theta_1)\leq t_\ell\phi(b_{01},b_{00}|\theta_0) \qquad (t_\ell,b_{01})$$

$$t_h\phi(b_{01},b_{00}|\theta_0)\leq(1-t_h)\phi(b_{01},b_{00}|\theta_1) \qquad (t_h,b_{01})$$

$$t_\ell\phi(b_{10},b_{00}|\theta_0)\leq(1-t_\ell)\phi(b_{10},b_{00}|\theta_1) \qquad (t_\ell,b_{10})$$

$$(1-t_h)\phi(b_{10},b_{00}|\theta_1)\leq t_h\phi(b_{10},b_{00}|\theta_0) \qquad (t_h,b_{10})$$

$$t_\ell\phi(b_{11},b_{00}|\theta_0)\leq(1-t_\ell)\phi(b_{11},b_{00}|\theta_1) \qquad (t_\ell,b_{11})$$

$$t_h\phi(b_{11},b_{00}|\theta_0)\leq(1-t_h)\phi(b_{11},b_{00}|\theta_1), \qquad (t_h,b_{11})$$

where the tag $(t,b)$ indicates that the corresponding constraint is for type $t$ to obey her action recommendation contained in $b$.

Now, by using $t_\ell<t_h$, it is immediate that constraint $(t_\ell,b_{11})$ is implied by $(t_h,b_{11})$, and that $(t_h,b_{00})$ is implied by $(t_\ell,b_{00})$. Furthermore, we must have

$$\phi(b_{01},b_{00}|\theta_0)=\phi(b_{01},b_{00}|\theta_1)=0.$$

To see why this is true, let us consider two cases. First, suppose $\phi(b_{01},b_{00}|\theta_0)=0$. By constraint $(t_\ell,b_{01})$, we must have $\phi(b_{01},b_{00}|\theta_1)=0$ as well. Consequently, both $(t_\ell,b_{01})$ and $(t_h,b_{01})$ would hold trivially with equality. Alternatively, suppose $\phi(b_{01},b_{00}|\theta_0)>0$. By

constraint $(t_h, b_{01})$, we must have $\phi(b_{01}, b_{00}|\theta_1) > 0$ as well. But then we obtain

$$\frac{\phi(b_{01}, b_{00}|\theta_0)}{\phi(b_{01}, b_{00}|\theta_1)} \leq \frac{1 - t_h}{t_h}$$

from $(t_h, b_{01})$ and

$$\frac{1 - t_\ell}{t_\ell} \leq \frac{\phi(b_{01}, b_{00}|\theta_0)}{\phi(b_{01}, b_{00}|\theta_1)}$$

from $(t_\ell, b_{01})$. These two inequalities cannot hold simultaneously because $t_\ell < t_h$ by assumption.

By symmetry and pivotality of $\sigma$, we have

$$\phi(b_{10}, b_{00}|\theta) + \phi(b_{11}, b_{00}|\theta) = \phi(b_{00}, b_{10}|\theta) + \phi(b_{00}, b_{11}|\theta) \tag{A.56}$$

and

$$\phi(b_{01}, b_{00}|\theta) + \phi(b_{11}, b_{00}|\theta) = \phi(b_{00}, b_{01}|\theta) + \phi(b_{00}, b_{11}|\theta). \tag{A.57}$$

Now, the obedience constraints $(t_\ell, b_{01})$ and $(t_h, b_{01})$ for player 2 are

$$(1 - t_\ell)\phi(b_{00}, b_{01}|\theta_1) \leq t_\ell \phi(b_{00}, b_{01}|\theta_0) \qquad (t_\ell, b_{01}, \text{ player 2})$$

and

$$t_h \phi(b_{00}, b_{01}|\theta_0) \leq (1 - t_h)\phi(b_{00}, b_{01}|\theta_1), \qquad (t_h, b_{01}, \text{ player 2})$$

respectively. By the same argument that we used for player 1, we can conclude that $\phi(b_{00}, b_{01}|\theta) = 0$. Since we also proved that $\phi(b_{01}, b_{00}|\theta) = 0$, it follows that (A.57) reduces to

$$\phi(b_{11}, b_{00}|\theta) = \phi(b_{00}, b_{11}|\theta),$$

and using the latter in (A.56) we also get

$$\phi(b_{10}, b_{00}|\theta) = \phi(b_{00}, b_{10}|\theta).$$

Notice that, by using all of the restrictions on $\phi$'s derived so far, the incentive constraints for player 1 and player 2 become exactly the same. Thus it is without loss of generality to focus on player 1 alone, as done in the main text.

Finally, using all the restrictions on $\phi$'s obtained above, the Sender's objective function

can be written as follows:

$$\begin{aligned}
U_S(\phi) = &\frac{1}{2}\Big\{p^2\left[2\phi(b_{10}, b_{00}|\theta_0) + 2\phi(b_{11}, b_{00})|\theta_0\right] \\
&+ 2p(1-p)\left[\phi(b_{10}, b_{00}|\theta_0) + 2\phi(b_{11}, b_{00})|\theta_0\right] \\
&+ (1-p)^2\left[2\phi(b_{11}, b_{00})|\theta_0\right]\Big\} \\
&+ \frac{1}{2}\Big\{p^2\left[2\phi(b_{10}, b_{00}|\theta_1) + 2\phi(b_{11}, b_{00})|\theta_1\right] \\
&+ 2p(1-p)\left[\phi(b_{10}, b_{00}|\theta_1) + 2\phi(b_{11}, b_{00})|\theta_1\right] \\
&+ (1-p)^2\left[2\phi(b_{11}, b_{00})|\theta_1\right]\Big\} \\
= &\sum_{\theta\in\Theta}\left[\phi(b_{11}, b_{00}|\theta) + p\phi(b_{10}, b_{00}|\theta)\right],
\end{aligned}$$

which is expression (4.36) in the main text.

## A.8 Proof of Proposition 13

Consider first the set of incentive constraints (4.37)-(4.40). The probability constraints are:

$$\begin{aligned}
\phi(b_{00}, b_{00}|\theta_0) + 2\phi(b_{10}, b_{00}|\theta_0) + 2\phi(b_{11}, b_{00}|\theta_0) = 1 \\
\phi(b_{00}, b_{00}|\theta_1) + 2\phi(b_{10}, b_{00}|\theta_1) + 2\phi(b_{11}, b_{00}|\theta_1) = 1.
\end{aligned}$$

At any solution, the following is true.

- Constraint (4.38) is binding. To see this, suppose by way of contradiction that it is slack. Since the left-hand side of (4.38) must be non-negative, it follows that $\phi(b_{10}, b_{00}|\theta_1)$ on the right-hand side must be strictly positive. Then one can decrease $\phi(b_{10}, b_{00}|\theta_1)$ by a sufficiently small amount $\epsilon > 0$ and increase $\phi(b_{11}, b_{00}|\theta_1)$ by the same amount $\epsilon$ so as to leave incentive and probability constraints unaffected. This shift of probability mass determines a change in the objective function equal to $\epsilon(1-p) > 0$, so reaching a contradiction.

- The fact that (4.38) is binding and that $t_\ell < t_h$ by assumption imply that constraint (4.39) is redundant.

- From the binding constraint (4.38) we obtain

$$\phi(b_{10}, b_{00}|\theta_0) = \frac{1-t_\ell}{t_\ell}\,\phi(b_{10}, b_{00}|\theta_1),$$

and plugging the latter expression into (4.37) we have that constraint (4.37) reduces to

$$(1-t_\ell)\phi(b_{00},b_{00}|\theta_1) \leq t_\ell \phi(b_{00},b_{00}|\theta_0). \tag{A.58}$$

- It must be the case that $\phi(b_{00},b_{00}|\theta_1)=0$. If not, then it is always feasible to decrease $\phi(b_{00},b_{00}|\theta_1)$ by an arbitrarily small amount $\epsilon>0$ and increase $\phi(b_{11},b_{00}|\theta_1)$ by $\frac{\epsilon}{2}$. This shift of probability mass relaxes incentive constraints (A.58) and (4.40) and entails a strict increase in the objective function, so contradicting optimality. Since we must have $\phi(b_{00},b_{00}|\theta_1)=0$, it follows that constraint (A.58) is implied by the non-negativity constraint on $\phi(b_{00},b_{00}|\theta_0)$. Furthermore, the probability constraint for state $\theta_1$ can be written as

$$\phi(b_{11},b_{00}|\theta_1) = \frac{1}{2} - \phi(b_{10},b_{00}|\theta_1).$$

Using the results obtained so far, the Sender's maximization problem reduces to the the following simplified problem:

$$\max \quad \frac{1}{2} + \phi(b_{11},b_{00}|\theta_0) + \frac{(p-t_\ell)}{t_\ell}\phi(b_{10},b_{00}|\theta_1)$$

with respect to $\phi(b_{11},b_{00}|\theta_0)$ and $\phi(b_{10},b_{00}|\theta_1)$ subject to non-negativity constraints and to the following:

$$\phi(b_{10},b_{00}|\theta_1) \leq \frac{1}{2} \tag{A.59}$$

$$\frac{(1-t_\ell)}{t_\ell}\phi(b_{10},b_{00}|\theta_1) + \phi(b_{11},b_{00}|\theta_0) \leq \frac{1}{2} \tag{A.60}$$

$$t_h\phi(b_{11},b_{00}|\theta_0) + (1-t_h)\phi(b_{10},b_{00}|\theta_1) \leq \frac{1-t_h}{2}, \tag{A.61}$$

where (A.59) and (A.60) are probability constraints, and (A.61) is a rewriting of (4.40). Now we argue that (A.61) must be binding at every solution. Suppose by way of contradiction that this is not the case. The first-order condition for optimality with respect to $\phi(b_{11},b_{00}|\theta_0)$ is $1-\lambda_2-\lambda_3 t_h \leq 0$, where $\lambda_2$ is the Lagrange multiplier associated to (A.60) and $\lambda_3$ is the multiplier associated to (A.61). Since (A.61) is slack by assumption, $\lambda_3 = 0$ by complementary slackness. Consequently, $1 \leq \lambda_2$, so that (A.60) must be binding. From the latter constraint, we obtain

$$\phi(b_{11},b_{00}|\theta_0) = \frac{1}{2} - \frac{(1-t_\ell)}{t_\ell}\phi(b_{10},b_{00}|\theta_1),$$

and substituting into (A.61), which is slack by assumption, we must have

$$\frac{t_h}{2} - \frac{(t_h - t_\ell)}{t_\ell}\phi(b_{10}, b_{00}|\theta_1) < \frac{1 - t_h}{2}. \tag{A.62}$$

Now, since $\phi(b_{10}, b_{00}|\theta_1) \leq \frac{1}{2}$ by (A.59), the left-hand side of (A.62) is always greater than or equal to $\frac{t_h}{2} - \frac{t_h - t_\ell}{2t_\ell}$. But then we have

$$\frac{t_h}{2} - \frac{t_h - t_\ell}{2t_\ell} < \frac{1 - t_h}{2},$$

which is impossible since $t_\ell > \frac{1}{2}$ by assumption.

Using the fact that (A.61) is binding, we obtain

$$\phi(b_{11}, b_{00}|\theta_0) = \frac{1 - t_h}{2t_h} - \frac{(1 - t_h)}{t_h}\phi(b_{10}, b_{00}|\theta_1). \tag{A.63}$$

Substituting the latter into (A.60), one can verify that (A.60) always holds if $\phi(b_{10}, b_{00}|\theta_1)$ satisfies (A.59). Finally, substituting (A.63) into the objective function, we can easily see that the maximization problem reduces to choose $\phi(b_{10}, b_{00}|\theta_1)$ between 0 and $\frac{1}{2}$ so as to maximize

$$U_S(\phi) = \frac{1}{2t_h} + \left(\frac{pt_h - t_\ell}{t_{\ell t_h}}\right)\phi(b_{10}, b_{00}|\theta_1).$$

Therefore, if $p < \frac{t_\ell}{t_h}$, the coefficient of $\phi(b_{10}, b_{00}|\theta_1)$ in $U_S(\phi)$ is strictly negative, so that it is optimal to set $\phi(b_{10}, b_{00}|\theta_1) = 0$. If $p > \frac{t_\ell}{t_h}$, the coefficient of $\phi(b_{10}, b_{00}|\theta_1)$ is strictly positive, and then $\phi(b_{10}, b_{00}|\theta_1) = \frac{1}{2}$ is optimal. If $p = \frac{t_\ell}{t_h}$, the coefficient is null, and $\phi(b_{10}, b_{00}|\theta_1)$ is optimal for any value between 0 and $\frac{1}{2}$.

# Bibliography


Ricardo Alonso and Odilon Câmara. Persuading voters. *The American Economic Review*, 106(11):3590–3605, 2016.

Luca Anderlini, Dino Gerardi, and Roger Lagunoff. Communication and learning. *The Review of Economic Studies*, 79(2):419–450, 2011.

Robert Aumann and Adam Brandenburger. Epistemic conditions for nash equilibrium. *Econometrica*, 63:1161–1161, 1995.

Robert J Aumann. Subjectivity and correlation in randomized strategies. *Journal of mathematical Economics*, 1(1):67–96, 1974.

Robert J Aumann. Agreeing to disagree. *The annals of statistics*, pages 1236–1239, 1976.

Robert J Aumann. Correlated equilibrium as an expression of bayesian rationality. *Econometrica*, pages 1–18, 1987.

Robert J Aumann and Sergiu Hart. Long cheap talk. *Econometrica*, 71(6):1619–1660, 2003.

David Austen-Smith and Jeffrey S Banks. Information aggregation, rationality, and the condorcet jury theorem. *American Political Science Review*, 90(01):34–45, 1996.

James Emil Avery, Jean-Yves Moyen, Pavel Ruzicka, and Jakob Grue Simonsen. Chains, antichains, and complements in infinite partition lattices. *Algebra universalis*, 79(2):37, 2018.

Christian W Bach and Andrés Perea. Two definitions of correlated equilibrium. EPICENTER Working Paper No. 18, Maastricht University, 2018.

Michael Bacharach. Some extensions of a claim of Aumann in an axiomatic model of knowledge. *Journal of Economic Theory*, 37(1):167–190, 1985.

Imre Bárány. Fair distribution protocols or how the players replace fortune. *Mathematics of Operations Research*, 17(2):327–340, 1992.

Paulo Barelli. Consistency of beliefs and epistemic conditions for nash and correlated equilibria. *Games and Economic Behavior*, 67(2):363–375, 2009.

Dirk Bergemann and Stephen Morris. Bayes correlated equilibrium and the comparison of information structures in games. *Theoretical Economics*, 11:487–522, 2016a.

Dirk Bergemann and Stephen Morris. Information design, bayesian persuasion, and bayes correlated equilibrium. *AEA Papers and Proceedings*, 106(5):586–591, 2016b.

Dirk Bergemann and Stephen Morris. Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95, 2019.

Dirk Bergemann, Alessandro Bonatti, and Alex Smolin. The design and price of information. *American Economic Review*, 108:1–48, 2018.

James Bergin. We eventually agree. *Mathematical social sciences*, 17(1):57–66, 1989.

Andreas Blume and Oliver Board. Language barriers. *Econometrica*, 81(2):781–812, 2013.

Giacomo Bonanno and Klaus Nehring. Agreeing to disagree: a survey. Working Paper Series 97-18, Department of Economics, University of California, Davis, 1997.

Adam Brandenburger, Eddie Dekel, and John Geanakoplos. Correlated equilibrium with generalized information structures. *Games and Economic Behavior*, 4(2):182–201, 1992.

Bernard Caillaud and Jean Tirole. Consensus building: How to persuade a group. *The American Economic Review*, pages 1877–1900, 2007.

David Cass and Karl Shell. Do sunspots matter? *Journal of political economy*, 91(2): 193–227, 1983.

Jonathan A K Cave. Learning to agree. *Economics Letters*, 12(2):147–152, 1983.

Jimmy Chan, Seher Gupta, Fei Li, and Yun Wang. Pivotal persuasion. *Journal of Economic Theory*, 180:178–202, 2019.

Martin W Cripps, Jeffrey C Ely, George J Mailath, and Larry Samuelson. Common learning. *Econometrica*, 76(4):909–933, 2008.

Boudewijn De Bruin. *Explaining games: the epistemic programme in game theory*, volume 346. Springer, 2010.

Kris De Jaegher. A game-theoretic rationale for vagueness. *Linguistics and Philosophy*, 26 (5):637–659, 2003.

Alfredo Di Tillio. A note on one-shot public mediated talk. *Games and Economic Behavior*, 46(2):425–433, 2004.

Ronald Fagin, Joseph Y Halpern, Yoram Moses, and Moshe Vardi. *Reasoning about knowledge*. MIT press, 2004.

Françoise Forges. Can sunspots replace a mediator? *Journal of Mathematical Economics*, 17(4):347–368, 1988.

Françoise Forges. Universal mechanisms. *Econometrica*, pages 1341–1364, 1990.

Françoise Forges. Five legitimate definitions of correlated equilibrium in games with incomplete information. *Theory and decision*, 35(3):277–310, 1993.

Françoise Forges. Correlated equilibrium in games with incomplete information revisited. *Theory and decision*, 61(4):329–344, 2006.

Paolo Galeazzi and Emiliano Lorini. Epistemic logic meets epistemic game theory: a comparison between multi-agent kripke models and type spaces. *Synthese*, 193(7):2097–2127, 2016.

John D Geanakoplos and Heraklis M Polemarchakis. We can't disagree forever. *Journal of Economic Theory*, 28:192–200, 1982.

Dino Gerardi and Leeat Yariv. Deliberative voting. *Journal of Economic Theory*, 134(1): 317–338, 2007.

Jacob Glazer and Ariel Rubinstein. On optimal rules of persuasion. *Econometrica*, 72(6): 1715–1736, 2004.

Jacob Glazer and Ariel Rubinstein. A study in the pragmatics of persuasion: a game theoretical approach. *Theoretical Economics*, 1(4):395–410, 2006.

Joseph Y Halpern. *Reasoning about uncertainty*. MIT press, 2003.

Joseph Y Halpern and Willemien Kets. A logic for reasoning about ambiguity. *Artificial Intelligence*, 209:1–10, 2014.

Joseph Y Halpern and Willemien Kets. Ambiguous language and common priors. *Games and Economic Behavior*, 90:171–180, 2015.

Aviad Heifetz. Comment on consensus without common knowledge. *Journal of Economic Theory*, 70(1):273–277, 1996.

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

Frédéric Koessler. Common knowledge and consensus with noisy communication. *Mathematical Social Sciences*, 42(2):139–159, 2001.

Anton Kolotilin, Tymofiy Mylovanov, Andriy Zapechelnyuk, and Ming Li. Persuasion of a privately informed receiver. *Econometrica*, 85(6):1949–1964, 2017.

Paul Krasucki. Protocols forcing consensus. *Journal of Economic Theory*, 70(1):266–272, 1996.

Ehud Lehrer. Mediated talk. *International Journal of Game Theory*, 25(2):177–188, 1996.

Ehud Lehrer and Sylvain Sorin. One-shot public mediated talk. *Games and Economic Behavior*, 20(2):131–148, 1997.

Barton L Lipman. Why is language vague? Boston University, 2009.

Emiliano Lorini and François Schwarzentruber. A modal logic of epistemic games. *Games*, 1(4):478–526, 2010.

Michael Maschler, Eilon Solan, and Shmuel Zamir. *Game Theory*. Cambridge University Press, 2013.

Laurent Mathevet, Jacopo Perego, and Ina Taneva. On information design in games. *Journal of Political Economy*, 2019. forthcoming.

Yoram Moses and Gal Nachum. Agreeing to disagree after all. In *Proceedings of the 3rd conference on Theoretical Aspects of Reasoning about Knowledge*, pages 151–168, 1990.

Manuel Mueller-Frank. A general framework for rational learning in social networks. *Theoretical Economics*, 8(1):1–40, 2013.

Roger B Myerson. Optimal coordination mechanisms in generalized principal–agent problems. *Journal of mathematical economics*, 10(1):67–81, 1982.

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

Lars Tyge Nielsen. Common knowledge, communication, and convergence of beliefs. *Mathematical Social Sciences*, 8(1):1–14, 1984.

Martin J Osborne and Ariel Rubinstein. *A course in game theory*. The MIT press, 1994.

Rohit Parikh. Finite and infinite dialogues. In Yiannis Moschovakis, editor, *Logic from Computer Science*, pages 481–497. Springer, 1992.

Rohit Parikh and Paul Krasucki. Communication, consensus, and knowledge. *Journal of Economic Theory*, 52(1):178–189, 1990.

Steven Roman. *Lattices and ordered sets*. Springer, 2008.

Ariel Rubinstein. The electronic mail game: Strategic behavior under "almost common knowledge". *The American Economic Review*, pages 385–391, 1989.

Dov Samet. Agreeing to disagree: The non-probabilistic case. *Games and Economic Behavior*, 69(1):169–174, 2010.

Ina Taneva. Information design. *American Economic Journal: Microeconomics*, 2019. forthcoming.

Bassel Tarbush. Counterfactuals in "agreeing to disagree" type results. *Mathematical Social Sciences*, 84:125–133, 2016.

Elias Tsakas and Mark Voorneveld. On consensus through communication without a commonly known protocol. *Journal of Mathematical Economics*, 47(6):733–739, 2011.

Yun Wang. Bayesian persuasion with multiple receivers. Working Paper, University of Pittsburgh, 2013.

RB Washburn and D Teneketzis. Asymptotic agreement among communicating decision-makers. *Stochastics*, 13(1-2):103–129, 1984.

Sonia Weyers. Three results on communication, information and common knowledge. CORE Discussion Papers 1992028, Université Catholique de Louvain, 1992.