NOTA DI LAVORO

28.2017

Information Design In Coalition Formation Games

Sareh Vosooghi, University of Oxford

Economic Theory
Series Editor: Carlo Carraro

Information Design In Coalition Formation Games

By Sareh Vosooghi, University of Oxford

Summary

I examine a setting, where an information sender conducts research into a payoff-relevant state variable, and releases information to agents, who consider joining a coalition. The agents' actions can cause harm by contributing to a public bad. The sender, who has commitment power, by designing an information mechanism (a set of signals and a probability distribution over them), maximises his payoff, which depends on the action taken by the agents, and the state variable. I show that the coalition size, as a function of beliefs of agents, is an endogenous variable, induced by the information sender. The optimal information mechanism from the general set of public information mechanisms, in coalition formation games is derived. I also apply the results to International Environmental Agreements (IEAs), where a central authority, as an information sender, attempts to reduce the global level of greenhouse gases (GHG) by communication of information on social cost of GHG.

Keywords: Coalition Formation, Learning, Information Persuasion, International Environmental Agreements

JEL Classification: D83, D70, C72, Q54

An earlier version of this paper circulated under the title "Optimal Communication of Climate Change With The Public". The paper is part of my PhD thesis at the University of Edinburgh. I thank my supervisors Tim Worrall and Jozsef Sakovics for their insightful guidance. In addition, I am grateful to Jeffery Ely, Jonathan Thomas, Hassan Benchekrou, Bipasa Datta, Ed Hopkins, Andy Zapechelnyuk, Ina Taneva, Alessandro Tavoni also seminar and conference participants at CTN, RES, EEA, Edinburgh, and Humboldt for their helpful comments and conversations.

This paper was presented at the 22nd Coalition Theory Network Workshop, which was held in Glasgow, UK, on 11 - 12 May 2017.

Address for correspondence:
Sareh Vosooghi
University of Oxford
St Edmund Hall
United Kingdom
E-mail: sareh.vosooghi@seh.ox.ac.uk

Information Design In Coalition Formation Games*

Sareh Vosooghi†

June, 2017

Abstract

I examine a setting, where an information sender conducts research into a payoff-relevant state variable, and releases information to agents, who consider joining a coalition. The agents' actions can cause harm by contributing to a public bad. The sender, who has commitment power, by designing an information mechanism (a set of signals and a probability distribution over them), maximises his payoff, which depends on the action taken by the agents, and the state variable. I show that the coalition size, as a function of beliefs of agents, is an endogenous variable, induced by the information sender. The optimal information mechanism from the general set of public information mechanisms, in coalition formation games is derived. I also apply the results to International Environmental Agreements (IEAs), where a central authority, as an information sender, attempts to reduce the global level of greenhouse gases (GHG) by communication of information on social cost of GHG.

Key words: Coalition Formation; Learning; Information Persuasion; International Environmental Agreements

JEL Classification: D83; D70; C72; Q54

1 Introduction

The possibility of affecting decision-making of groups of agents in writing cooperative agreements is an important issue, and has significant implications for everyday phenomena. In particular, the efficiency

*An earlier version of this paper circulated under the title "Optimal Communication of Climate Change With The Public". The paper is part of my PhD thesis at the University of Edinburgh. I thank my supervisors Tim Worrall and Jozsef Sakovics for their insightful guidance. In addition, I am grateful to Jeffery Ely, Jonathan Thomas, Hassan Benchekrou, Bipasa Datta, Ed Hopkins, Andy Zapechelnyuk, Ina Taneva, Alessandro Tavoni also seminar and conference participants at CTN, RES, EEA, Edinburgh, and Humboldt for their helpful comments and conversations.

†sareh.vosooghi@seh.ox.ac.uk, St Edmund Hall, University of Oxford, United Kingdom

implications are compounded where the agents' collective actions may lead to prevention of public bads. In this paper, a theoretical framework is constructed to explain the significance of communication of a research central authority (information sender) with agents (receivers) in coalition formations. Using the approach of information design, the paper suggest a new method to implement desired outcome in n-player coalition formation games with public bads.

Specifically, I consider a model in which the sender maximises its payoff which depends on the action undertaken by the agents and a payoff-relevant state variable. Before any decision making by the agents, the sender initiates the research by choosing an optimal information mechanism (a set of signals, as recommended actions, and an information policy, which is a probability distribution over signals) to maximise his expected payoff. Choosing a research strategy is modelled as choosing a probability distribution over a signal. Before choosing the information policy, the sender is uninformed about the state. Furthermore, the sender is assumed to have commitment power. So, after conducting the research on the state variable, he commits to the information mechanism and sends a public signal. The payoffs of both sender and receivers are common knowledge.

Finding the optimal information mechanism is studied in the recent but growing literature on information design, introduced by Kamenica and Gentzkow (2011). This paper generalises Kamenica and Gentzkow (2011) to coalition formation games, where there are multiple receivers who are either non-signatories or signatories of a treaty, and the information sender can achieve the formation of a desired coalition by designing an optimal information mechanism.

The interaction of the agents is modelled as a linear n-player coalition game with two stages: the membership stage and the "action" stage, where they decide about prevention or contribution to the public bad. The timeline is as follows. The sender announces an information policy, and sends a public signal to the agents. Given the information mechanism and the public signal, the agents update their (common) prior belief and, in the membership stage, they decide whether to join a coalition or remain non-signatories. Finally, the non-signatories and signatories of the coalition choose their actions.

Subgame perfection implies that given their posterior belief, the agents decide about their actions and subsequently their membership strategies. Finally, given the best responses of the agents in the action and membership subgames, the sender chooses an optimal information mechanism.

The information mechanism leads to a probability distribution over posterior beliefs of the agents, which in turn determines the action and membership outcomes. Kamenica and Gentzkow (2011) use the

terminology of "persuasion" to refer to affecting the receiver's action by inducing a certain distribution over her posterior beliefs. Here, it reflects the fact that the sender's choice of information policy influences the coalition formation choices of the agents. In other words, the sender can persuade the agents as to what action and membership choices to make.

Therefore, beliefs are endogenous variables, and in contrast to the literature, posterior beliefs are not fixed parameters. Indeed, in our analysis, the beliefs about the state variable are the critical variables in determining the profitability and stability (self-enforceability for signatories and non-signatories) of a coalition, and in this paper, the threshold behaviours of the signatories and non-signatories of an agreement are dissected with respect to their beliefs about the state variable.

It is shown that the number of signatories of a coalition, depends on the beliefs, and therefore, it is an endogenous variable induced by the information sender. I consider a simple and tractable model, where the state space and the action choices are binary. Therefore, strategies take a threshold form with respect to beliefs about the social cost. In the action stage, assume the agents choose between causing harm by generating a public bad, or preventing harm. In addition, let the state variable reflect the social cost of public bad, and the state space be either high or low. Then, for example, by focusing on the support of common belief about high social cost, it is shown that such beliefs have three distinct ranges, divided by the stability thresholds of different coalitions. The lowest threshold belief is the threshold of grand coalition, below which all agents choose causing harm. The largest threshold is the threshold of singleton coalition structure (coalitions formed by one agent only), above which all agents coordinate on preventing harm. The range of beliefs between the thresholds of grand coalition and coalition of singletons, is where signatories of stable coalitions choose prevention, while non-signatories cause harm. This range itself is divided into sub-partitions, where in each sub-partition of beliefs, a unique coalition of a size, which varies between two and full participation, is stable.

In two applications, I examine two different payoff specifications for the sender, where either his expected payoff coincides with the expected payoff of a group of agents, or he minimises total stock of the public bad. First, the cases are investigated where the expected payoff of the sender coincides with the expected payoff of signatories, non-signatories, or a combination of both groups. This application exemplifies situations where the information sender supports a group of agents, so maximises their payoffs. It is shown that in such cases, the unique optimal information mechanism takes the form of full learning by the agents. The fact that alignment of the payoffs of the sender and receiver leads to full revelation,

is a known result in the literature on information design. But here, the sender faces a problem where the two groups of signatories and non-signatories of the agreement have different threshold behaviours, and may have conflict of interests. Moreover, given any of the expected payoffs of the sender mentioned, the induced equilibrium action of the agents, coincides with the socially optimal action outcome, as if the expected payoff of a potential grand coalition were maximised.

In addition, the problem of a sender whose target is reducing the total level of public bad is examined, and the model suggests that the optimal information mechanism is imperfect learning of the social cost. The application epitomises the historic climate change agreement, which was adopted in Paris in December 2015. This led to the formation of a grand coalition of over 190 agents with the objective of reducing the level of GHG emissions, with the consequent hope of limiting the rise in the average global temperature. The success in obtaining full participation in the international environmental agreement (IEA) was undoubtedly in part due to the international research institutes (the IPCC and other partners of the UN), which had conducted research on climate change, and communicated the results to the countries. The analysis here provides a simple setup through which to understand the underlying logic behind the observed behaviour of the counties and the research authorities involved in the Paris communications. More interestingly, it is shown that the optimal information persuasion, given the objective of minimising the total level of public bad, also leads to the selection of the socially optimal action and membership strategies.

The remainder of the paper is organised as follows. The related literature is reviewed in the next section, and the model is presented in section 3. Section 4 provides analysis of the persuasion problem in a coalition formation setting. The two applications are introduced in sections 5 and 6 respectively. Finally, section 7 concludes and the appendix contains all proofs.

2 Related Literature

Coalition formation has been extensively studied in game theory. Ray (2007) and Ray and Vohra (2015) provide a broad perspective on the main theoretical advancement of the topic. The current paper contributes to the non-cooperative strand of the coalition theory. The dominant part of the recent literature of non-cooperative coalition formation is focused on finding endogenous channels of coalition formations. The literature has been greatly enriched by consideration of inter-coalitional communications.

Chwe (1994) introduces farsighted stability, where agents take into account the reaction of others in case of deviation. Using the concept of "largest consistent set", he suggests a method to overcome the myopia problem, which is prevalent under the d'Aspremont stability criterion. Along the same lines, Bloch (1999) and Ray and Vohra (1997, 1999) suggest using "value" of coalitions in strategic games, to capture the fact that payoffs of agents depend on the entire coalition structure.

As another endogenous channel, Kosfeld et al (2009) investigate endogenous formation of sanctioning institutions in a linear n-player public good game. They show that establishment of such institutions by members of an "organisation" leads to overcoming the free-riding problem, and increasing payoffs for the participants.

In the current research, the coalition formation is endogenous due to the introduction of an information sender, as a new player in the collective problem.[1] Here the sender is able to design and implement a coalition structure which maximises his payoffs.

This paper contributes to the literature on non-cooperative coalition formation with public goods (bads), which is mainly studied in relation to the IEAs. Seminal papers about coalition formation in IEAs include Carrao and Siniscalco (1993) and Barrett (1994). The literature on IEAs is better reviewed by Toman (1998), Finus (2002), Wagner (2002), Kolstad and Toman (2001), Aldy and Stavins (2009) and Benchekroun and Long (2012), among others.

The role of uncertainty in the cooperative approach to coalition formation has been extensively studied. For example, see Allen and Yannelis (2001), Forges et al (2002), Dutta and Vohra (2005) for the literature on asymmetric information in coalition formation. In a seminal paper Myerson (2007) generalises the "core" concept to games of incomplete information.

Within non-cooperative game theory, uncertainty and strategic coalition formation have been given less attention. A relatively small subset of the literature on IEAs investigates the role of uncertainty and learning. In stochastic IEAs, the uncertainty is basically modelled as uncertainty about an unknown parameter in the payoff function of the countries involved, which is mainly the social cost of GHG (known as the cost-benefit ratio). A seminal paper in this strand of literature is that of Na and Shin (1998), which introduces asymmetric uncertainty (distributional uncertainty) in a coalition game of three countries. Ulph (2004) extends the idea of variable membership to stochastic IEAs, and Kolstad (2007) suggests learning after the membership stage and before the emission stage. Kolstad and Ulph (2011)

[1] This is in addition to the assumption of the farsightedness of agents.

investigate the case of ex-post asymmetry of social cost of GHG among countries. In addition, Finus and Pintassilgo (2013) assume a continuum choice variable, and by comparing the effects of learning in different stages of a coalition game for a symmetric and asymmetric state variable derive more general conclusions. Moreover, Barrett (2013) introduces uncertainty about a catastrophic threshold in a climate change coalition game, where if the threshold is met the countries suffer a catastrophic cost in addition to the conventional social cost of GHG.

In this paper, the coalition game is solved with respect to the beliefs about the social cost of public bad, where information may be noisy and learning can be imperfect. Hence, the full-learning and no-learning cases studied in the literature on stochastic IEAs, are considered as specific cases. Furthermore, most literature on the stochastic IEA concludes that the "veil of uncertainty" helps the formation of climate coalitions. In our analysis, where the information designer can implement the equilibrium outcome, the coalition formation is endogenous and depends on the expected payoff of the designer. In other words, if the payoff of the sender and (a group of) agents coincide, the optimal information policy takes the form of perfect learning. On the other hand, if the objective of the research central authority is minimising the total level of public bad, then optimal information policy results in partial learning of the state by the agents.

In addition to the literature on stochastic coalition formation, the paper also relates to the literature on information design. The seminal paper in this area is that of Kamenica and Gentzkow (2011), which introduces information persuasion of one agent by a sender. The literature is growing extensively in various dimensions, such as dynamic information design with public and private signals (Ely, 2015), persuasion of an informed receiver (Kolotilin et al, 2015), private persuasion by the sender (Taneva, 2016), costly persuasion for the sender (Gentzkow and Kamenica, 2014), and multiple senders and one receiver (Gentzkow and Kamenica, 2012).

This paper contributes to the strand of public persuasion. The existing literature mostly relates to the literature on media communication in a political economy environment. Gehlbach and Sonin (2014) use information design, where they model a government as a strategic player, which controls the media output to a mass of citizens. In a binary model (binary state and binary action choices for the citizens) the sender tries to influence the beliefs of citizens by sending public signals, and they derive the equilibrium level of media bias.

Shapiro (2016) applies the idea of persuasion by multiple senders to climate change. Referring to

the fact that media balances both scientific and political views on climate change, and the public does not reach a consensus on global warming, they suggest a political economy model, where the receiver is a voter, who has binary choices, and seeks information about a binary state variable. A mass of experts is divided into two groups of informed experts and "wrongly" informed experts who belong to the "opposition party". The voter receives messages indirectly through a media journalist, who may combine the messages of two experts, who could be from either parties. The opposition party is indeed the sender, which is maximising its payoff by allocating its own expert in the media to influence the journalists' report. They examine the effect of the cost of the opposition party hiring an expert on the informativeness of signals in equilibrium.

As a public persuasion model, instead of the effect of media on voters, this paper focuses on coalition formation. In contrast to the papers mentioned, there is no cost of information distortion, and the communication strategies of the sender in equilibrium lead to a socially optimal outcome. Moreover, the receivers of the public signal are heterogeneous with respect to being signatories or non-signatories of the coalition.

3 The Model

A sender releases public information to the agents about an (uncertain) payoff-relevant state variable, in order to induce some specific action. Let I refer to the set of agents. Agent $i \in I$ has linear payoff of $u_i(\mathbf{q}, \gamma) = q_i - \gamma Q$, where $\mathbf{q} = (q_1, q_2, ..., q_N)$ refers to the vector of "actions" chosen by the N agents, and assume that the action space is binary. Let $q_i \in \{0, 1\}$, and we label $q_i = 0$ as preventing harm and $q_i = 1$ as causing harm (or contributing to the public bad). Furthermore, Q is the accumulated stock of public bad, i.e. $Q = \sum_i q_i$. In addition, $\gamma \in \Gamma$ is the state variable about which the agents receive information. The state space is also taken to be a binary set and it is either low or high, i.e. $\Gamma = \{\gamma_l, \gamma_h\}$, where $\gamma_l < \gamma_h$. The state variable indeed reflects the marginal social cost of public bad for agent i, and clearly we focus on risk-neutral agents, which are ex ante symmetric.[2]

Furthermore, let γ_h be a finite real number, and $\gamma_h > 1$. As is shown in the next sections, this

[2]The functional form is commonly used in the literature, for example see Kosfeld et al (2009). As becomes clear in the next section, the assumptions of binary state variable and binary action space lead to the threshold behaviour of signatories and non-signatories with respect to their beliefs. In terms of robustness of the results with respect to these assumptions and the functional form, although the results depend on the single-crossing property of the expected payoffs, the linearity of the payoff function is for simplicity.

assumption ensures the existence of a possible large social cost, which leads to coordination of all agents on prevention of harm. For simplicity, let us assume $\gamma_l = 0$, which implies that if the social cost is γ_l, contribution to the public bad is not harmful.

The sender has payoff of $\boldsymbol{\nu}(\mathbf{q}(\gamma))$ which depends on the action chosen by the agents, that indirectly depends on the state variable. The payoffs of both sides are common knowledge.

At the beginning of the game, nature draws one parameter from set Γ for all agents. Before the agents choose any action, in order to affect their beliefs about the state variable, the sender, who has commitment power, can initiate research about the unknown state variable. Choosing a research strategy is modelled as choosing an information mechanism, which consists of the set of signals, S, and the information policy. Set S is finite, and the information policy is a map from the state, γ, to the probability distribution over signals when the true state is γ, i.e. $\gamma \to \pi(. \mid \gamma) \in \Delta(S)$. In other words, if the observed state by the sender is γ, the information policy specifies that the sender chooses a signal according to the rule of $\pi(. \mid \gamma)$. There is no need to specify the details of research, as it all reduces to the information mechanism. The sender commits to the chosen information policy. Furthermore, it is assumed that the research is costless for the sender.

The timeline is as follows. At the beginning of the game, and before observing the true state, the sender announces the information mechanism, and commits to it. By conducting the research, he observes the state[3]. Then, the results of research are presented as a public signal, $s \in S$, according to the information policy. The agents, interpret the information policy and the signal, using Bayes rule. So, given the information policy and the observed signal realisation, the agents update their belief about the state, and decide about joining a coalition and consequently, they choose an "action" to maximise their payoffs. We refer to this process as "persuasion".

The agents and the sender share a common prior, $p(\gamma)$. In addition, every piece of information that the agents ever get, comes from the sender. So, sending signal $s \in S$, the information sender knows the agents' updated belief, $\mu_s(\gamma) \in \Delta(\Gamma)$.

Therefore, for any state, each signal realisation, s, induces a posterior belief, $\mu_s(\gamma)$.[4] The total probability of a given signal s is $\pi(s) = \sum_{\gamma'} p(\gamma')\pi(s \mid \gamma')$, which is the sum of conditional probability

[3]If the sender does not observe the state realisation, he has access to a limited set of policies, for example, the truthful policy is not available. But given that limited set, it is still possible to derive the optimal policy. Hence, observation of the state is not a crucial assumption in the analysis. As I do not restrict the set of public mechanisms, and any probability distribution is available to the sender, it is assumed that he observes the state after setting the information policy.

[4]Using the Bayes rule, $\mu_s(\gamma) = \frac{p(\gamma)\pi(s|\gamma)}{\sum_{\gamma' \in \Gamma} p(\gamma')\pi(s|\gamma')}$ for all γ and s.

distributions of the signal, given all states, weighted by the prior belief. Hence, we can think of any mechanism as inducing a probability distribution over updated beliefs. Let $\tau(\mu_s(\gamma)) = \sum_s \sum_{\gamma'} p(\gamma')\pi(s \mid \gamma')$, for all $\mu_s(\gamma)$, represent a distribution over posteriors, given the information mechanism. Since the action of the agents depend on their beliefs, all that matters is the probability distributions about beliefs, which is summarised by $\tau(\mu_s(\gamma))$. From this, the problem of finding the optimal information policy reduces to directly choosing the optimal distribution of posteriors.

To simplify notations, let μ_s refer to the probability of $\gamma = \gamma_h$, which is potentially an updated posterior belief after observation of signal s, also p is the prior belief about such a state.

Here, we extend the work of Kamenica and Gentzkow (2011) to the case of multiple agents, where we examine a coalition model from the perspective of induced beliefs and information persuasion, and we derive the optimal information policies given different payoffs of the sender.

3.1 A coalition-formation setting

Consider a coalition formation model, where every agent $i \in I$ decides about joining a coalition. Let the number of agents, N, be a finite Natural number and $N \geq 2$. We restrict attention to single-coalition games. In other words, it is assumed that along an equilibrium path, only one coalition forms, and we do not study simultaneous formation of more than one coalition. This is of course in addition to the coalition of singletons, which is always a Nash equilibrium, and hence self-enforceable. Let us label the set of coalition members by M, and the set of non-members by F, where $M \subseteq I$, $F \subseteq I$, $M \cup F = I$, and $M \cap F = \emptyset$.

The game consists of two stages: the membership stage and the stage of decision about contribution to the public bad. The membership is open (voluntary), so the agents can freely join and no player can be excluded from joining by other coalition members. But the membership is fixed. In other words, once the agreement is reached it is implemented at no further cost[5].

Here we analyse the model from the perspective of information design, so we are interested in finding the optimal action of the agents for any belief. Indeed, in contrast to the literature, where beliefs are given parameters, and the profitability and stability of a coalition are studied with respect to the number of signatories of a treaty, in this paper, the threshold behaviour of the signatories and non-signatories of

[5] For example through legal devices, joining implies ratifying the treaty, which makes compliance compulsory, and signatories cannot leave the agreement

the agreement are examined with respect to their beliefs about the state variable. In other words, the stable number of signatories is a function of the endogenous beliefs.

4 The Solution

The model is solved according to the sender's-preferred subgame perfect equilibrium (SPE). In other words, if the agents are indifferent between two actions, they choose the sender's preferred action.[6] Specifically, given their beliefs, in the action stage, the agents' optimal decisions construct a vector of Nash strategies. The agents use pure strategies, as the only uncertainty is about the unknown state variable. Their action decisions form the expected payoff of (potential) signatories of the agreement and non-signatories, which determines the profitability of joining a coalition, and its stability in the membership stage. Due to the assumptions of fixed membership and single-coalition game, we check unilateral deviations. Hence, the membership decisions also form Nash strategies. Then, upon the optimal behaviour of the agents in action and membership stages, the sender designs an optimal information mechanism, which results in his most preferred outcome. The problems of the information receivers and the sender are respectively investigated in the next three subsections.

4.1 The action stage

4.1.1 Non-signatories' decision

Consider a coalition of n members, where $n \leq N$. In the action stage, a non-signatory takes the number of coalition members, and the action chosen by signatories and other non-signatories as given and individually maximises its expected payoff of $\mathbb{E}_\mu u_i^{fn}(\mathbf{q}, \gamma) = q_i^{fn} - \mathbb{E}_\mu(\gamma)[nq_i^m + \sum_{-i} q_{-i}^{fn} + q_i^{fn}]$, where superscript f is for the non-signatories or fringe agents and m for the coalition members, also superscript n shows the dependence of the payoff (or in future, other variables) to the number of coalition members. Thus superscript fn refers to the fringe where the number of coalition members is n. Furthermore, subscript $-i$ is the index for the other agent in the group (of non-signatories here).

By choosing $q_i^{fn} = 0$, the expected payoff of the fringe agent is $-\mathbb{E}_\mu(\gamma)[nq_i^m + \sum_{-i} q_{-i}^{fn}]$, and choosing $q_i^{fn} = 1$ leads to expected payoff of $1 - \mathbb{E}_\mu(\gamma)[nq_i^m + \sum_{-i} q_{-i}^{fn} + 1]$. Therefore, independent of the decision

[6] The tie-breaking rule is a standard trick, and as the problem is well-defined and the solution exists, the tie is broken in this way.

of coalition and other fringe agents, if a fringe agent believes that $\mu_s \geq \mu^f \equiv \frac{1}{\gamma_h}$, then it chooses $q_i^{*fn} = 0$, and if $\mu_s < \mu^f$, then $q_i^{*fn} = 1$ is chosen. Thus, the non-signatories have a dominant strategy which only depends on their belief about the social cost, and not the action undertaken by other agents, or number of coalition members, i.e. $q_i^{*fn} = q_i^{*f}$. Hence, in studying the underlying coalition formation game, both assumptions of Stackelberg and Cournot lead to the same result.

Hence, the coalition of singletons chooses prevention above μ^f, and below it, all individual agents choose contribution. Thus, we may refer to μ^f as the threshold of coalition of singletons. The threshold μ^f is positive and strictly less than one, i.e. $0 < \mu^f < 1$. Furthermore, it is clear that the larger γ_h is, the smaller the threshold belief of a fringe agent is, which leads to a prevention decision for a larger range of support of μ_s. Finally, μ^f is independent of the number of coalition members, and given the parameter value of γ_h, is fixed.

4.1.2 Signatories' decision

The coalition members in the action stage, act as a singleton. Given the action chosen by fringe agents, and the public belief, they share the same expected payoff.[7] In fact, they compare expected payoff of $-n\mathbb{E}_\mu(\gamma)\sum_i q_i^f$ by choosing $q_i^m = 0$ each, and $n - n\mathbb{E}_\mu(\gamma)[n + \sum_i q_i^f]$ by choosing $q_i^m = 1$. Therefore, independent of the action of non-signatories, they will have a common threshold of belief, $\mu^n \equiv \frac{1}{\gamma_h n}$, above which they prevent harm and below which they choose to contribute to the public bad. Specifically, if $\mu_s \geq \mu^n$, they prevent and if $\mu_s < \mu^n$, the signatories choose causing harm.[8] Their threshold is decreasing in γ_h and the number of coalition members. Furthermore, $0 < \mu^n \leq \mu^f$, where the equality of two thresholds occur if and only if $n = 1$.

4.1.3 Socially optimal decision

Before proceeding to the membership stage, let us verify the actions corresponding to the social optimum. The social-optimal action is a profile of strategies that maximises the sum of agents' expected payoffs. This is a benchmark, and later the equilibrium outcome of the model is compared with the social optimum. The next lemma verifies the social-optimal vector of actions, $\mathbf{q}^{so}$, which includes both imperfect (non-degenerate μ_s) and perfect-learning (degenerate μ_s) situations.

[7] For the symmetric agents, it is plausible to assume an equal sharing rule.

[8] In future, we examine different payoffs for the sender, and the equilibrium is always the sender's preferred SPE. Given the expected payoff of the sender, the tie-breaking rule of signatories and non-signatories may be revised, if necessary.

Lemma 1. *Given a common belief $\mu_s \in [0,1]$, the social optimum for all $i \in I$, implies selection of $q_i^{so} = 1$ for all posterior beliefs $0 \leq \mu_s < \mu^N \equiv \frac{1}{\gamma_h N}$, and selection of $q_i^{so} = 0$ for all beliefs $\mu^N \leq \mu_s \leq 1$.*

Proof. Examining the sum of expected payoffs under the two actions, implies an agent obtains expected payoff of $1 - N\gamma_h\mu_s$, if all agents choose to contribute to the public bad, and expected payoff of zero, if they all cooperate on prevention. This leads to the threshold behaviour specified in the lemma. □

In addition, $0 < \mu^N \leq \mu^n$, where these two thresholds are equal if and only if $n = N$. In fact, the socially optimal action strategies are equivalent to the action strategy of a potential grand coalition ($n = N$) at each level of beliefs. In other words, if a grand coalition can be formed for all levels of beliefs, then μ^N is the prevention threshold of such a coalition. Furthermore, as the number of agents increases, the socially optimal threshold, μ^N, decreases, implying that achieving the social optimum requires cooperation on prevention for a larger range of beliefs.

4.2 The membership stage

In the membership subgame, given the optimal decision and expected payoffs of the two groups of signatories and non-signatories in the action stage, the agents consider joining a coalition. This analysis consists of examining the profitability and self-enforceability (stability) of a coalition. Given the dominant strategies in the action stage, the sequence of membership decision is irrelevant, and we assume that the agents simultaneously choose their membership strategies.

A coalition is stable if the strategies in the membership stage construct a Nash equilibrium. Sufficient conditions for stability are the internal and external stability conditions, which are the Nash equilibrium conditions for signatories and non-signatories in the membership subgame. Recall that M is the set of coalition members, also let $n^*(\mu_s)$ be the expected number of signatories of the stable coalition, which depends on belief. The internal stability condition implies that no signatory has an incentive to leave the coalition, i.e. $\mathbb{E}_\mu u_i^{n^*}(M) \geq \mathbb{E}_\mu u_i^{n^*}(M \backslash \{i\})$, for all $i \in M$. The external stability refers to the condition that no non-signatory has an incentive to join the coalition, i.e. $\mathbb{E}_\mu u_i^{n^*}(M) > \mathbb{E}_\mu u_i^{n^*}(M \cup \{i\})$, for all $i \notin M$. Given the common beliefs, and the fact that all signatories have the same expected payoff, as well as all non-signatories, the internal stability implies that the size of coalition cannot be smaller than $n^*(\mu_s)$, otherwise the coalition becomes ineffective, i.e. $\mathbb{E}_\mu u_i^{mn^*}(n^*(\mu_s)) \geq \mathbb{E}_\mu u_i^f(n^*(\mu_s) - 1)$. Furthermore, the

external stability is reduced to the condition that the size of coalition cannot be greater than $n^*(\mu_s)$, otherwise deviation becomes profitable, i.e. $\mathbb{E}_\mu u_i^f(n^*(\mu_s)) > \mathbb{E}_\mu u_i^{mn^*}(n^*(\mu_s)+1)$.

To derive the stable number of members in the coalition formation subgame, the action strategies, and hence the expected payoffs of signatories and non-signatories for each level of beliefs must be examined. Given the ranking of thresholds of coalition of singletons and the grand coalition, $0 < \mu^N < \mu^f < 1$, the support of belief admits three distinct ranges.

First, if $\mu_s \geq \mu^f$, then both groups signatories and non-signatories optimally choose to prevent, i.e. $q_i^* = 0$ for all $i \in I$, and each agent gets expected payoff of zero. So in that range of posterior beliefs, no coalition is as good as full cooperation, which corresponds to the socially optimal action. Hence, in this range of beliefs the stable coalition is the coalition of singletons.

Second, for any belief $\mu_s < \mu^N$, all agents choose to contribute to the public bad, or $q_i^* = 1$ for all $i \in I$. As explained, μ^N is the preventing threshold of the grand coalition, and each agent receives expected payoff of $\mathbb{E}_\mu u_i^{mn^*} = 1 - \mu_s \gamma_h N$.

The final range of posterior beliefs is $\mu^N \leq \mu_s < \mu^f$. The threshold of signatories is between these two thresholds, i.e. for any $2 \leq n^*(\mu_s) \leq N$, the thresholds are $\mu^N \leq \mu^{n^*} < \mu^f$.[9] Given any $n^*(\mu_s)$, it may be profitable for the $n^*(\mu_s)$ members of the stable coalition to sign a treaty which specifies prevention for all members, while the fringe agents can be causing harm (if $n^*(\mu_s) < N$). Hence, the payoffs of signatories and non-signatories are $\mathbb{E}_\mu u_i^{mn^*} = -\mu_s \gamma_h (N - n^*(\mu_s))$ and $\mathbb{E}_\mu u_i^f = 1 - \mu_s \gamma_h (N - n^*(\mu_s))$, respectively. The stability of a coalition with size $n^*(\mu_s)$ is examined in the appendix, and the results of the membership stage are summarised in the next proposition. But, first let us zoom deeper into the range of beliefs where $\mu^N \leq \mu_s < \mu^f$.

Two facts can be directly verified: (i) the condition $\mu^{n^*} \leq \mu_s$ implies that a stable coalition of size $I(\frac{1}{\mu_s \gamma_h})$ can be formed, where $I(.)$ is the smallest integer which is no smaller than its argument; (ii) for any $2 \leq n^*(\mu_s) \leq N$, the threshold μ^{n^*} depends on $n^*(\mu_s)$. So, every $n^*(\mu_s)$ leads to a different threshold for the signatories. Facts (i) and (ii) imply that the coalition formation is endogenous, and there exists a mapping from the beliefs to the set of size of stable coalitions. Given that $n^*(\mu_s)$ is an integer, it is not a one-to-one mapping, but every range of beliefs between every two successive thresholds corresponds to a unique $n^*(\mu_s)$. Hence, the size of a stable coalition can be uniquely pinned down by choosing the

[9] It has already been discussed that if $n^*(\mu_s) = 1$, then $\mu^{n^*} = \mu^f$, which implies that the threshold of coalition of singletons is fixed at μ^f.

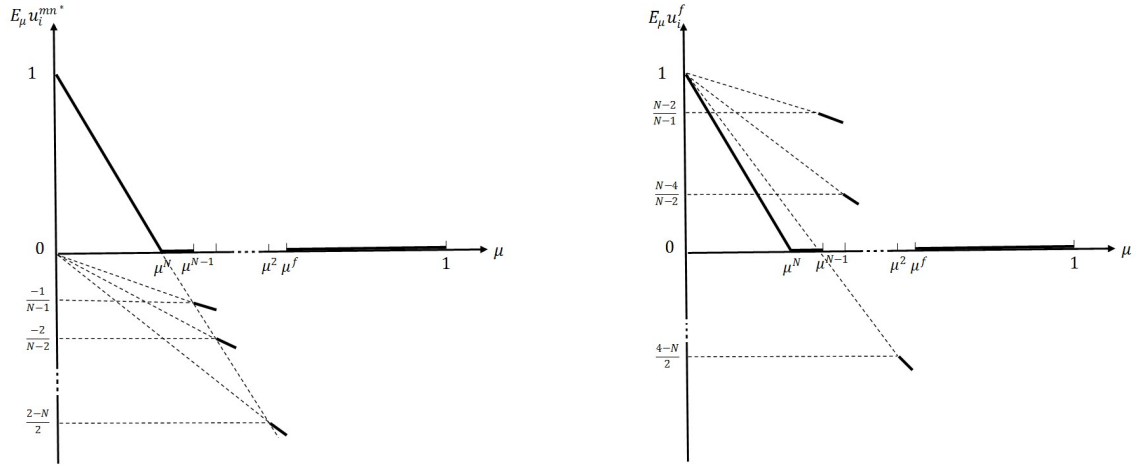
posterior beliefs.

For any belief between the thresholds of grand coalition and coalition of singletons, i.e. $\mu^N \leq \mu_s < \mu^f$, the model admits (possibly) finitely many thresholds for the signatories of different coalitions. Therefore, the intermediate partition of posterior beliefs is itself partitioned by smaller ranges, say sub-partitions, where in each sub-partition, the stable preventing coalition has a unique size. For any generic $2 \leq n^*(\mu_s) \leq N$, if $\mu^{n^*} \leq \mu_s < \frac{1}{\gamma_h(n^*(\mu_s)-1)}$, the stable coalition has $n^*(\mu_s)$ members. If $\mu_s < \mu^{n^*}$, the coalition of size $n^*(\mu_s)$ is not profitable, and the incentive for cooperation is cancelled out by the incentive to free ride, as the signatories leave such a coalition and make it worthless, i.e. $-\mu_s\gamma_h(N - n^*(\mu_s)) < 1 - \mu_s\gamma_h N$. Similarly, if $n^*(\mu_s) > 2$, for any belief $\frac{1}{\gamma_h(n^*(\mu_s)-1)} \leq \mu_s < \frac{1}{\gamma_h(n^*(\mu_s)-2)}$, the preventing coalition of size $n^*(\mu_s) - 1$ is stable, and so on.

Accordingly, the minimum threshold of signatories is μ^N, thus if $\mu^N \leq \mu_s < \frac{1}{\gamma_h(N-1)} \equiv \mu^{N-1}$, then the preventing grand coalition is stable. Also, the maximum threshold is $\frac{1}{2\gamma_h} \equiv \mu^2$, implying that for any belief $\mu^2 \leq \mu_s < \mu^f$, preventing coalition of size $n^*(\mu_s) = 2$ is stable. Clearly, for $N = 2$, the minimum and maximum signatories' thresholds coincide, implying that for any $\mu_s \geq \mu^N$, the coalition with $n^*(\mu_s) = 2$ is stable, where both agents cooperate on prevention. In that case, there would not be any difference between the equilibrium action strategies and the socially optimal strategies for any belief.

Hence, in the range of $\mu^N \leq \mu_s < \mu^f$, the expected number of members of the stable coalition, $n^*(\mu_s)$, is a (weakly) decreasing function of the posterior belief about high state, μ_s. The negative relationship of the social cost (parameter) and the number of members of a stable coalition is known from Barrett (1994), but in contrast to the literature, where belief is a constant parameter, and a single-coalition model admits one n^*, here we examine the situation from a design perspective, where the outcomes of different such games, with different stable coalitions are studied. If $\mu^N \leq \mu_s < \mu^f$, the endogenous size of stable coalitions, $n^*(\mu_s)$, varies from one sub-partition of beliefs to the other. As shown in the appendix, as belief about the high state increases, formation of smaller coalitions becomes profitable. However, this fact creates the free-riding incentive and makes only the smallest coalition internally stable.

Figure 4.1 depicts the expected payoffs of a representative signatory and non-signatory under stable membership strategies. The two panels are different only in the range of beliefs $\mu^{N-1} \leq \mu_s < \mu^f$. It can be verified that within this range of beliefs, for each $n^*(\mu_s)$, the difference in corresponding expected payoff for the two groups is constant at one. In other words, by fixing the beliefs in one sub-partition, the two groups have dominant strategies with respect to each other.



(a) (b)

Figure 4.1: The stable expected payoffs of a representative signatory, and non-signatory are illustrated respectively in panel (a) and (b) for $N > 4$.

Furthermore, note that if $\mu^{N-1} \leq \mu_s < \mu^f$, across the sub-partitions, as belief about high state decreases, and the number of signatories of the stable coalition increases, the expected payoffs of both signatories and non-signatories increase, as does the total payoff of $n^*(\mu_s)\mathbb{E}_\mu u_i^{*m} + (N - n^*(\mu_s))\mathbb{E}_\mu u_i^{*f}$. This is revisiting the "global efficiency" property in the literature, which refers to the positive relationship between the number of signatories and the total payoff. However, due to free-riding in the range of $\mu^{N-1} \leq \mu_s < \mu^f$, the actions undertaken by the agents do not satisfy the social optimum. While, if $0 \leq \mu_s \leq \mu^{N-1}$, or if $\mu^f \leq \mu_s \leq 1$, as explained in section 4.1.3, because all agents take similar actions, the action outcomes coincide with the social optimum.

The following proposition summarises our results on the stable strategies of the membership subgame.

Proposition 1. *(i) If $\mu_s < \mu^N$, then $q_i^* = 1$ for all $i \in I$. In this range of beliefs, prevention is not profitable for any coalition.*

(ii) If $\mu^N \leq \mu_s < \mu^f$, there are $N-1$ thresholds of signatories, μ^{n^*}, and for any $2 \leq n^*(\mu_s) \leq N$, the beliefs in between every two successive thresholds, $\mu^{n^*} \leq \mu_s < \frac{1}{\gamma_h(n^*(\mu_s)-1)}$, map to the unique $n^*(\mu_s)$, where $q_i^{*m} = 0$ and (if $n^*(\mu_s) < N$) $q_i^{*f} = 1$. In this range of beliefs, the coalition of singletons which contributes is also stable.

(iii) If $\mu_s \geq \mu^f$, then the preventing coalition of singletons is stable, where $q_i^* = 0$ for all $i \in I$.

For the proof see Appendix 8.1.

4.3 Sender's persuasion

Before the membership stage, the sender can communicate an information policy and conduct research, which leads to sending a signal, before any decision about the membership is made by the agents. Indeed, the sender, given the best response of the agents in the action and membership stages, chooses a lottery over posterior beliefs such that it results in his most preferred stable coalition sizes and action strategies. To fix ideas, in this section, by "equilibrium", we refer to the sender's-preferred SPE, and "stable" coalition, as explained above, refers to any coalition which satisfies the internal and external stability conditions, and may or may not be the "equilibrium" outcome.

Recall that the sender and the agents share a common prior. Even if the agents can predict the optimal information mechanism, the signal is a random draw according to the information policy. Furthermore, the sender does not know yet which agent is going to be a signatory and which one will be a fringe. Hence, we focus on public signals. Since, each group of signatories and non-signatories have different thresholds, the sender faces a problem where the receivers of the public signal (may) have different payoffs, and accordingly different best-response actions.

Upon the fact that the country's belief is μ_s, the expected value of sender's payoff for a given belief, can be written directly as a function of μ_s. Let $\nu(\mu_s) = \mathbb{E}_\mu \boldsymbol{\nu}(\mathbf{q}(\gamma))$.

To derive the feasible subset of policies, a necessary condition is that the expected value of the posteriors over τ, must be equal to the prior, i.e. $\mathbb{E}_\tau \mu = p$, which is the law of total probability. Kamenica and Gentzkow (2011) refer to it as the Bayes-plausibility condition, and they show that this is the only restriction. In other words, this canonical subset of policies captures all possibilities. Indeed a distribution τ over posterior beliefs can be generated by an information mechanism if and only if it satisfies the law of total probability.

Therefore, in our binary setting, the sender selects a binary signal space, which implies that there is a one-to-one relationship between the signal space and the agents' action space. Hence, the sender by sending a signal, recommends a certain action, and the agents accordingly take the corresponding action. This is known as "direct" information mechanism, and in fact, the sender could not do better than that. The Bayes-plausibility of the signals implies that the agents follow the recommendation of the sender.

Let $S = \{0, 1\}$, where $s = 0$ refers to state γ_l, and $s = 1$ corresponds to state γ_h. Hence, for example

signal $s = 1$ leads to prevention action, $q_i = 0$. So, the notation μ_1 refers to the probability that the state is high conditional on receiving signal $s = 1$. Furthermore, let τ represent the probability of μ_1.

Because each belief is associated with a payoff, and as the sender by choosing an information policy is going to choose a distribution τ, every policy yields the expected payoff of $\tau\nu(\mu_1) + (1-\tau)\nu(\mu_0)$ for the sender. So any information policy can be represented by τ and we can write the expected payoff of the sender as a function of τ.

So, the sender by choosing τ maximises

$$\begin{aligned} &\tau\nu(\mu_1) + (1-\tau)\nu(\mu_0) \\ &\text{subject to } \tau\mu_1 + (1-\tau)\mu_0 = p \end{aligned} \tag{4.1}$$

The solution can be illustrated in $(\mu, \nu)-$space. Let us denote $V(\mu) \equiv \max_\tau[\tau\nu(\mu_1) + (1-\tau)\nu(\mu_0) \mid \tau\mu_1 + (1-\tau)\mu_0 = p]$, which given any Bayes-plausible lottery over beliefs, represents the maximum expected payoff over the beliefs. The function $V(\mu)$ is the supremum of convex hull of the graph of $\nu(\mu)$, or the smallest concave function that is no smaller than $\nu(\mu)$ at every belief. According to Kamenica and Gentzkow (2011) and Aumann and Maschler (1995), this determines the optimal policy for a given support of the beliefs. In addition, given the constraint of the optimisation, the lottery over beliefs should be on average equal to the prior, p. So, it is possible to read off the optimised value of the information policy, which implements a desired coalition size and action regarding the public bad, as $V(p)$.

In the next two applications, we examine the problem of finding the optimal information mechanism under different preference specifications for the sender. We examine each case in turn.

5 Application 1: Supporting potential signatories or non-signatories

In the next two subsections, we consider cases where the preferences of the sender coincide with the signatories and non-signatories, respectively. However, it is shown that no matter whether the sender has the same preferences as the signatories or non-signatories of the coalition, for any $p \in [0, 1]$, the unique optimal public information policy, is full revelation of the state of the world.

Then, the results are generalised to a case where the expected payoff of the sender is a combination of the expected payoffs of the two groups of coalition members and fringe agents, and it is shown that the unique information policy is again perfect learning.

5.1 Sender with the same preference as the potential signatories

Suppose the expected payoff of the sender coincides with the expected payoff of a representative signatory, then upon the fact that the belief of agents about the high state is μ_s,

$$\nu(\mu_s) = \mathbb{E}_\mu u_i^{mn^*}(\mathbf{q}^*(\gamma)) = \begin{cases} 0 & \text{if } \mu^f \leq \mu_s \leq 1 \\ -\mu_s\gamma_h(N - n^*(\mu_s)) & \text{if } \mu^{n^*} < \mu_s \leq \frac{1}{(n^*(\mu_s)-1)\gamma_h} \\ 1 - \mu_s\gamma_h N & \text{if } 0 \leq \mu_s \leq \mu^N \end{cases} \tag{5.1}$$

where $2 \leq n^*(\mu_s) \leq N$, and for the case of $n^*(\mu_s) = 2$, the range of beliefs does not include any thresholds, i.e. $\mu^2 < \mu_s < \mu^f$, to ensure the existence of the sender's-preferred equilibrium.

As explained, the sender chooses an optimal distribution of posteriors, τ, which satisfies the law of total probability and maximises his payoff as described in (5.1). This in turn reduces the problem to finding a lottery over the common belief μ. The problem is analysed formally in Appendix 8.2.[10] Furthermore, from panel (a) of Figure 4.1, it can be verified that the unique smallest concave function, which is no smaller than $\nu(\mu)$ is a straight line connecting $\mu = 0$ and $\mu = 1$, which implies that for any interior prior, the full revelation is the unique optimal mechanism.

Lemma 2. *If the payoff of the sender is $\nu(\mu_s) = \mathbb{E}_\mu u_i^{mn^*}(\mathbf{q}^*(\gamma))$, then for any prior belief $p \in (0,1)$, the unique optimal conditional probability of signals are $\pi(1 \mid \gamma_h) = 1$ and $\pi(0 \mid \gamma_l) = 1$.*

In addition, the convexity in $\mathbb{E}_\mu u_i^{mn^*}(\mu_s)$ implies that for any $p \in (0,1)$, $V(p)$, which is the value of optimal mechanism, is greater than $\nu(p)$, which is the sender's expected payoff in the absence of any persuasion. In other words

Corollary 1. *If the payoff of sender is $\nu(\mu_s) = \mathbb{E}_\mu u_i^{mn^*}(\mathbf{q}^*(\gamma))$, then for any prior belief $p \in (0,1)$, the sender strictly benefits from the persuasion.* [11]

[10] The proofs of the following lemmas and propositions are provided for the general case of $N > 3$, where there are at least three sub-partitions between the thresholds of grand coalition and coalition of singletons. Clearly, the proofs for cases of $N = 2$ and $N = 3$ are subsets of the proofs for case $N > 3$.

[11] If the prior is degenerate, i.e $p = 1$, when $\gamma = \gamma_h$, and $p = 0$ when $\gamma = \gamma_l$, then in fact the sender need not do anything. But formally the optimal mechanism is not unique is such cases, as choosing the full-revelation policy is still optimal, although the sender does not benefit from it.

5.2 Sender with the same preference as the potential non-signatories

In this section, we check the sensitivity of our results to the preference of the sender over non-signatories and signatories. Assume before the membership decision, the sender chooses an information policy to maximise the expected payoff of a potential representative non-signatory, in other words,

$$\nu(\mu_s) = \mathbb{E}_\mu u_i^f(\mathbf{q}^*(\gamma)) = \begin{cases} 0 & \text{if } \mu^f \leq \mu_s \leq 1 \text{ or if } \mu^N < \mu_s < \mu^{N-1} \\ 1 - \mu_s\gamma_h(N - n^*(\mu_s)) & \text{if } \mu^{n^*} < \mu_s \leq \frac{1}{(n^*(\mu_s)-1)\gamma_h} \\ 1 - \mu_s\gamma_h N & \text{if } 0 \leq \mu_s \leq \mu^N \end{cases} \tag{5.2}$$

where in this equation by $n^*(\mu_s)$ we refer to any $2 \leq n^*(\mu_s) \leq N-1$. Furthermore, the tie-breaking rule is such that for the case of $n^*(\mu_s) = 2$, the range of beliefs is $\mu^2 < \mu_s < \mu^f$, and for the case of $n^*(\mu_s) = N-1$, the range of beliefs is $\frac{1}{(N-1)\gamma_h} \leq \mu_s \leq \mu^{N-2}$.

It turns out that the optimal information mechanism is the same as if the expected payoff of signatories was chosen. In fact, the following lemma and corollary are parallel to the results in the previous subsection.

Lemma 3. *If the payoff of the sender is $\nu(\mu_s) = \mathbb{E}_\mu u_i^f(\mathbf{q}^*(\gamma))$, then for any prior belief $p \in (0,1)$, the unique optimal conditional probability of signals are $\pi(1 \mid \gamma_h) = 1$ and $\pi(0 \mid \gamma_l) = 1$.*

See Appendix 8.3 for the proof of lemma 3. Panel(b) of Figure 4.1 also confirms the results. From the figure, it can be verified that for all parameter values of the model, and all interior posterior beliefs, $\mathbb{E}_\mu u_i^f(\mu)$ is below $V(\mu)$. In other words,

Corollary 2. *If the payoff of the sender is $\nu(\mu_s) = \mathbb{E}_\mu u_i^f(\mathbf{q}^*(\gamma))$, then for any prior belief $p \in (0,1)$, the sender strictly benefits from the persuasion.*

Given the last two lemmas, for both cases that the sender maximises the expected payoff of signatories or non-signatories, the optimal information mechanism is full revelation of the state of the world. Accordingly, any non-degenerate μ_s is out of the equilibrium path, and the action equilibrium outcome will be reduced to selection of prevention for all $i \in I$ if $\mu_1 = 1$, also contributing to the public bad for all $i \in I$, where $\mu_0 = 0$. Thus, the action vectors in the action stage coincide with the social optimum, defined in lemma 1, even though the sender maximises the expected payoff of one group.

Proposition 2. *If $\nu(\mu_s) = \mathbb{E}_\mu u_i^f(\mathbf{q}^*(\gamma))$, or if $\nu(\mu_s) = \mathbb{E}_\mu u_i^{mn^*}(\mathbf{q}^*(\gamma))$, the optimal information mechanism of full revelation leads to the socially optimal action outcome.*

Finally, in both cases, the equilibrium (action strategies, membership strategies, and the information mechanism) is independent of the level of γ_h, and the prior p.

5.3 Sender and a combination of preferences of both signatories and non-signatories

Suppose

$$\nu(\mu_s) = \alpha n^*(\mu_s)\mathbb{E}_\mu u_i^{mn^*}(\mathbf{q}^*(\gamma)) + (N - n^*(\mu_s))\mathbb{E}_\mu u_i^f(\mathbf{q}^*(\gamma)) \tag{5.3}$$

where α is a constant and let $\alpha \geq 1$. In other words, assume that the sender maximises the summation of expected payoffs of both groups of signatories and non-signatories, and it may weight signatories' preferences more, knowing that $n^*(\mu_s)$ varies over different partitions of beliefs.

Proposition 3. *If the expected payoff of the sender is* $\nu(\mu_s) = \alpha n^*(\mu_s)\mathbb{E}_\mu u_i^{mn^*}(\mathbf{q}^*(\gamma))+(N-n^*(\mu_s))\mathbb{E}_\mu u_i^f(\mathbf{q}^*(\gamma))$, *then for any* $\alpha \geq 1$, *and any prior* $p \in (0,1)$, *the unique optimal information mechanism is full revelation of the state, and the equilibrium action strategies coincide with the socially optimal outcome.*

Formal proof is in Appendix 8.4. Intuitively, if $0 \leq \mu_s < \mu^{N-1}$, or if $\mu^f \leq \mu_s \leq 1$, then the sender maximises the summation of expected payoffs of the grand coalition, or equivalently, the expected payoff of a representative agent[12]. While, (for the case of $N > 2$) if beliefs belong to, $\mu^{N-1} \leq \mu_s < \mu^f$, then the expected payoffs of signatories and non-signatories are different. Indeed, the expected payoff of signatories in range of beliefs $\mu^{N-1} \leq \mu_s < \mu^f$, is negative, and the maximum possible expected payoff of non-signatories is $1-\mu_s\gamma_h$, where $n^*(\mu_s) = N-1$. Given that the posterior beliefs on average should be equal to the prior, in lemma 3, it was shown that the maximum value of the expected payoff of a representative non-signatory, $1-p\gamma_h$, is below $1-p$ for interior prior beliefs, specifically, $1-p\gamma_h < 1-p$. In this problem, perfect learning prescribed by the optimal information mechanism, i.e. $\mu_0 = 0$ and $\mu_1 = 1$, suggests that $V(p) = N(1-p)$. Thus, in the current problem, the summation of maximum value of expected payoffs of non-signatories is clearly below $V(p)$ of the optimal mechanism, i.e. $1 - p\gamma_h < N(1-p)$. Accordingly, the maximum value of expected payoff of both signatories and non-signatories is below $V(p) = N(1-p)$, ensuring that for no prior, the sender could benefit from a partially uninformative policy, also, the sender gains from the optimal policy of full revelation for all interior prior beliefs.

[12]The tie-breaking rule is for $N > 2$.

6 Application 2: IEA and minimising the total level of greenhouse gases

In this section, we apply the results to IEA in climate change. Hence, the public bad can be interpreted as the GHG, where the agents' (countries') membership and emission decisions depend on the social cost of GHG as the state variable. A central authority, as the information sender, conducts research on the true social cost of climate change, and releases information to the agents. Here a case is studied, where the sender's objective is minimising the total level of GHG. If the main concerns of the sender are the global consequences of the emission of agents, for example if the incentives of the sender are driven from the affect of increased GHG on the average temperature of the planet, e.g. meeting the two-degree threshold, or if the preferences of the sender reflect the wide difference of social and private discount factors, then it is reasonable to assume that the sender minimises the total level of public bad. Clearly, in such a situation, the expected payoff of the sender, does not coincide with the payoff of the agents.

Assume that the sender's objective is minimising the total level of public bad, or maximising $\boldsymbol{\nu}(\mathbf{q}(\gamma)) = -Q$. So, given the best-response strategies of the agents in the action and membership stages, the expected payoff of the sender can be written as

$$\nu(\mu_s) = \begin{cases} 0 & \text{if } \mu^f \leq \mu_s \leq 1 \\ n^*(\mu_s) - N & \text{if } \mu^{n^*} < \mu_s \leq \frac{1}{(n^*(\mu_s)-1)\gamma_h} \\ -N & \text{if } 0 \leq \mu_s < \mu^N \end{cases} \tag{6.1}$$

where in this equation by $n^*(\mu_s)$ we refer to any $2 \leq n^*(\mu_s) \leq N$. Furthermore, the tie-breaking rule is such that for the case of $n^*(\mu_s) = 2$, the range of beliefs is $\mu^2 < \mu_s < \mu^f$, and for the case of $n^*(\mu_s) = N$, the range of beliefs is $\mu^N \leq \mu_s \leq \mu^{N-1}$.

The supremum of convex hull of graph of $\nu(\mu)$ suggests that the optimal information policy is not globally unique, and depending on p, it may take a form of partial learning, either by selection of a non-degenerate lottery over posteriors, or leaving the agents at their prior beliefs.

Proposition 4. *If $\boldsymbol{\nu}(\mathbf{q}(\gamma)) = -Q$, the optimal information mechanism prescribes imperfect learning for $p \in (0, 1)$, where*

(i) if $\mu^f \leq p < 1$, or if $\mu^N \leq p \leq \mu^{N-1}$, then a degenerate lottery over posteriors which is equal to the

prior with probability one, i.e. $\mu_1 = p$ *and* $\tau = 1$*, is an optimal information policy. This is equivalent to communication of no signal.*

(ii) if $0 < p < \mu^N$*, then the unique optimal policy consists of a Bayes-plausible randomisation over* $\mu_0 = 0$ *and* $\mu_1 = \mu^N$*.*

(iii) for any $\mu^N \leq p < 1$*, the information policy of a Bayes-plausible randomisation over* $\mu_0 \in [\mu^N, \mu^{N-1}]$ *and* $\mu_1 \in [\mu^f, 1]$*, is one of the optimal information policies.*

The proof is in Appendix 8.5, and it includes all possible optimal information policies for different prior beliefs. In contrast to the other expected payoffs of the sender, which were examined in application 1, here the optimal information policy may lead to imperfect learning.

If $\mu^f \leq p < 1$, or if $\mu^N \leq p \leq \mu^{N-1}$, the agents in the absence of any persuasion, choose the sender's preferred action, which is prevention by all agents. Hence, in such cases, sending no signal is an optimal information policy.

In situations where non-signatorie(s) contribute to the public bad, and signatories do not, i.e. if $\mu^{N-1} < p < \mu^f$, all possible optimal policies include selection of a Bayes-plausible lottery over posteriors from $\mu_0 \in [\mu^N, \mu^{N-1}]$ (where a preventing grand coalition is formed) and $\mu_1 \in [\mu^f, 1]$ (where all agents coordinate on preventing).

Furthermore, if $0 < p < \mu^N$, by selection of a randomisation over the minimum posterior associated with the preventing grand coalition, i.e. μ^N, and the posterior of $\mu_0 = 0$, the sender strictly gains. Specifically,

Corollary 3. *For any* $0 < p < \mu^N$ *and* $\mu^{N-1} < p < \mu^f$*, the sender strictly benefits from the persuasion. Furthermore, for all* $p \in [0, 1]$*, the sender's-preferred action and membership SPE coincides with the socially optimal outcome.*

Hence, in cases where sending no signal is an optimal information policy, the sender does not benefit from persuasion, while for all other interior prior beliefs, the sender strictly gains from persuasion.

In terms of equivalence of the equilibrium outcome with the social optimum, here although the sender's expected payoff does not coincide with the preferences of (any or none of) the agents, and may lead to partial learning, minimising the total level of public bad by the sender, for all interior prior beliefs, results in the socially optimal action outcome, as specified in lemma 1. This is in fact because in all cases of the two applications, the assumed expected payoffs of the sender increases with the expected number of

signatories of the coalition.

According to our model, formation of the grand coalition in Paris, with the focus on minimising the total level of GHG, could be the result of the adoption of two communication strategies by the IPCC and other partners of the UN. Given the fact that in contrast to previous meetings, from the beginning of the Paris conference, the leaders of many of participating states were optimistic about achieving an agreement, it can be said that the agents' prior belief about the social cost of GHG had already implied believing in a high level of social cost of GHG.[13] If this was the case, then our simple model suggests that one possible optimal information policy would be sending no signal. In fact, the strategy of IPCC from months prior to the conference, was not to communicate the social cost of GHG, although prominent research was carried out by the IPCC on the social cost of carbon.[14] The conference negotiations in December 2015 were mainly focused around the contributions of natural sciences to the level of GHG and the average global temperature. Another possible communication policy could be that the international research authorities involved, referring to the wide uncertainty about the state variable, optimally using a randomisation of signals, which corresponds to the posterior beliefs which lead to the formation of the preventing grand coalition.

7 Conclusions

The question of the possibility of achieving a grand coalition or social optimum in collective decision-making has important policy implications. The paper suggests a new approach affecting the membership decision of agents in coalition formation, which is a central question in the literature. Here, the problem of coalition formation is examined from the perspective of information design, where the designer selects an optimal information structure to implement a desirable coalition.

A theoretical model is developed, where a research central authority, as an information sender, communicates an information policy and signals about a payoff-relevant state variable to the agents. In order to affect the decisions of the agents regarding contribution to a public bad, the sender persuades the agents to form a desirable coalition. Through the updated beliefs of the agents about the state variable (the social cost of public bad), the communication by the sender, who has commitment power, determines

[13]The high prior belief of the agents could be a result of private pre-communication with many individual countries from months before the conference.

[14] This is according to a public lecture at the University of Edinburgh by professor Ottmar Edenhofer, co-chair of Working Group III of the IPCC, in May 2015.

the decisions of the signatories and non-signatories to the agreement.

Specifically, in a coalition formation model where the beliefs are not exogenous variables, it is shown that the equilibrium strategies of signatories and non-signatories take a threshold form with respect to their beliefs, and this results in partitioning the support of beliefs such that every partition of beliefs maps to a unique coalition size. Hence, the size of the stable coalition, as an endogenous variable, can be uniquely pinned down by choosing the posterior beliefs. Knowing this, the sender designs an information policy to maximises its expected payoff.

As a direction for future research, the current coalition model is simplified in various dimensions. A possible generalisation could be extension of the model to a dynamic setting, where the coalition formation can incorporate variable membership, and the optimal time of communication and delay can be derived. Again, this has been proved to be an important issue in practice, and is given attention in the theoretical literature. Furthermore, private information acquisition by the agents, private persuasion versus public persuasion by the sender, and costly research for the sender are other pathways to generalise the analysis.

8 Appendix

8.1 Proof of proposition 1

First, the coalition of singletons is always stable, as no unilateral deviation can make anyone better off. Accordingly, the resultant outcome of the coalition of singletons is the same as that derived for the fringe agents, i.e. $q_i^* = 1$ for any $\mu_s < \mu^f$, and $q_i^* = 0$ for any $\mu_s \geq \mu^f$. Now we continue the proof for $n^*(\mu_s) \geq 2$ and the three different ranges of beliefs.

i) If $\mu_s < \mu^N$, choosing $q_i^* = 1$, by all $i \in I$, leads to expected payoff of $\mathbb{E}_\mu u_i^* = 1 - \mu_s \gamma_h N$ for agent i. No unilateral deviation can make anyone better off, i.e. $1 - \mu_s \gamma_h N > -\mu_s \gamma_h (N-1)$, which implies that there is no stable preventing coalition in this range of beliefs.

ii) Consider the case where $\mu^N \leq \mu_s < \mu^f$. Fix $n^*(\mu_s)$, also assume $\mu^{n^*} \leq \mu_s < \frac{1}{(n^*(\mu_s)-1)\gamma_h}$. We claim that in this range of beliefs, there is a stable coalition of unique size $n^*(\mu_s) = I(\frac{1}{\mu_s \gamma_h})$, where $q_i^{*m} = 0$, and $q_i^{*f} = 1$.

If $n \geq n^*(\mu_s)$, then $\mathbb{E}_\mu u_i^{mn^*} = -\mu_s \gamma_h (N-n)$, and $\mathbb{E}_\mu u_i^f = 1 - \mu_s \gamma_h (N-n)$. While if $n < n^*(\mu_s)$, farsightedness of coalition members implies $\mathbb{E}_\mu u_i^f = 1 - \mu_s \gamma_h N$.

The profitability of prevention for every coalition member implies that $-\mu_s \gamma_h (N - n^*(\mu_s)) \geq 1 - \mu_s \gamma_h N$.

This is always satisfied as in this range, $\mu_s \geq \frac{1}{n^*(\mu_s)\gamma_h} \equiv \mu^{n^*}$.

In addition, the external stability condition implies that the stable number of signatories cannot be greater than $n^*(\mu_s)$, i.e.

$$\mathbb{E}_\mu u_i^{mn^*}(n^*(\mu_s)+1) < \mathbb{E}_\mu u_i^f(n^*(\mu_s)) \tag{8.1}$$

This is satisfied if and only if $-\mu_s\gamma_h(N-n^*(\mu_s)-1) < 1-\mu_s\gamma_h(N-n^*(\mu_s))$, which is the case as in this range $\mu_s < \frac{1}{\gamma_h} \equiv \mu^f$. In other words, it does not pay any non-signatory to change its decision.

By crossing the threshold of $\frac{1}{(n^*(\mu_s)-1)\gamma_h}$, a smaller preventing coalition with $n^*(\mu_s)-1$ members is also profitable. But as explained in section 4.2, the internal stability condition implies that the stable number of signatories is determined by the smallest possible number of members. Therefore, if $\mu^{n^*} \leq \mu_s < \frac{1}{(n^*(\mu_s)-1)\gamma_h}$, no coalition member has an incentive to leave the coalition of size $n^*(\mu_s)$. Similarly, (in case that $n^*(\mu_s) > 2$) if $\frac{1}{(n^*(\mu_s)-1)\gamma_h} \leq \mu_s < \frac{1}{(n^*(\mu_s)-2)\gamma_h}$, then preventing coalition of size n^*-1 is stable, which results in expected payoff of $\mathbb{E}_\mu u_i^{mn^*} = -\mu_s\gamma_h(N-n^*(\mu_s)+1)$ for the signatories of the coalition.

If $\mu_s < \mu^{n^*}$, then a coalition of $n^*(\mu_s)$ with decision of $q_i^{*m}=0$ is not internally stable and signatories have an incentive to leave the coalition as $-\mu_s\gamma_h(N-n^*(\mu_s)) < 1-\mu_s\gamma_h N$.

Therefore, in the range $\mu^N \leq \mu_s < \mu^f$, given that the minimum threshold of signatories is μ^N, and (in the case that $N>2$,) the maximum threshold is μ^2, there exists $N-1$ thresholds for signatories, μ^{n^*}, and accordingly $N-1$ sub-partitions of beliefs, $\mu^{n^*} \leq \mu_s < \frac{1}{(n^*(\mu_s)-1)\gamma_h}$. Thus, the size of the stable coalition varies between two and N, i.e. $2 \leq n^*(\mu_s) \leq N$.

iii) Now suppose $\mu_s \geq \mu^f$. Then, causing harm is not profitable, and therefore it is not internally stable, i.e. $1-\mu_s\gamma_h N < 0$. Checking the external stability condition is superfluous here.

8.2 Proof of lemma 2

First, recall that the law of total probability implies $\tau = \frac{p-\mu_0}{\mu_1-\mu_0}$. Therefore, $\frac{\partial\tau}{\partial\mu_1} = \frac{\mu_0-p}{(\mu_1-\mu_0)^2}$, which is always non-positive, as $\mu_0 \leq p$. In addition, $\frac{\partial\tau}{\partial\mu_0} = \frac{p-\mu_1}{(\mu_1-\mu_0)^2}$, this is also non-positive, as $p \leq \mu_1$. The corner solution, which is found in most of the following problems, relies on these properties. Furthermore, using direct mechanism and our payoff structures imply that on the equilibrium path, $\mu_0 \leq \mu_1$.

Given proposition 1 and the corresponding equilibrium payoff of signatories in (5.1), the problem of the sender as specified in (4.1) can be written as maximising

$$\tau \begin{cases} 0 & \text{if } \mu^f \leq \mu_1 \leq 1 \\ -\mu_1\gamma_h(N - n^*(\mu_1)) & \text{if } \mu^{n^*} < \mu_1 \leq \frac{1}{(n^*(\mu_1)-1)\gamma_h} \\ 1 - \mu_1\gamma_h N & \text{if } 0 \leq \mu_1 \leq \mu^N \end{cases} \tag{8.2}$$
$$+ (1-\tau) \begin{cases} 0 & \text{if } \mu^f \leq \mu_0 \leq 1 \\ -\mu_0\gamma_h(N - n^*(\mu_0)) & \text{if } \mu^{n^*} < \mu_0 \leq \frac{1}{(n^*(\mu_0)-1)\gamma_h} \\ 1 - \mu_0\gamma_h N & \text{if } 0 \leq \mu_0 \leq \mu^N \end{cases}$$

with respect to μ_1 and μ_0, subject to $\tau\mu_1 + (1-\tau)\mu_0 = p$. Also, as mentioned, $2 \leq n^*(\mu_s) \leq N$, and for the case of $n^*(\mu_s) = 2$, the range of beliefs is $\mu^2 < \mu_s < \mu^f$. Moreover, the expected payoff in (8.2) can be further decoded to

$$\begin{aligned}
&\tau\nu(\mu_1) + (1-\tau)\nu(\mu_0) = [(1-\tau)(1-\mu_0\gamma_h N)]\mathbf{1}_{\substack{\mu^f \leq \mu_1 \leq 1 \\ 0 \leq \mu_0 \leq \mu^N}} \\
&- [(1-\tau)\mu_0\gamma_h(N - n^*(\mu_0))]\mathbf{1}_{\substack{\mu^f \leq \mu_1 \leq 1 \\ \mu^{n^*} < \mu_0 \leq \frac{1}{(n^*(\mu_0)-1)\gamma_h}}} \\
&+ [0]\mathbf{1}_{\substack{\mu^f \leq \mu_1 \leq 1 \\ \mu^f \leq \mu_0 \leq 1}} + [(1-\tau)(1-\mu_0\gamma_h N) - \tau\mu_1\gamma_h(N - n^*(\mu_1))]\mathbf{1}_{\substack{\mu^{n^*} < \mu_1 \leq \frac{1}{(n^*(\mu_1)-1)\gamma_h} \\ 0 \leq \mu_0 \leq \mu^N}} \\
&- [\tau\mu_1\gamma_h(N - n^*(\mu_1)) + (1-\tau)\mu_0\gamma_h(N - n^*(\mu_0))]\mathbf{1}_{\substack{\mu^{n^*} < \mu_1 \leq \frac{1}{(n^*(\mu_1)-1)\gamma_h} \\ \mu^{n^*} < \mu_0 \leq \frac{1}{(n^*(\mu_0)-1)\gamma_h}}} \\
&- [\tau\mu_1\gamma_h(N - n^*(\mu_1))]\mathbf{1}_{\substack{\mu^{n^*} < \mu_1 \leq \frac{1}{(n^*(\mu_1)-1)\gamma_h} \\ \mu^f \leq \mu_0 \leq 1}} \\
&+ [\tau(1 - \mu_1\gamma_h N) + (1-\tau)(1-\mu_0\gamma_h N)]\mathbf{1}_{\substack{0 \leq \mu_1 \leq \mu^N \\ 0 \leq \mu_0 \leq \mu^N}} \\
&+ [\tau(1 - \mu_1\gamma_h N) - (1-\tau)\mu_0\gamma_h(N - n^*(\mu_0))]\mathbf{1}_{\substack{0 \leq \mu_1 \leq \mu^N \\ \mu^{n^*} < \mu_0 \leq \frac{1}{(n^*(\mu_0)-1)\gamma_h}}} \\
&+ [\tau(1 - \mu_1\gamma_h N)]\mathbf{1}_{\substack{0 \leq \mu_1 \leq \mu^N \\ \mu^f \leq \mu_0 \leq 1}}
\end{aligned} \tag{8.3}$$

In order to prove the lemma, and find the unique optimal mechanism, it is required to compare the maximised expected payoff of all of the above nine partitions of posterior beliefs with each other, and select the policy corresponding to the maximum expected payoff.

Let us label each term of the above expected payoff by ascending numbers, e.g. case 1 refers to the

first term where $\mu^f \leq \mu_1 \leq 1$ and $0 \leq \mu_0 \leq \mu^N$.

It is easy to verify that selection of posterior beliefs corresponding to cases 6, 8, and 9, where $\mu_0 > \mu_1$, are dominated as the persuasion would be worthless. For example, in case 9, where $0 \leq \mu_1 \leq \mu^N$ and $\mu^f \leq \mu_0 \leq 1$. The resulted expected payoff of $\tau(1 - \mu_1\gamma_h N)$ is maximised if $\mu_1 = 0$, to obtain expected payoff of τ. Then, the constraint of law of total probability implies $1 - \frac{p}{\mu_0} = \tau$. So the maximum τ is achieved by setting $\mu_0 = 1$. Hence $\pi(1 \mid \gamma_h) = 0$ and $\pi(0 \mid \gamma_l) = 0$, which is absolute lying and such a policy will be ignored by the agents.

In addition, in cases that the corresponding expected payoff depends on $n^*(\mu_s)$, it is an increasing function of $n^*(\mu_s)$. Therefore, selection of posteriors which lead to $n^*(\mu_s) = N$ maximises the sender's payoff. Furthermore, the expected payoffs of cases 2, 3, and 5 are inferior relative to cases 1, 4, and 7, which lead to positive expected payoffs. Now we compare the maximum possible values of these positive candidates.

Case 1: $\mu^f \leq \mu_1 \leq 1$ and $0 \leq \mu_0 < \mu^N$. Then, the expected payoff of $(1-\tau)(1-\mu_0\gamma_h N)$ is maximised if $\mu_0 = 0$. Thus, the law of total probability implies $\tau = \frac{p}{\mu_1}$. Therefore, the resulted expected payoff of $(1 - \tau)$ is maximised if τ is minimised, which is achieved by choosing $\mu_1 = 1$. Hence, $V(p) = 1 - p$ is a potential value of the optimal mechanism.

Case 4: $\mu^{n^*} < \mu_1 \leq \frac{1}{(n^*(\mu_1)-1)\gamma_h}$ and $0 \leq \mu_0 < \mu^N$. To maximise the associated expected payoff, it should be that $\mu_0 = 0$. Also, ideally, μ_1 should be such that $n^*(\mu_1) = N$. In other words, $\mu^N < \mu_1 \leq \mu^{N-1}$. Hence, the resulted expected payoff, which is decreasing in τ, so increasing in μ_1, is maximised if $\mu_1 = \mu^{N-1}$. However, first, it may not be Bayes-plausible if $p > \mu^{N-1}$. Second, $\mu_1 = \mu^{N-1}$ leads to an expected payoff which is less than case 1, where $\mu_1 = 1$.

Case 7: $0 \leq \mu_1 \leq \mu^N$ and $0 \leq \mu_0 \leq \mu^N$. Then, the expected payoff of $\tau(1-\mu_1\gamma_h N)+(1-\tau)(1-\mu_0\gamma_h N)$ is maximised if $\mu_0 = \mu_1 = 0$, but this is not Bayes-plausible, for an interior p. If we search for a pair of Bayes-plausible (μ_0, μ_1), and let $\mu_0 = 0$, then given the law of total probability, $\tau = \frac{p}{\mu_1}$, and the expected payoff of $\tau(1-\mu_1\gamma_h N)+(1-\tau)$, is reduced to $1-p\gamma_h N$. But for any $p > 0$, this is less than the maximised expected payoff of case 1, which is $1 - p$. Moreover, as the expected payoff is linear in μ_0 and μ_1, any other combination of the two variables in this range leads to an expected payoff which is less than $1 - p$.

Therefore, case 1 has the unique maximum expected payoff of $1 - p$, by choosing the corner solution of $\mu_0 = 0$ and $\mu_1 = 1$. Hence, optimally $\pi(1 \mid \gamma_h) = 1$ and $\pi(0 \mid \gamma_l) = 1$.

8.3 Proof of lemma 3

Based on the expected payoff of non-signatories for any belief in equation (5.2), the problem of sender can be formalised as maximising

$$\begin{aligned}
&\tau \begin{cases} 0 & \text{if } \mu^f \leq \mu_1 \leq 1 \text{ or if } \mu^N < \mu_1 < \mu^{N-1} \\ 1-\mu_1\gamma_h(N-n^*(\mu_1)) & \text{if } \mu^{n^*} < \mu_1 \leq \frac{1}{(n^*(\mu_1)-1)\gamma_h} \\ 1-\mu_1\gamma_h N & \text{if } 0 \leq \mu_1 \leq \mu^N \end{cases} \\
&+(1-\tau) \begin{cases} 0 & \text{if } \mu^f \leq \mu_0 \leq 1 \text{ or if } \mu^N < \mu_0 < \mu^{N-1} \\ 1-\mu_0\gamma_h(N-n^*(\mu_0)) & \text{if } \mu^{n^*} < \mu_0 \leq \frac{1}{(n^*(\mu_0)-1)\gamma_h} \\ 1-\mu_0\gamma_h N & \text{if } 0 \leq \mu_0 \leq \mu^N \end{cases}
\end{aligned} \tag{8.4}$$

with respect to μ_1 and μ_0, subject to $\tau\mu_1 + (1-\tau)\mu_0 = p$. Also, as mentioned, in this equation by $n^*(\mu_s)$ we refer to any $2 \leq n^*(\mu_s) \leq N-1$. In addition, for the case of $n^*(\mu_s) = 2$, the range of beliefs are $\mu^2 < \mu_s < \mu^f$, and for the case of $n^*(\mu_s) = N-1$, the range of beliefs are $\mu^{N-1} \leq \mu_s \leq \mu^{N-2}$. The expected payoff in (8.4) can be rewritten as

$$
\begin{aligned}
\tau\nu(\mu_1)+(1-\tau)\nu(\mu_0) &= [(1-\tau)(1-\mu_0\gamma_h N)]\mathbf{1}_{\substack{\mu^f\le\mu_1\le 1\\ 0\le\mu_0\le\mu^N}} + [0]\mathbf{1}_{\substack{\mu^f\le\mu_1\le 1\\ \mu^N<\mu_0<\mu^{N-1}}} \\
&+ [(1-\tau)(1-\mu_0\gamma_h(N-n^*(\mu_0)))]\mathbf{1}_{\substack{\mu^f\le\mu_1\le 1\\ \mu^{n^*}<\mu_0\le\frac{1}{(n^*(\mu_0)-1)\gamma_h}}} + [0]\mathbf{1}_{\substack{\mu^f\le\mu_1\le 1\\ \mu^f\le\mu_0\le 1}} \\
&+ [(1-\tau)(1-\mu_0\gamma_h N)+\tau(1-\mu_1\gamma_h(N-n^*(\mu_1)))]\mathbf{1}_{\substack{\mu^{n^*}<\mu_1\le\frac{1}{(n^*(\mu_1)-1)\gamma_h}\\ 0\le\mu_0\le\mu^N}} \\
&+ [\tau(1-\mu_1\gamma_h(N-n^*(\mu_1)))]\mathbf{1}_{\substack{\mu^{n^*}<\mu_1\le\frac{1}{(n^*(\mu_1)-1)\gamma_h}\\ \mu^N<\mu_0<\mu^{N-1}}} \\
&+ [\tau(1-\mu_1\gamma_h(N-n^*(\mu_1)))+(1-\tau)(1-\mu_0\gamma_h(N-n^*(\mu_0)))]\mathbf{1}_{\substack{\mu^{n^*}<\mu_1\le\frac{1}{(n^*(\mu_1)-1)\gamma_h}\\ \mu^{n^*}<\mu_0\le\frac{1}{(n^*(\mu_0)-1)\gamma_h}}} \\
&+ [\tau(1-\mu_1\gamma_h(N-n^*(\mu_1))]\mathbf{1}_{\substack{\mu^{n^*}<\mu_1\le\frac{1}{(n^*(\mu_1)-1)\gamma_h}\\ \mu^f\le\mu_0\le 1}} + [(1-\tau)(1-\mu_0\gamma_h N)]\mathbf{1}_{\substack{\mu^N<\mu_1<\mu^{N-1}\\ 0\le\mu_0\le\mu^N}} \\
&+ [0]\mathbf{1}_{\substack{\mu^N<\mu_1<\mu^{N-1}\\ \mu^N<\mu_0<\mu^{N-1}}} + [(1-\tau)(1-\mu_0\gamma_h(N-n^*(\mu_0)))]\mathbf{1}_{\substack{\mu^N<\mu_1<\mu^{N-1}\\ \mu^{n^*}<\mu_0\le\frac{1}{\gamma_h(n^*(\mu_0)-1)}}} \\
&+ [0]\mathbf{1}_{\substack{\mu^N<\mu_1<\mu^{N-1}\\ \mu^f\le\mu_0\le 1}} + [\tau(1-\mu_1\gamma_h N)+(1-\tau)(1-\mu_0\gamma_h N)]\mathbf{1}_{\substack{0\le\mu_1\le\mu^N\\ 0\le\mu_0\le\mu^N}} \\
&+ [\tau(1-\mu_1\gamma_h N)]\mathbf{1}_{\substack{0\le\mu_1\le\mu^N\\ \mu^N<\mu_0<\mu^{N-1}}} \\
&+ [\tau(1-\mu_1\gamma_h N)+(1-\tau)(1-\mu_0\gamma_h(N-n^*(\mu_0)))]\mathbf{1}_{\substack{0\le\mu_1\le\mu^N\\ \mu^{n^*}<\mu_0\le\frac{1}{(n^*(\mu_0)-1)\gamma_h}}} \\
&+ [\tau(1-\mu_1\gamma_h N)]\mathbf{1}_{\substack{0\le\mu_1\le\mu^N\\ \mu^f\le\mu_0\le 1}}
\end{aligned}
\tag{8.5}
$$

Again we label each term of the above expected payoff by ascending numbers, e.g. case 1 refers to the first term where $\mu^f \le \mu_1 \le 1$ and $0 \le \mu_0 \le \mu^N$. So, there are 16 cases, where cases 8, 11, 12, 14, 15, and 16 cannot be equilibrium because in these cases $\mu_0 > \mu_1$.

The cases with potential positive payoffs are as follows:

Case 1: $\mu^f \le \mu_1 \le 1$ and $0 \le \mu_0 \le \mu^N$. This can be a candidate, as by setting $\mu_0 = 0$, the Bayes-plausibility implies $\tau = \frac{p}{\mu_1}$. Hence, the expected payoff of $1-\tau$ is maximised by setting $\mu_1 = 1$. Thus the full-revelation policy leads to the expected payoff of $1-p$.

Case 3: $\mu^f \le \mu_1 \le 1$ and $\mu^{n^*} < \mu_0 \le \frac{1}{(n^*(\mu_0)-1)\gamma_h}$. Ideally, choosing μ_0 in the range that $\mu^{N-1} < \mu_0 \le \mu^{N-2}$, leads to $n^*(\mu_0) = N-1$, and expected payoff of $(1-\tau)(1-\mu_0\gamma_h)$, which is increasing in μ_1, as

$\mu_0 \leq p$. Therefore, setting $\mu_1 = 1$ leads to expected payoff of $\frac{1-p}{1-\mu_0}(1-\mu_0\gamma_h)$, which is strictly less than the maximised expected payoff of case 1, $1-p$, for any μ_0 in this range.

Case 5: $\mu^{n^*} < \mu_1 \leq \frac{1}{(n^*(\mu_1)-1)\gamma_h}$ and $0 \leq \mu_0 \leq \mu^N$. The expected payoff is maximised if $\mu_0 = 0$, and Bayes-plausibility implies $\tau = \frac{p}{\mu_1}$. So, the resulted expected payoff of $\tau(1-\mu_1\gamma_h(N-n^*(\mu_1))) + (1-\tau)$ can be simplified to $1-p\gamma_h(N-n^*(\mu_1))$. Again, to maximise this, μ_1 can be selected such that $\mu^{N-1} < \mu_1 \leq \mu^{N-2}$. Thus, $n^*(\mu_1) = N-1$, and the expected payoff is reduced to $1-p\gamma_h$, but this is always less than $V(p) = (1-p)$ in case 1, for any $p > 0$.

Case 6: $\mu^{n^*} < \mu_1 \leq \frac{1}{(n^*(\mu_1)-1)\gamma_h}$ and $\mu^N \leq \mu_0 \leq \mu^{N-1}$. The corresponding expected payoff is maximised by choosing $\mu^{N-1} < \mu_1 \leq \mu^{N-2}$ to the expected payoff of $\tau(1-\mu_1\gamma_h)$. The law of total probability for μ_0 and μ_1, restricts the resulted expected payoff to $\frac{p-\mu_0}{\mu_1-\mu_0}(1-\mu_1\gamma_h)$, which is decreasing in μ_1. The minimum possible value of μ_1 is p, but it leads to expected payoff of $1-\mu_1\gamma_h$, which is less than $1-p$ for any $p > 0$.

Case 7: $\mu^{n^*} < \mu_1 \leq \frac{1}{(n^*(\mu_1)-1)\gamma_h}$ and $\mu^{n^*} < \mu_0 \leq \frac{1}{(n^*(\mu_0)-1)\gamma_h}$. The associated expected payoff is maximised if $\mu^{N-1} < \mu_1 \leq \mu^{N-2}$ and $\mu^{N-1} < \mu_0 \leq \mu^{N-2}$. The resulted expected payoff of $\tau(1-\mu_1\gamma_h) + (1-\tau)(1-\mu_0\gamma_h)$, given a pair of Bayes-plausible (μ_0, μ_1), and replacing $\mu_1 = \frac{p}{\tau} - \frac{1-\tau}{\tau}\mu_0$, can be simplified to $1-p\gamma_h$, which for any $p > 0$, is less than the payoff of case 1.

Case 9: $\mu^N < \mu_1 \leq \mu^{N-1}$ and $0 \leq \mu_0 \leq \mu^N$. Similar to case 1, it is possible to set $\mu_0 = 0$ and obtain expected payoff of $1-\tau$. This is increasing in μ_1, but in this range of beliefs, it cannot be as maximised as case 1, where it was possible to select $\mu_1 = 1$.

Case 13: $0 \leq \mu_1 < \mu^N$ and $0 \leq \mu_0 < \mu^N$. This case is ruled out as it is identical to case 7 of the signatories problem in the proof of lemma 2.

Therefore, case 1 provides the unique maximum expected payoff of $V(p) = 1-p$, and the optimal policy is full revelation of the state.

8.4 Proof of proposition 3

Given equations (8.2) and (8.4), and adjusting the tie-breaking rule (for $N > 2$), the expected payoff of sender specified in (5.3) is

$$
\begin{aligned}
&\tau\nu(\mu_1)+(1-\tau)\nu(\mu_0)=[\alpha N(1-\tau)(1-\mu_0\gamma_h N)]\mathbf{1}_{\substack{\mu^f\le\mu_1\le 1\\ 0\le\mu_0\le\mu^N}}+[0]\mathbf{1}_{\substack{\mu^f\le\mu_1\le 1\\ \mu^N<\mu_0<\mu^{N-1}}}\\
&+[(1-\tau)(N-n^*(\mu_0))(1-\mu_0\gamma_h(N-n^*(\mu_0)))\\
&-(1-\tau)\alpha n^*(\mu_0)\mu_0\gamma_h(N-n^*(\mu_0))]\mathbf{1}_{\substack{\mu^f\le\mu_1\le 1\\ \mu^{n^*}<\mu_0\le\frac{1}{(n^*(\mu_0)-1)\gamma_h}}}\\
&+[0]\mathbf{1}_{\substack{\mu^f\le\mu_1\le 1\\ \mu^f\le\mu_0\le 1}}+[\alpha N(1-\tau)(1-\mu_0\gamma_h N)\\
&-\alpha\tau n^*(\mu_1)\mu_1\gamma_h(N-n^*(\mu_1))+\tau(N-n^*(\mu_1))(1-\mu_1\gamma_h(N-n^*(\mu_1)))]\mathbf{1}_{\substack{\mu^{n^*}<\mu_1\le\frac{1}{(n^*(\mu_1)-1)\gamma_h}\\ 0\le\mu_0\le\mu^N}}\\
&+[\tau(N-n^*(\mu_1))(1-\mu_1\gamma_h(N-n^*(\mu_1))-\alpha\tau n^*(\mu_1)\mu_1\gamma_h(N-n^*(\mu_1))]\mathbf{1}_{\substack{\mu^{n^*}<\mu_1\le\frac{1}{(n^*(\mu_1)-1)\gamma_h}\\ \mu^N<\mu_0<\mu^{N-1}}}\\
&+[(1-\tau)(N-n^*(\mu_0))(1-\mu_0\gamma_h(N-n^*(\mu_0)))+\tau(N-n^*(\mu_1))(1-\mu_1\gamma_h(N-n^*(\mu_1)))\\
&-\alpha\tau n^*(\mu_1)\mu_1\gamma_h(N-n^*(\mu_1))-\alpha(1-\tau)n^*(\mu_0)\mu_0\gamma_h(N-n^*(\mu_0))]\mathbf{1}_{\substack{\mu^{n^*}<\mu_1\le\frac{1}{(n^*(\mu_1)-1)\gamma_h}\\ \mu^{n^*}<\mu_0\le\frac{1}{(n^*(\mu_0)-1)\gamma_h}}}\\
&+[\tau(N-n^*(\mu_1))(1-\mu_1\gamma_h(N-n^*(\mu_1)))\\
&-\tau\alpha n^*(\mu_1)\mu_1\gamma_h(N-n^*(\mu_1))]\mathbf{1}_{\substack{\mu^{n^*}<\mu_1\le\frac{1}{(n^*(\mu_1)-1)\gamma_h}\\ \mu^f\le\mu_0\le 1}}\\
&+[\alpha N(1-\tau)(1-\mu_0\gamma_h N)]\mathbf{1}_{\substack{\mu^N<\mu_1<\mu^{N-1}\\ 0\le\mu_0\le\mu^N}}+[0]\mathbf{1}_{\substack{\mu^N<\mu_1<\mu^{N-1}\\ \mu^N<\mu_0<\mu^{N-1}}}\\
&+[(1-\tau)(N-n^*(\mu_0))(1-\mu_0\gamma_h(N-n^*(\mu_0)))\\
&-(1-\tau)\alpha n^*(\mu_0)\mu_0\gamma_h(N-n^*(\mu_0))]\mathbf{1}_{\substack{\mu^N<\mu_1<\mu^{N-1}\\ \mu^{n^*}<\mu_0\le\frac{1}{\gamma_h(n^*(\mu_0)-1)}}}+[0]\mathbf{1}_{\substack{\mu^N<\mu_1<\mu^{N-1}\\ \mu^f\le\mu_0\le 1}}\\
&+[\alpha N(\tau(1-\mu_1\gamma_h N)+(1-\tau)(1-\mu_0\gamma_h N))]\mathbf{1}_{\substack{0\le\mu_1\le\mu^N\\ 0\le\mu_0\le\mu^N}}+[\alpha N\tau(1-\mu_1\gamma_h N)]\mathbf{1}_{\substack{0\le\mu_1\le\mu^N\\ \mu^N<\mu_0<\mu^{N-1}}}\\
&+[\alpha N\tau(1-\mu_1\gamma_h N)-\alpha(1-\tau)n^*(\mu_0)\mu_0\gamma_h(N-n^*(\mu_0))\\
&+(1-\tau)(N-n^*(\mu_0))(1-\mu_0\gamma_h(N-n^*(\mu_0)))]\mathbf{1}_{\substack{0\le\mu_1\le\mu^N\\ \mu^{n^*}<\mu_0\le\frac{1}{(n^*(\mu_0)-1)\gamma_h}}}\\
&+[\alpha N\tau(1-\mu_1\gamma_h N)]\mathbf{1}_{\substack{0\le\mu_1\le\mu^N\\ \mu^f\le\mu_0\le 1}}
\end{aligned}
\tag{8.6}
$$

where in this equation, $n^*(\mu_s)$ refers to $2\le n^*(\mu_s)\le N-1$.

Again by labelling each term of the above expected payoff by ascending numbers, (e.g. case 1 refers

to the first term where $\mu^f \leq \mu_1 \leq 1$ and $0 \leq \mu_0 \leq \mu^N$) and ruling out cases where $\mu_0 > \mu_1$ from the analysis, i.e., cases 8, 11, 12, 14, 15, and 16, we compare potential positive payoffs:

Case 1: $\mu^f \leq \mu_1 \leq 1$ and $0 \leq \mu_0 < \mu^N$. Selecting $\mu_0 = 0$, and $\mu_1 = 1$, leads to expected payoff of $\alpha N(1-p)$.

Case 3: $\mu^f \leq \mu_1 \leq 1$ and $\mu^{n^*} \leq \mu_0 < \frac{1}{\gamma_h(n^*(\mu_0)-1)}$. Given that the maximum possible expected payoff in this range is obtained by selecting $n^*(\mu_0) = N-1$. So, the corresponding expected payoff can be written as $(1-\tau)(1-\mu_0\gamma_h) - (1-\tau)(N-1)\alpha\mu_0\gamma_h$. For any $\alpha \geq 1$ and $N > 2$, the last term is non-positive. For any positive μ_0, this is less that the expected payoff of case 1.

Case 5: $\mu^{n^*} \leq \mu_1 < \frac{1}{\gamma_h(n^*(\mu_1)-1)}$ and $0 \leq \mu_0 < \mu^N$. Undoubtedly, it is optimal to set $\mu_0 = 0$, thus $\tau = \frac{p}{\mu_1}$ implies that the corresponding expected payoff is $N\frac{p}{\mu_1}(1-\mu_1\gamma_h) + \alpha N(1-\frac{p}{\mu_1}) + (1-\alpha)(N-1)p\gamma_h$. This expected payoff can be further simplified to $N\alpha - p\gamma_h(\alpha(N-1)+1) + N\frac{p}{\mu_1}(1-\alpha)$, where the last two terms are non-positive for any $\alpha \geq 1$. So, this is always less than the expected payoff of case 1, $N\alpha - N\alpha p$.

Case 6: $\mu^{n^*} \leq \mu_1 < \frac{1}{\gamma_h(n^*(\mu_1)-1)}$ and $\mu^N < \mu_0 < \mu^{N-1}$. Selection of μ_1 such that $n^*(\mu_1) = N-1$ leads to expected payoff of $\tau(1-\mu_1\gamma_h[1+\alpha(N-1)])$. It is decreasing in both μ_0 and μ_1, but relative to case 1, the corner solutions where either of these could be zero is not available in this range of beliefs. Also since $\alpha \geq 1$ and $N-1 > 1$, this expected payoff is less than case 1.

Case 7: $\mu^{n^*} \leq \mu_1 < \frac{1}{\gamma_h(n^*(\mu_1)-1)}$ and $\mu^{n^*} \leq \mu_0 < \frac{1}{\gamma_h(n^*(\mu_0)-1)}$. The last two terms of the expected payoff are negative. Given the fact that in this range of beliefs, the expected payoffs of signatories and non-signatories are maximised if $n^*(\mu_s) = N-1$, the first two terms are identical to case 7 in the proof of lemma 3, where a Bayes-plausible pair of beliefs leads to expected payoff of $1-p\gamma_h$, which is inferior to the case 1, for any $\alpha \geq 1$.

Case 9: $\mu^N < \mu_1 < \mu^{N-1}$ and $0 \leq \mu_0 < \mu^N$. The corresponding expected payoff is similar to case 1 if $\mu_0 = 0$. However, here it is not possible to set $\mu_1 = 1$ and achieve the maximised value of expected payoff of case 1.

Case 13: $0 \leq \mu_1 \leq \mu^N$ and $0 \leq \mu_0 \leq \mu^N$. This case is also ruled out, as it is similar to case 7 in proof of lemma 2.

Hence, the expected payoff of case 1, given $\mu_0 = 0$ and $\mu_1 = 1$, provides a unique maximum expected payoff. Furthermore, selection of $\mu_0 = 0$ leads to selection of causing harm by all agents, while selection of $\mu_1 = 1$ implies coordination of all on preventing. This in turn coincides with the socially optimal action

outcome specified in 1.

8.5 Proof of proposition 4

Given the expected payoff of (6.1) for any belief, the problem of sender can be formalised as maximising

$$\tau \begin{cases} 0 & \text{if } \mu^f \leq \mu_1 \leq 1 \\ n^*(\mu_1) - N & \text{if } \mu^{n^*} < \mu_1 \leq \frac{1}{(n^*(\mu_1)-1)\gamma_h} \\ -N & \text{if } 0 \leq \mu_1 < \mu^N \end{cases} \tag{8.7}$$

$$+(1-\tau) \begin{cases} 0 & \text{if } \mu^f \leq \mu_0 \leq 1 \\ n^*(\mu_0) - N & \text{if } \mu^{n^*} < \mu_0 \leq \frac{1}{(n^*(\mu_0)-1)\gamma_h} \\ -N & \text{if } 0 \leq \mu_0 < \mu^N \end{cases}$$

with respect to μ_1 and μ_0, such that $\tau\mu_1 + (1-\tau)\mu_0 = p$. In addition, in this equation by $n^*(\mu_s)$ we refer to any $2 \leq n^*(\mu_s) \leq N$. Furthermore, the tie-breaking rule is such that for the case of $n^*(\mu_s) = 2$, the range of beliefs are $\mu^2 < \mu_s < \mu^f$, and for the case of $n^*(\mu_s) = N$, the range of beliefs are $\mu^N \leq \mu_s \leq \mu^{N-1}$. The expected payoff in (8.7) can be rewritten as

$$
\begin{aligned}
\tau\nu(\mu_1) + (1-\tau)\nu(\mu_0) = &-[(1-\tau)N]\mathbf{1}_{\substack{\mu^f \le \mu_1 \le 1 \\ 0 \le \mu_0 < \mu^N}} \\
&- [(1-\tau)(N - n^*(\mu_0))]\mathbf{1}_{\substack{\mu^f \le \mu_1 \le 1 \\ \mu^{n^*} < \mu_0 \le \frac{1}{(n^*(\mu_0)-1)\gamma_h}}} \\
&+ [0]\mathbf{1}_{\substack{\mu^f \le \mu_1 \le 1 \\ \mu^f \le \mu_0 \le 1}} - [\tau(N - n^*(\mu_1)) + (1-\tau)N)]\mathbf{1}_{\substack{\mu^{n^*} < \mu_1 \le \frac{1}{(n^*(\mu_1)-1)\gamma_h} \\ 0 \le \mu_0 < \mu^N}} \\
&- [\tau(N - n^*(\mu_1)) + (1-\tau)(N - n^*(\mu_0))]\mathbf{1}_{\substack{\mu^{n^*} < \mu_1 \le \frac{1}{(n^*(\mu_1)-1)\gamma_h} \\ \mu^{n^*} < \mu_0 \le \frac{1}{(n^*(\mu_0)-1)\gamma_h}}} \\
&- [\tau(N - n^*(\mu_1))]\mathbf{1}_{\substack{\mu^{n^*} < \mu_1 \le \frac{1}{(n^*(\mu_1)-1)\gamma_h} \\ \mu^f \le \mu_0 \le 1}} - [N]\mathbf{1}_{\substack{0 \le \mu_1 < \mu^N \\ 0 \le \mu_0 < \mu^N}} \\
&- [\tau N + (1-\tau)(N - n^*(\mu_0))]\mathbf{1}_{\substack{0 \le \mu_1 < \mu^N \\ \mu^{n^*} < \mu_0 \le \frac{1}{(n^*(\mu_0)-1)\gamma_h}}} - [\tau N]\mathbf{1}_{\substack{0 \le \mu_1 < \mu^N \\ \mu^f \le \mu_0 \le 1}}
\end{aligned}
\tag{8.8}
$$

Let us first label each term of the above expected payoff by ascending numbers, e.g. case 1 refers to the first term where $\mu^f \le \mu_1 \le 1$ and $0 \le \mu_0 < \mu^N$. Furthermore, note that cases 6, 8, and 9 are ruled out as in these case $\mu_0 > \mu_1$.

In order to prove the proposition, let us fix p, and consider three possibilities for the prior belief.

First, $\mu^f \le p < 1$. A Bayes-plausible lottery over posteriors should satisfy the order of $\mu_0 \le p \le \mu_1$. Thus, the corresponding expected payoffs of cases 1, 2, and 3 must be compared. As in this range of prior beliefs, $\mu^f \le p \le 1$, the agents in the absence of any persuasion, take the sender's-preferred action, the optimal policy can be sending no signal at all. This is not a unique optimal information policy, and it is equivalent to a degenerate randomisation of posteriors, which is equal to the prior with probability one, i.e. $\mu_1 = p$ and $\tau = 1$. This can be obtained by a degenerate randomisation in either of cases 1, 2, or 3, leading to the expected payoff of zero. Another optimal policy is available in case 3, by selecting a Bayes-plausible randomisation over $\mu_0 = \mu^f$ and $\mu_1 = 1$.

Second, $\mu^{n^*} \le p < \frac{1}{(n^*(\mu_s)-1)\gamma_h}$. Searching for a maximum payoff should include expected payoffs of cases 2, 4, and 5. The expected payoff of case 2 is maximised if $\mu_0 = \mu^N$, which implies $n^*(\mu_0) = N$, and expected payoff of zero. Therefore, for any prior belief in $\mu^N \le p \le 1$, a Bayes-plausible randomisation of $\mu_0 = \mu^N$ and $\mu_1 = 1$ is optimal. If $\mu^N \le p \le \mu^f$, another optimal policy is randomisation of $\mu_0 = \mu^N$ and $\mu_1 = \mu^f$. Similarly if $\mu^{N-1} \le p \le \mu^f$, then a randomisation of $\mu_0 = \mu^{N-1}$ and $\mu_1 = \mu^f$ (or $\mu_1 = 1$),

which satisfies the law of total probability is also optimal. In short, for any $\mu^N \leq p < 1$, a Bayes-plausible randomisation over $\mu_0 \in [\mu^N, \mu^{N-1}]$ and $\mu_1 \in [\mu^f, 1]$ is an optimal information policy. In case 5, it is possible to obtain $V(p) = 0$, if and only if $\mu^N \leq p \leq \mu^{N-1}$. In such a case, that the agents choose the sender's preferred action in the absence of any persuasion, the optimal policy is not unique, and there can be no persuasion, or a degenerate lottery of $\mu_1 = p$ and $\tau = 1$, or the policy of selection of a Bayes-plausible randomisation over $\mu_0 = \mu^N$ and $\mu_1 = \mu^{N-1}$. All of these policies lead to expected payoff of zero for the sender. Finally, in case 4, if and only if $\mu^N \leq p \leq \mu^{N-1}$, the optimal policy of $\mu_1 = p$ and $\tau = 1$ leads to $V(p) = 0$.

Third, $0 < p < \mu^N$. This implies comparison of expected payoffs of cases 1, 4, and 7. Let us start the analysis with comparison of expected payoffs of cases 1 and 4. In both cases the expected payoffs are increasing in τ, and therefore decreasing in both μ_0 and μ_1. Hence, for case 1, choosing $\mu_0 = 0$ and $\mu_1 = \mu^f$ obtains maximum expected payoff of $\frac{p}{\mu^f} N - N$. While for case 4, choosing the minimum possible posteriors of $\mu_0 = 0$ and $\mu_1 = \mu^N$, leads to maximum expected payoff of $\frac{p}{\mu^N} N - N$. The expected payoff of 4 is more than 1 if $N \geq 2$, and the two cases lead to the same expected payoff. In addition, for any non-degenerate randomisation over posteriors, case 7 has the least expected payoff. Therefore, the unique optimal information policy is a Bayes-plausible randomisation of $\mu_0 = 0$ and $\mu_1 = \mu^N$.

References

Aldy, J. E., and R. N. Stavins (2009): *Post-Kyoto international climate policy: implementing architectures for agreement*. Cambridge University Press.

Allen, B., and N. C. Yannelis (2001): "Differential Information Economies: Introduction," *Economic Theory*, 18(2), 263–273.

Aumann, R. J., M. Maschler, and R. E. Stearns (1995): *Repeated games with incomplete information*. MIT press.

Ayres, I., S. Raseman, and A. Shih (2013): "Evidence from two large field experiments that peer comparison feedback can reduce residential energy usage," *Journal of Law, Economics, and Organization*, 29(5), 992–1022.

Barrett, S. (1994): "Self-enforcing international environmental agreements," *Oxford Economic Papers*, pp. 878–894.

Benchekroun, H., and N. V. Long (2012): "Collaborative Environmental Management: A Review Of The Literature," *International Game Theory Review*, 14(04).

Benchekroun, H., and A. Ray Chaudhuri (2011): "Environmental policy and stable collusion: The case of a dynamic polluting oligopoly," *Journal of Economic Dynamics and Control*, 35(4), 479–490.

Bloch, F. (1996): "Sequential formation of coalitions with fixed payoff division and externalities," *Games and Economic Behavior*, 14, 90–123.

Breton, M., L. Sbragia, and G. Zaccour (2010): "A dynamic model for international environmental agreements," *Environmental and Resource Economics*, 45(1), 25–48.

Carraro, C., and D. Siniscalco (1993): "Strategies for the international protection of the environment," *Journal of public Economics*, 52(3), 309–328.

Chwe, M. S.-Y. (1994): "Farsighted coalitional stability," *Journal of Economic Theory*, 63(2), 299–325.

Costa, D. L., and M. E. Kahn (2013): "Energy conservation nudges and environmentalist ideology: evidence from a randomized residential electricity field experiment," *Journal of the European Economic Association*, 11(3), 680–702.

Diamantoudi, E., and E. S. Sartzetakis (2006): "Stable international environmental agreements: An analytical approach," *Journal of public economic theory*, 8(2), 247–263.

Dutta, B., and R. Vohra (2005): "Incomplete information, credibility and the core," *Mathematical Social Sciences*, 50(2), 148–165.

Ecchia, G., and M. Mariotti (1998): "Coalition formation in international environmental agreements and the role of institutions," *European Economic Review*, 42(3), 573–582.

Ely, J. C. (2017): "Beeps," *The American Economic Review*, 107(1), 31–53.

Ferraro, P. J., J. J. Miranda, and M. K. Price (2011): "The persistence of treatment effects with norm-based policy instruments: evidence from a randomized environmental policy experiment," *The American Economic Review*, pp. 318–322.

Finus, M. (2002): "Game theory and international environmental cooperation: any practical application?," *Controlling global warming: Perspectives from economics, game theory and public choice*, pp. 9–104.

——— (2008): "Game theoretic research on the design of international environmental agreements: Insights, critical remarks, and future challenges," *International Review of Environmental and Resource Economics*, 2(1), 29–67.

Finus, M., and P. Pintassilgo (2013): "The role of uncertainty and learning for the success of international climate agreements," *Journal of Public Economics*, 103, 29–43.

Galle, B. (2013): "Tax, Command or Nudge: Evaluating the New Regulation," *Tex. L. Rev.*, 92, 837.

Gehlbach, S., and K. Sonin (2014): "Government control of the media," *Journal of Public Economics*, 118, 163–171.

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

——— (2012): "Disclosure of endogenous information," *University of Chicago mimeo*.

——— (2014): "Costly persuasion," *The American Economic Review*, 104(5), 457–462.

Kolotilin, A., M. Li, T. Mylovanov, and A. Zapechelnyuk (2015): "Persuasion of a Privately Informed Receiver," Discussion paper, Working paper.

Kolstad, C. D. (2007): "Systematic uncertainty in self-enforcing international environmental agreements," *Journal of Environmental Economics and Management*, 53(1), 68–79.

Kolstad, C. D., and M. Toman (2001): "The economics of climate policy," *Handbook of Environmental Economics*, 2.

Kolstad, C. D., and A. Ulph (2011): "Uncertainty, learning and heterogeneity in international environmental agreements," *Environmental and Resource Economics*, 50(3), 389–403.

Kosfeld, M., A. Okada, and A. Riedl (2009): "Institution formation in public goods games," *The American Economic Review*, 99(4), 1335–1355.

Na, S.-l., and H. S. Shin (1998): "International environmental agreements under uncertainty," *Oxford Economic Papers*, 50(2), 173–185.

Ray, D. (2007): *A game-theoretic perspective on coalition formation*. Oxford University Press.

Ray, D., and R. Vohra (2015): "Coalition formation," *Handbook of Game Theory*, 4, 239–326.

Rubio, S. J., and A. Ulph (2007): "An infinite-horizon model of dynamic membership of international environmental agreements," *Journal of Environmental Economics and Management*, 54(3), 296–310.

Shapiro, J. M. (2016): "Special interests and the media: Theory and an application to climate change," *Journal of Public Economics*, 144, 91–108.

Taneva, I. (2016): "Information Design," Discussion paper, Discussion paper, University of Edinburgh.

Toman, M. (1998): "Research frontiers in the economics of climate change," *Environmental and Resource Economics*, 11(3-4), 603–621.

Ulph, A. (2004): "Stable international environmental agreements with a stock pollutant, uncertainty and learning," *Journal of Risk and Uncertainty*, 29(1), 53–73.

Wagner, U. J. (2001): "The design of stable international environmental agreements: Economic theory and political economy," *Journal of economic surveys*, 15(3), 377–411.

NOTE DI LAVORO DELLA FONDAZIONE ENI ENRICO MATTEI

Fondazione Eni Enrico Mattei Working Paper Series

Our Note di Lavoro are available on the Internet at the following addresses:

http://www.econis.eu/LNG=EN/FAM?PPN=505954494
http://ageconsearch.umn.edu/handle/35978
http://www.bepress.com/feem/
http://labs.jstor.org/sustainability/

NOTE DI LAVORO PUBLISHED IN 2017

SAS	1.2017	
ET	2.2017	
SAS	3.2017	
ESP	4.2017	
ET	5.2017	
ET	6.2017	
MITP	7.2017	
MITP	8.2017	Samuel Carrara, Thomas Longden: Freight Futures: The Potential Impact of Road Freight on Climate Policy
ET	9.2017	Claudio Morana, Giacomo Sbrana: Temperature Anomalies, Radiative Forcing and ENSO
ESP	10.2017	Valeria Di Cosmo, Laura Malaguzzi Valeri: Wind, Storage, Interconnection and the Cost of Electricity Generation
EIA	11.2017	Elisa Delpiazzo, Ramiro Parrado, Gabriele Standardi: Extending the Public Sector in the ICES Model with an Explicit Government Institution
MITP	12.2017	Bai Chen Xie Jie Gao, Shuang Zhang, ZhongXiang Zhang: What Factors Affect the Competiveness of Power Generation Sector in China? An Analysis Based on Game Cross Efficiency
MITP	13.2017	Stergios Athanasoglou, Valentina Bosetti, Laurent Drouet: A Simple Framework for Climate Change Policy under Model Uncertainty
MITP	14.2017	Loïc Berger and Johannes Emmerling: Welfare as Simple(x) Equity Equivalents
ET	15.2017	Christoph M. Rheinberger, Felix Schläpfer, Michael Lobsiger: A Novel Approach to Estimating the Demand Value of Road Safety
MITP	16.2017	Giacomo Marangoni, Gauthier De Maere, Valentina Bosetti: Optimal Clean Energy R&D Investments Under Uncertainty
SAS	17.2017	Daniele Crotti, Elena Maggi: Urban Distribution Centres and Competition among Logistics Providers: a Hotelling Approach
ESP	18.2017	Quentin Perrier: The French Nuclear Bet
EIA	19.2017	Gabriele Standardi, Yiyong Cai, Sonia Yeh: Sensitivity of Modeling Results to Technological and Regional Details: The Case of Italy's Carbon Mitigation Policy
EIA	20.2017	Gregor Schwerhoff, Johanna Wehkamp: Export Tariffs Combined with Public Investments as a Forest Conservation Policy Instrument
MITP	21.2017	Wang Lu, Hao Yu, Wei Yi Ming: How Do Regional Interactions in Space Affect China's Mitigation Targets and Economic Development?
ET	22.2017	Andrea Bastianin, Paolo Castelnovo, Massimo Florio: The Empirics of Regulatory Reforms Proxied by Categorical Variables: Recent Findings and Methodological Issues
EIA	23.2017	Martina Bozzola, Emanuele Massetti, Robert Mendelsohn, Fabian Capitanio: A Ricardian Analysis of the Impact of Climate Change on Italian Agriculture
MITP	24.2017	Tunç Durmaz, Aude Pommeret, Ian Ridley: Willingness to Pay for Solar Panels and Smart Grids
SAS	25.2017	Federica Cappelli: An Analysis of Water Security under Climate Change
ET	26.2017	Thomas Demuynck, P. Jean-Jacques Herings, Riccardo D. Saulle, Christian Seel: The Myopic Stable Set for Social Environments
ET	27.2017	Joosung Lee: Mechanisms with Referrals: VCG Mechanisms and Multilevel Mechanisms
ET	[redacted]	Sareh Vosooghi: Information Design In Coalition Formation Games