Job Market Paper

Pluralistic Ignorance and Social Change*

A Model of Sequential Decisions with Second Order Conformity

Mauricio Fernández Duque†

September 13, 2017

Abstract

I develop a theory of group interaction in which individuals who act sequentially are concerned about signaling what they believe is the majority group preference. The framework allows me to study three features of collective action. First, equilibrium dynamics may result in a perverse situation where most individuals reluctantly act in a way they mistakenly believe is cooperative, a situation known as 'pluralistic ignorance'. Second, behavior may be affected by leaders, laws or surveys that influence what is thought to be the majority preference, possibly creating pluralistic ignorance. Third, abrupt social change may come about through the action of an obscure, politically inactive and uninformed individual whose brash actions reveal what everyone wishes they were doing. The model formalizes insights from scholarship that emphasizes how social meaning is constructed, and then applies these insights to political phenomena such as the Arab Spring, climate change beliefs, and the impact of get out the vote campaigns.

[*] I would like to thank Samuel Asher, Nava Ashraf, Robert Bates, Iris Bohnet, Jeffry Frieden, Benjamin Golub, Michael Hiscox, Scott Kominers, Kenneth Shepsle and seminar participants at Harvard University.

[†] Harvard University. E-mail: duque@fas.harvard.edu.

1 Introduction

When group members conform to what they think others want, they may end up doing what nobody wants. In a classic paper, O'Gorman (1975) shows that the majority of whites in the U.S. in 1968 did not favor segregation, about half believed that the majority of whites did favor segregation, and those who overestimated support for segregation were more willing to support housing policies that allowed others to segregate. O'Gorman was studying *pluralistic ignorance*, a social situation where 'a majority of group members privately reject a norm, but incorrectly assume that most others accept it, and therefore go along with it' (Katz and Allport, 1931). Pluralistic ignorance can be broken down into three facts. First, group members are acting *reluctantly* – many whites' public support for segregationist housing did not reflect their private views. Second, group members underestimate others' reluctance – whites believed others' support for segregationist housing did reflect their private views. Third, underestimating reluctance makes individuals more willing to act reluctantly – whites were more willing to support segregationist housing when they overestimated its support. Pluralistic ignorance can result in a perverse failure of collective action in which individuals reluctantly do what they mistakenly think is socially optimal. Although pluralistic ignorance is a well studied phenomenon found in topics as varied as climate change (Mildenberger and Tingley, 2016), political correctness (Van Boven, 2000), and tax evation (Wenzel, 2005), existing formal models do not capture its main features.

In this paper I capture pluralistic ignorance with a model which introduces *second order conformity*, in which group members are motivated to act according to what they believe the group believes is the majority preference. I use this framework to analyze social change by asking when leaders and policies affect what is considered the majority preference, and how abrupt change may come through an everyman who reveals what everyone wishes they were doing.

To develop intuition for the model, consider a stylized conversation between a group of acquaintances in which they take turns stating whether they support or oppose segregationist housing. A concern for conforming motivates them to express a view that matches their acquaintances' views (similar to Bernheim, 1994), but individuals are not sure what their acquaintances' views are.[1] An individual must use *context* to form beliefs about others' views – for instance, the probability an acquaintance is anti-segregationist is more likely in a college campus than in a white supremacist rally.

[1] The model is of *second order* conformity because individuals are uncertain over what the majority prefers – they are motivated to signal what they believe is the majority preference. Past models of conformity have assumed that individuals know which preference others consider valuable (e.g. Bernheim, 1994, Ellingsen et al., 2008, Benabou and Tirole, 2012).

In order to capture context, suppose group members are either drawn from a population of mostly anti-segregationists, or from a population of mostly pro-segregationists. Individuals are uncertain over which population their group was drawn from, but they do know the probability with which they were drawn from the pro-segregationist population – this commonly known probability is the context. In a supremacist rally, for instance, individuals know group members were drawn from the pro-segregationist population with a high probability. The first speaker in the stylized conversation will express an opinion which depends on context. When the context strongly indicates that acquaintances are pro-segregationist, the first speaker will conform by favoring segregationist housing independent of her views, leaving the second speaker in the same position. However, when context is completely uninformative about others' views, the first speaker will prefer to express her private views. The first speaker's revelation of her private views impacts the opinions expressed by others: if the first speaker truthfully reveals she is pro-segregationist, the second speaker will be more likely to conform to a pro-segregationist view, as she'll believe the group is more pro-segregationist than what the first speaker believed.[2] The informativeness of context, therefore, leads to different dynamics of opinion expression, and therefore to different conditions under which pluralistic ignorance arises. Note further that the argument so far has not specified the size of the group, so we can use this model to think of opinion expression with groups of arbitrary size. We can think of whites in the U.S. as a large group, and use the dynamics of their conversation as a way to capture public opinion formation. Alternatively, it could be a conversation between a small group of whites. I show the probability of pluralistic ignorance is lowest in small groups when context is uninformative, and lowest in large groups when context is informative.

To show how I use the model to analyze social change, I will now briefly describe it in more general terms. Members of a group take turns making a decision, which could be the opinion they express, or whether to protest, vote, or contribute to a public good. Group members judge whether the private views reflected in an individual's behavior reflect those of the group majority. However, since views are private, the group majority view is uncertain. Individuals do not know which population group members were drawn from, but there is common knowledge over the context. Decision makers take into account how others' judgments will depend on their private information and on what is known about preferences

[2] The uninformative context dynamic is analogous to *herding* (Banerjee, 1992) or *information cascades* (Bikhchandani et al., 1992). Individuals herd on what they think the majority group preference is. Unlike in a standard herding model, individuals are concerned with what the private view they reveal says about them, although they are perfectly informed of their private view. This leads to different behavior when they think they will be judged and when they do not. I discuss experimental evidence for this behavior in section 4.2.2.

given the context and past behaviors.

By reinterpreting the model as a sequence of group decisions more generally, second order conformity provides a framework for analyzing social change. The approach captures dynamics which are hard to explain with standard rational choice accounts, and which have found compelling explanations from scholars that emphasize how social meaning is constructed (Bates et al., 1998, Finnemore, 1996, Tarrow, 2011, McAdam et al., 1996). Leaders, laws and surveys often affect social change by redefining which behaviors signal the majority preference, or are 'appropriate' (March and Olsen, 2004). I now provide some empirical examples that illustrate this phenomenon, and which suggest tensions with standard rational choice explanations. I will capture these examples in my model by considering the impact of adding different leaders and policies to the beginning of the game.

Gandhi was able to solve a large collective action problem by convincing his followers that a true Indian non-violently responds to being beaten by the British. It was not sufficient that Gandhi know that non-violent protest would topple the regime, since followers had a high chance of being beaten – even in large numbers (Guha, 2017). To overcome the free rider problem, Gandhi needed to have the support of followers who would judge others negatively for not wanting to follow the behavior he favored.

Laws to stop undesirable social practices must make it difficult to follow those practices and behave appropriately. Dueling by the U.S. elite fell when a law was passed that prohibited those who duel from fulfilling their gentlemanly duty of holding public office. Notably, laws with harsher punishments that did not affect appropriateness were much less successful (Lessig, 1995).

Social information campaigns, in which a group's aggregate behavior is publicized, are a policy for increasing voluntary public good contributions such as voting (Gerber and Rogers, 2009), energy consumption (Allcott, 2011) and donations to public spaces (Alpizar et al., 2008). However, its success has varied widely (Kenny et al., 2011). This puzzling variation may be due in part to whether the intended audience believes respondents answered what they believed the surveyor wanted to hear – the so-called 'surveyor demand effect' (Melson et al., 2011, Kypri and Maclennan, 2011, Karp and Brockington, 2005, Gonzalez-Ocantos et al., 2012).

Second order conformity goes beyond formalizing insights found outside rational choice, and provides novel results on how leaders and policies affect pluralistic ignorance. A leader that redefines appropriate behavior can increase pluralistic ignorance by promoting behavior that supporters will follow reluctantly. This would not be the case under a common rational choice approach to leadership, in which individuals follow a leader if they believe she provides unbiased information (e.g. Cukierman and Tommasi, 1998, Canes-Wrone et al., 2001). Social information campaigns will be most successful when the intended audience knows surveyors avoided demand effects – for example by making it hard for respondents to know what the surveyors wanted to hear. However, the presence of a social campaign itself may make it hard to conceal the surveyor's preference.

Abrupt social change can also be profitably analyzed with second order conformity. When there is pluralistic ignorance, change can begin with an obscure, politically inactive and uninformed everyman. A brash action by someone whose private views are thought to reflect the majority view can be sufficient to reveal what the majority privately prefers. I will argue that Mohammed Bouazizi's impact on the Arab Spring followed this logic.

Rational explanations of pluralistic ignorance can be found in Kuran (1997), Bicchieri (2005), Chwe (2013), while psychological explanations are reviewed in Kitts (2003) and sociological explanations in Shamir and Shamir (2000). Rational choice explanations are related to a large rational choice literature on collective action (Olson, 1965, Hardin, 1982, Esteban, 2001, Siegel, 2009). Past models do not capture equilibrium misunderstandings about what the majority prefers, a point I develop in more detail in section 4.2. There is a pervasive idea in the social sciences that social norms are often efficiency enhancing (Fang, 2001, Arrow, 1970, Schotter et al., 1981, Sugden, 1989, Coleman, 1990), which is frequently rebutted by examples of inefficient social norms (Bicchieri, 2005, Elster, 1989, Edgerton, 1992, Centola et al., 2005). The paper shows how norms may not only be inefficient, but they may be sustained by individuals who believe they are cooperating. My work also relates to a literature on equilibrium misaggregation of preferences (Acemoglu et al., 2011, Bikhchandani et al., 1992, Banerjee, 1992, Golub and Jackson, 2010, Eyster and Rabin, 2010) . This is the first paper to provide a dynamic in which individuals who are motivated to conform reach mistaken beliefs over which preference to conform to. Finally, this work is related to an experimental literature on reluctance, which shows that pro-social motivations are often driven by what an individual thinks others expect (Dana et al., 2006, 2007, DellaVigna et al., 2012).

Section 2 sets up the basic model. Section 3 presents the main results. Section 4 shows second order conformity arises in a variety of settings, that the model's predictions can explain several empirical findings, and provides an experimental design to test a novel prediction of the model. Section 5 then extends the model to consider how principals or policies can affect which behavior is considered appropriate. Section 6 concludes. All proofs are in the appendix.

2 Setup

Here I set up the basic model, interpret it in terms of white segregationists, and define the equilibrium concept.

2.1 Model

There is a group with $|I|$ members and $2 + |I|$ periods, with generic members $i, j, k, l \in \{1, 2, ..., I\} \equiv I$, and generic period $t \in \{-1, 0, 1, 2, ..., I\} \equiv T$. Group members, equivalently referred to as 'individuals', or 'players', will be assigned a privately observed ideal point or private view $\phi_i \in \{0, 1\}$, which represents their private views. These ideal points are drawn from an unobserved population distribution $\theta \in \{0, 1\}$. Nature assigns the population distribution θ with probability $P(\theta) \in [0, 1]$. $P(\theta)$ will capture commonly held priors over the private views of the group, and I will refer to it as the 'context'. Group members are randomly drawn from this population, and are randomly indexed from 1 to I. Individuals' preferences $\phi_i \in \{0, 1\}$ are drawn with probability $P(\phi_i = \phi \mid \theta = \phi) > 1/2$ – population $\theta = \phi$ is more likely to produce groups with majority type ϕ. Although θ is not observed and ϕ_i is private information, both $P(\theta)$ and $P(\phi_i \mid \theta = \phi_i)$ are common knowledge.

Once the group is assigned, individuals take turns making decisions and judging each others' decisions. In period i, individual i chooses $a_i \in \{0, 1\} \equiv A$ after observing the history of play $h_i = \{a_1, a_2, ..., a_i\}$, with H the set of possible histories. I will refer to a_i as an 'action', 'decision', 'announcement' or 'opinion expression'. This sequential decision-making captures in a stylized way how public opinion is formed – individual announcements become public little by little, and past announcements inform later announcements. The timeline is denoted in Figure 1.

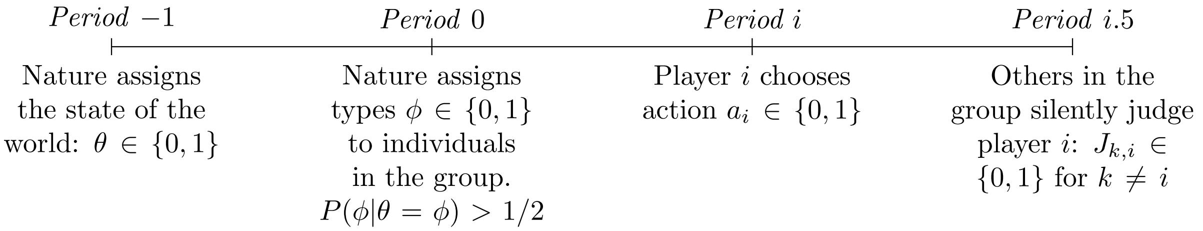

Period -1
Period 0
Period i
Period $i.5$
Nature assigns the state of the world: $\theta \in \{0, 1\}$
Nature assigns types $\phi \in \{0, 1\}$ to individuals in the group. $P(\phi|\theta = \phi) > 1/2$
Player i chooses action $a_i \in \{0, 1\}$
Others in the group silently judge player i: $J_{k,i} \in \{0, 1\}$ for $k \neq i$

Figure 1: Timeline

After i takes an action, and before $i+1$ does, all players other than i judge i based on her action. Let $-i$ be the set of players in the group other than i, or the 'group without i'. Player i trades off her ideal point with how she expects she will be judged by others. Let $\mathcal{J}_{j,i} \in [0, 1]$ be j's judgment of i. i's payoff is affected by the average of judgments $\mathbb{E}(\mathcal{J}_{k,i} \mid \hat{\phi}_{-i})$, where $\hat{\phi}_{-i}$ is the distribution of preference in the group without i. If i knew the true distribution of others' preferences, $\mathbb{E}(\mathcal{J}_{k,i} \mid \hat{\phi}_{-i})$ would be a deterministic function of a_i. Since i does not know the distribution, she maximizes her expected utility: $a_i^* = \arg\max_{a_i \in \{0,1\}} \mathbb{E}u(a_i; \phi_i, h_i, |I|) =$

$$\underset{a_i \in \{0,1\}}{\arg\max} -(\phi_i - a_i)^2 + \beta \mathbb{E}_{\hat{\phi}_{-i}} \left(\mathbb{E} \left(\mathcal{J}_{j,i} \mid \hat{\phi}_{-i} \right) \mid h_i, \phi_i \right) \tag{1}$$

where the outer expectation operator in the second summand is over all the distributions of preferences $\hat{\phi}_{-i}$. Whites are trading off acting according to their ideal point with acting according to how they think others in the group want them to act. I refer to the expectation over average judgments as the '*expected judgment*' or the '*social expectation*'. I assume $\beta \in (1, 2)$, as otherwise individuals would either always choose their ideal point or face very strong incentives to act according to how they expect to be judged. Note that I will generally use i for the current period's decision maker, and j for a judge of i.

When j judges i, j determines whether i's preference matches the preferences of the majority in the group without i. If judge j believes both i's preference and the average preference of the group without i are more likely to be $x \in \{0, 1\}$, then j judges i's preference to match those of the group. These judgments are not observed by others. Let ϕ_{-i}^m be majority preference in group $I \backslash i$. Then I can write j's judgment of i as

$$\mathcal{J}_{j,i} \begin{cases} = 1 & \text{if} \quad P(\phi_i = x \mid h_i, a_i, \phi_j) > 1/2 \;\&\; P(\phi_{-i}^m = x \mid h_i, a_i, \phi_j) > 1/2,\; x \in \{0,1\} \\ = 0 & \text{if} \quad P(\phi_i = x \mid h_i, a_i, \phi_j) > 1/2 \;\&\; P(\phi_{-i}^m = x \mid h_i, a_i, \phi_j) < 1/2,\; x \in \{0,1\} \\ \in [0,1] & \text{if} \quad P(\phi_i = x \mid h_i, a_i, \phi_j) = 1/2 \text{ or } P(\phi_{-i}^m = x \mid h_i, a_i, \phi_j) = 1/2,\; x \in \{0,1\} \end{cases}$$

The appendix shows how $\mathcal{J}_{j,i}$ can be derived from a maximization problem for judges in which they are trying to accurately guess whether i's preference match those of the majority of the group without i. Indeed, this judgment is the heart of second order conformity. If there were no uncertainty over the distribution of group preferences, judges would all agree on their judgment of i, and therefore i would not face any uncertainty in how she'll be judged. The result would be a simplified model of conformity as in Bernheim (1994).

As a shorthand, whenever there is no ambiguity I will say j 'believes preferences match' if $\mathcal{J}_{j,i} = 1$. Given the assumptions I make in the body of the text, the case where either beliefs over the individual $P(\phi_i = x \mid h_i, a_i, \phi_j)$ or the group's preference $P(\phi_{-i}^m = x \mid h_i, a_i, \phi_j)$ is equal to 1/2 will be a knife-edge scenario, so the value chosen for $\mathcal{J}_{j,i}$ is inessential. These assumptions will be relaxed in the appendix.

2.2 Interpretation

Before moving on to defining strategies and equilibrium, I will relate the setup of the model to the example given in the introduction.

There is a group of $|I|$ whites who will, one by one, announce whether they support segregationist housing policies ($a_i = 1$) or whether they do not ($a_i = 0$). Each individual i knows whether they privately support these policies ($\phi_i = 1$) or do not ($\phi_i = 0$), but they do not observe what others' private views are. There is common knowledge that group members were drawn either from a population which consists of mostly pro-segregationists ($\theta = 1$, with $P(\phi_i = 1 \mid \theta = 1) > 1/2$), or from a population which consists mostly of anti-segregationists ($\theta = 0$, with $P(\phi_i = 0 \mid \theta = 0) > 1/2$), but group members do not observe which population they were drawn from. Further, there is common knowledge over the probability they were drawn from population $P(\theta)$ – I refer to this probability as the 'context'. For example, a white supremacist rally has a strongly pro-segregationist context ($P(\theta = 1)$ is high).

Right after an individual i announces her opinion on segregationist policies, the other group members judge her. They are judging whether individual i's private views match those of the majority of the group. To make this judgment, judges take into account what they know about others' private views and about individual i's private views. Judges' information comes from the context, their own private views and from the private views revealed by past decision makers' actions. Since judges have different private views, their judgments may differ. In order to form social expectations, decision makers consider how they'll be judged on average given the different possible distributions of judges' private judgment.

2.3 Strategies and equilibrium

Denote by $G \equiv \{P(\theta), P(\phi = \theta \mid \theta), \beta, I\}$ the game defined above with context $P(\theta)$, informativeness of signal $P(\phi = \theta \mid \theta)$, weight on social expectations β and group size I. A pure strategy of i is $\alpha_i : H \times \{0,1\} \to \{0,1\}$, which maps a history h_i and a type ϕ_i to an action. I focus on pure strategies here, and consider mixed strategies in the appendix. I will be interested in Perfect Bayesian Equilibrium, and use the D1 criterion to refine out-of-equilibrium beliefs (Fudenberg and Tirole, 1991). I say that k *pools* on $x \in \{0,1\}$ if $\alpha_k(h_k, \phi_k) = x$ for all ϕ_k, and I call k a *withholder* if k pooled or whose turn to make an announcement has not come at some history h_i. If $\alpha(h_i, \phi_i) \neq \alpha(h_i, 1 - \phi_i)$, I say i *separates*, and I call k a *revealer* if k separated at some history h_i.

An equilibrium is *non-reversing* if, whenever i of type ϕ_i chooses an action with certainty, ϕ_i chooses the action corresponding to her type: $\alpha_i^*(h_i, \phi_i) = \phi_i$. It is easy to show that whenever there is a 'reverse' separating strategy in which type ϕ_i chooses action $1-\phi_i$, there exists a non-reverse separating strategy. A *preferentially pure* equilibrium is an equilibrium in which i chooses a separating or a pooling strategy unless no pure strategies are best responses. An *informative* equilibrium selects the equilibrium strategy that reveals most information about a player's type – if pooling and separating can be sustained in equilibrium, I select the separating strategy.

Definition 1. *An* ***equilibrium*** *is a non-reversing, preferentially pure, informative Perfect Bayesian Equilibrium with the D1 criterion for refining out-of-equilibrium beliefs.*

Requiring the equilibrium to be preferentially pure and informative means that I select a separating equilibrium whenever it exists, and a semi-separating equilibrium only if no pure strategy equilibrium exists. This allows us to obtain results that are biased against information stagnation (i.e., subjects not revealing their type), while retaining tractability.

3 Dynamics And Pluralistic Ignorance

This section presents the main results of the model. After laying some groundwork, I provide a result that describes the equilibrium dynamics of the model, as well as some comparative dynamics. This result is the cornerstone of all other results. In subsection 3.2, I give a formal definition of pluralistic ignorance and show how the probability of pluralistic ignorance depends on group size and context. I further show how pluralistic ignorance can lead to a perverse breakdown of collective action in which individuals reluctantly do what they mistakenly believe is socially optimal.

3.1 Equilibrium Dynamics

I first define strong and uninformative contexts, which are subsets of $P(\theta)$ I will focus on. I then introduce some terms to describe equilibrium dynamics, and conclude this subsection with a characterization of equilibrium dynamics.

3.1.1 Defining Strong and Uninformative Context

I begin with a more specific way of thinking about context. I'll focus on contexts where individuals have a 'strong' sense of what the majority preference is, and on 'uninformative' contexts where individuals are very uncertain of the majority preference.

Definition 2. *Let* $\phi \in \{0,1\}$ *throughout.*

- *The context $P(\theta)$ is **uninformative** if and only if, at period 0, an individual of type ϕ believes a randomly chosen group member is of type ϕ:*

$$\frac{P(\theta=0)}{P(\theta=1)} \in \left(\frac{P(\phi_i=0 \mid \theta=1)}{P(\phi_i=0 \mid \theta=0)}, \frac{P(\phi_i=1 \mid \theta=1)}{P(\phi_i=1 \mid \theta=0)}\right)$$

- *Suppose that if $j \neq 1$ at period 0 knew that the population was $\theta=\phi$, she would believe the majority of players other than 1 are of type ϕ: $P(\phi^m_{-1}=\phi \mid \phi_j = 1-\phi, \theta=\phi) > 1/2$. Then I say **private views yield**. Otherwise, I say **private views sway**.*

- *Suppose $j \neq 1$ with private view ϕ observed a second signal of type ϕ at period 0, and private views yield. Then I say the context $P(\theta)$ is **strongly $\mathbf{1-\phi}$ with yielding private views** if and only if j at period 0 believes most players without 1 are of type $1-\phi$:*

$$\frac{P(\theta=\phi)}{P(\theta=1-\phi)} \in \left(0, \left[\frac{P(\phi_i=\phi \mid \theta=1-\phi)}{P(\phi_i=\phi \mid \theta=\phi)}\right]^2 \left\{\frac{P(\phi^m_{-1}=1-\phi \mid \phi_j=\phi, \theta=1-\phi)-1/2}{1/2-P(\phi^m_{-1}=1-\phi \mid \phi_j=\phi, \theta=\phi)}\right\}\right)$$

- *Suppose player 1 observed $|I|$ signals of type ϕ at period 0, and private views sway. Then I say the context $P(\theta)$ is **strongly $\mathbf{1-\phi}$ with swaying private views** if and only if player 1 believes a randomly chosen player is of type $1-\phi$:*

$$\frac{P(\theta=\phi)}{P(\theta=1-\phi)} \in \left(0, \left[\frac{P(\phi_i=\phi \mid \theta=1-\phi)}{P(\phi_i=\phi \mid \theta=\phi)}\right]^I\right)$$

- *I say the context $P(\theta)$ is **strongly $\boldsymbol{\phi}$** if it is either strongly ϕ with yielding private views or strongly ϕ with swaying private views.*

Uninformative contexts and contexts that are strongly $\phi \in \{0,1\}$ are of first order importance, since they respectively capture high uncertainty and high certainty about the distribution of preferences from which the group was drawn. Furthermore, equilibrium strategies are pure in uninformative contexts and in contexts that are strongly $\phi \in \{0,1\}$ with yielding private views. I will focus on pure strategy equilibria in the text for clarity. The appendix substantially generalizes the results, and discusses the technical complications that arise from considering mixed strategies.

There is a robust set of parameters for which private views yield. Private views sway when the information $j \neq 1$ has about her own preference ϕ_j is sufficient for her to believe that, no matter what she knows about the population, the majority preference of the group other than 1 is ϕ_j. For example, if the group is of size 2 ($|I|=2$), then player 2 is certain that the majority preference of the group other than 1 is equal to her own preference, ϕ_2.

However, by the law of large numbers there is always a value of $|I|$ large enough that, if the population was $\theta \neq \phi_j$, j would believe the majority of the group other than 1 is most likely of type θ. Therefore, private views yield for $|I|$ large enough, or for any $|I| \geq 4$ as long as it is sufficiently likely the majority preference of the group other than 1 is equal to the population θ ($P(\phi = \theta \mid \theta)$ sufficiently high).

3.1.2 Action Lead and Phase: Terms for Describing Equilibrium Dynamics

Now that I've defined the contexts I will be focusing on, I follow Ali and Kartik (2012) in defining some terms which will allow me to talk about the equilibrium dynamics. In order to keep track of the information that has been revealed by players' actions, I define the *action lead* for action 1 at period h_i, $\Delta(h_i)$:

$$\Delta(h_i) \equiv \sum_{k=1}^{i-1} \mathbb{1}\{a_k = 1\} - \mathbb{1}\{a_k = 0\}$$

This summary statistic keeps a tally of how many individuals have chosen action 1 at period i, and subtracts how many have chosen action 0. I will refer to $-\Delta(h_i)$ as the action lead for action 0.

I will show that the equilibrium strategies follow a threshold rule, in which individuals in the first periods of the the game separate until the action lead for action $\phi \in \{0, 1\}$ is greater or equal than a certain value, at which point all players pool on ϕ. In order to capture this dynamic, I define the *threshold action leads* for each action, $n_\phi(i; P(\theta))$, such that i is the first player to pool on ϕ only if the action lead for action ϕ reaches $n_\phi(i; P(\theta))$. That is, for any $i \geq 1$, $n_1(i; P(\theta))$ is the smallest action lead such that for any history h_i with $\Delta(h_i) = n_1(i; P(\theta))$, the unique equilibrium strategy is for i to pool on 1. $n_0(i; P(\theta))$ is defined analogously.

Given the equilibrium dynamics, players will either be in a phase of the game where they are separating, or in a phase where they are pooling on $\phi \in \{0, 1\}$. In order to capture this, I define the *phase map* $\Psi : h_i \to \{L, 0, 1\}$, for all $i > 1$, by

$$\Psi(h_i) \equiv \begin{cases} \Psi(h_i) & \text{if } \Psi(h_{i-1}) \in \{0,1\} \\ 1 & \text{if } \Psi(h_{i-1}) \in \{0,1\} = L \text{ and } \Delta(h_i) = n_1(i) \\ 0 & \text{if } \Psi(h_{i-1}) \in \{0,1\} = L \text{ and } \Delta(h_i) = n_0(i) \\ L & \text{otherwise,} \end{cases}$$

with the initial condition $\Psi(h_1) = x \in \{0, 1\}$ if the first agent's equilibrium strategy is to pool on x, and $\Psi(h_1) = L$ otherwise. If the phase map at period $i-1$ is $\Psi(h_{i-1}) = \phi \in \{0, 1\}$, then player i pools on ϕ and the phase map at period i is $\Psi(h_i) = \phi$. Suppose the phase map at period $i-1$ is $\Psi(h_{i-1}) = L$. Then if at period i the action lead for action ϕ reaches the threshold action lead, i pools on ϕ ($\Psi(h_i) = \phi$), and otherwise i separates ($\Psi(h_i) = L$).

3.1.3 Characterizing Equilibrium Dynamics

I can now state the following characterization of equilibrium dynamics:

Result 1. *If context is uninformative or if context is strongly $\phi \in \{0,1\}$ with yielding private views, there is a unique equilibrium. In either case, the phase map $\Psi(h_i)$ is such that $i \in I$ separates if $\Psi(h_i) = L$, and pools on $x \in \{0,1\}$ if $\Psi(h_i) = x$.*

I can further characterize equilibrium strategies depending on the context:

- *If the context is uninformative, the first player separates: $\Psi(h_1) = L$. Furthermore,*
    - *$n_\phi(i; P(\theta)) > 0$. That is, for i to pool on ϕ there must be more revealed signals of type ϕ than of type $1-\phi$.*
    - *If $P(\theta = \phi) > 1/2$, $n_\phi(i; P(\theta)) \leq 2$ and $n_{1-\phi}(i; P(\theta)) \leq 3$. That is, the threshold action lead for ϕ is less or equal to 2 if the population is more likely to be ϕ. Further, the threshold action lead for $1-\phi$ is less or equal to 3.*
    - *$n_\phi(i; P(\theta))$ weakly decreases in β, $P(\theta = \phi)$ and i, and weakly increases in $|I|$ for $\phi \in \{0,1\}$. The larger the weight individuals put on following social expectations or the more likely the population is ϕ, the smaller the threshold action lead for action ϕ. The threshold action lead of ϕ becomes weakly smaller for players who make decisions later in the game. The threshold action lead of ϕ is weakly larger for larger groups.*
- *If the context is strongly $\phi \in \{0,1\}$ with yielding private views, all players pool on ϕ: $\Psi(h_1) = \phi$.*

I now provide an informal discussion of the logic behind Result 1. I analyze the uninformative context first, and then the much more straightforward strongly ϕ context with yielding private views. A reader interested in the applications of the model may wish to skip this discussion. For those interested in the technical details, the appendix provides a more detailed discussion of the results under a relaxed set of assumptions.

Before proceeding, it is perhaps worth reminding the reader that social expectations for player i are what i expects is the average judgment players other than i make about whether 'preferences match', which means that the majority preference in the group other than i is equal to the preference i is most likely to have. Whoever is making a decision is a 'decision maker' who gets utility from her social expectation, and the rest of the group are her judges.

Uninformative context. When the context is uninformative, player 1 of type ϕ believes most judges are of type ϕ, and that judges of type ϕ believe a randomly chosen group member is most likely of type ϕ. Further, by revealing her type, all judges would have stronger beliefs that other group members are of type ϕ. Therefore, whether she reveals her type by separating, or does not reveal her type by pooling, she believes at least half of judges will believe preferences match: most believe most group members other than 1 are of type ϕ, as they would believe player 1 is if she pooled. Since $\beta \in (1, 2]$, players will choose their ideal point whenever they believe half of judges believe preferences match. But then player 1 of type ϕ would not deviate from a separating strategy, nor from a strategy of pooling on ϕ for any out-of-equilibrium beliefs. I can then use the D1 criterion to conclude that, if player 1 pools on ϕ, a deviation must come from player 1 of type $1 - \phi$. But by symmetry, player 1 of type $1 - \phi$ believes at least half of judges believe preferences match if she reveals to be of type $1 - \phi$. Therefore, player 1 of type $1 - \phi$ deviates from a strategy of pooling on ϕ.

Player 2 will therefore observe player 1's type before making a decision, and knows that others observed it as well. Player 1's revealed type affects group members' belief over the population, and yields direct information about the preference of player 1, who is one of the judges of player 2. If player 2 is of the same type ϕ as player 1, she will then have stronger beliefs than player 1 did that most judges are of type ϕ, and that judges believe that other judges are most likely of type ϕ. Her incentives to reveal her type therefore increase. In contrast, if player 2 is of type $1 - \phi$, she will have observed a signal ϕ and a signal $1 - \phi$. Incentives to deviate from separating are higher for player 2 of type $1 - \phi$, as are incentives to pool on ϕ. These incentives depend on the context: they increase with $P(\theta = \phi)$.

More generally, the first players to make a decision will separate, which reveals information to other decision makers and judges about the distribution of preferences in the group. The proof proceeds by calculating the social expectations for each integer value of the action lead $\Delta(h_i)$. The analysis is complicated by the fact that the single crossing property generally does not hold, nor does the D1 criterion always yield a unique out-of-equilibrium belief. In general however, as the action lead of action ϕ increases, the incentives to pool on ϕ increase – both decision makers and judges believe group members are more likely to be of type ϕ, so the social expectation of revealing ϕ increases and of revealing $1 - \phi$ decreases. If i pools on ϕ, then $i + 1$ pools on ϕ, since $i + 1$ has the same information about judges' preferences and beliefs than i did.

Comparative dynamics follow from this logic. First, a lower weight on social expectations (lower β) implies that i of type $1-\phi$ must receive a higher social expectations benefit in order to pool on ϕ. Second, for a fixed history h_i such that player i separates ($\Psi(h_i) = L$), a larger i or a smaller $|I|$ implies that group members have more certainty about the distribution of group preferences. Therefore, group members will hold stronger beliefs that the group majority is equal to whichever action $a \in \{0, 1\}$ has a positive action lead, which increases incentives to pool on a.

Individuals may begin to pool on ϕ when as few as the first two individuals have revealed to be of type ϕ, independently of group size. This result is related to the literature on 'herding' (Banerjee, 1992) or 'information cascades' (Bikhchandani et al., 1992), in which the signals revealed by a few types leads others to not reveal their type. In many of those models, the state of the word impacts individuals' utility directly – a classic example is of individuals looking to others' choices to figure out which of two restaurants is objectively better. In my second order conformity model, the state of the world provides information about the distribution of preferences in a group, which individuals use to calculate social expectations. This direct interest in the group's preferences means that the size of the group $|I|$ and the number of individuals who have taken an action i affects the threshold action leads, unlike in past herding models such as Ali and Kartik (2012), Bikhchandani et al. (1992), Banerjee (1992)

Context strongly ϕ with yielding private views. When the context is strongly $\phi \in \{0, 1\}$ with yielding private views, player 1 of type $1 - \phi$ believes most judges are of type ϕ, that all judges believe a randomly chosen group member is of type ϕ, and furthermore knows that $1 - \phi$ type judges will believe most judges are of type ϕ even if they saw a second signal of type $1 - \phi$. Therefore, social expectations from revealing $1 - \phi$ are zero (no judges would believe preferences match), while social expectations from pooling or revealing ϕ are one (all judges would believe preferences match). But since $\beta \in (1, 2]$, players do not choose their ideal point if the difference in social expectations is one, so player 1 pools on ϕ.

3.2 Pluralistic Ignorance in Large and Small Groups

Result 1 described equilibrium dynamics for a wide range of parameter values. In this subsection, I relate these dynamics to pluralistic ignorance. I will first define pluralistic ignorance as a situation where individuals are reluctantly acting in a way they mistakenly believe is the majority preference. Then I show that the probability that pluralistic ignorance arises in a group depends on the context and the group size. I will discuss the significance of this result in Section 4. I then show how pluralistic ignorance can lead to a perverse failure of collective action.

3.2.1 Defining Pluralistic Ignorance

I say decision maker i of type ϕ_i '*acts reluctantly*' at history h_i if her equilibrium action differs from her ideal point: $\phi_i \neq \alpha_i^*(\phi_i, h_i)$. If i is acting reluctantly, i would change her behavior if she did not care about social expectations ($\beta = 0$). Whites who act reluctantly announce they support segregationist housing policies when they don't support segregation, or announce they don't support housing policies when they do support segregation.

Definition 3. *I say there is* ***pluralistic ignorance*** *for realization* $\overline{\phi} \equiv (\phi_1, \phi_2, ...\phi_I)$ *if and only if (a) most agents act reluctantly,* $P(\phi_i \neq \alpha_i^*(\phi_i, h_i) \mid \overline{\phi}) > 1/2$*, and (b) most agents believe most others are not acting reluctantly:* $\mathbb{E}\big(P(\phi_j \neq \alpha_j^*(\phi_j, h_j) \mid \phi_i, h_i, j \neq i) \mid \overline{\phi}\big) < 1/2$.

If there is pluralistic ignorance, most individuals in the group act reluctantly, but believe most others are not acting reluctantly. Pluralistic ignorance among whites would be for most to reluctantly announce they (do not) support segregation, and for most to believe most are announcing they (do not) support segregation non-reluctantly. Substituting the strict inequalities for weak inequalities in the definition leads to qualitatively similar results. The existence of pluralistic ignorance in the model is a corollary of Result 1.

3.2.2 Context, Group Size and Pluralistic Ignorance

I will now provide a result that sheds some light on how group size and the context affects pluralistic ignorance. Note that since pluralistic ignorance is defined for a realization $\overline{\phi}$ of types, the probability of pluralistic ignorance for a game G is the proportion of realization of types that result in pluralistic ignorance.

Result 2. *There exist thresholds* $\tau_1(G) \geq 2$ *and* $\tau_2(G) < \infty$ *such that*

- *For all* $I \geq \tau_2(G)$*, the probability of pluralistic ignorance is lower when the context is strongly* ϕ *and* $P(\theta = \phi)$ *is above some threshold* p *than for any other context – i.e., for any other value of* $P(\theta)$*. When* $P(\theta)$ *is above* p*, the probability of pluralistic ignorance is lower if* $I \geq \tau_2(G)$ *than if* $I \leq \tau_1(G)$*.*

- *For all* $I \leq \tau_1(G)$*, the probability of pluralistic ignorance is lower when the context is uninformative than for any other context – i.e., for any other value of* $P(\theta)$*. When the context is uninformative, the probability of pluralistic ignorance is lower if* $I \leq \tau_1(G)$ *than if* $I \geq \tau_2(G)$*.*

Result 2 shows that the conditions under which we should find pluralistic ignorance will be very different depending on whether we are considering a large group interaction or a small group interaction. I now discuss this result, and relate it to the motivating example of whites' segregationist preferences.

The context $P(\theta)$ captures the common knowledge over the type of people the group is likely composed of. Notice there are two levels of uncertainty – the first over from which population the group was drawn, and the second over the group which was generated from this population. When beliefs over the population $\theta \in \{0, 1\}$ are arbitrarily strong, group members hold arbitrarily strong beliefs that other group members are drawn from θ. From Result 1, we know that strong enough beliefs over θ leads to pooling on θ. Furthermore, by the law of large numbers, the majority preference of a sufficiently large group arbitrarily approximates the majority preference of the population θ – the probability of pluralistic ignorance is therefore arbitrarily small for arbitrarily large groups with arbitrarily strong beliefs over θ. However, small groups with strong beliefs over θ also pool on θ, and their majority preference will often deviate from the majority preference of the population. To illustrate, suppose there is a strongly pro-segregationist context among whites in the United States. Public opinion among whites would accurately reflect the majority pro-segregationist view, even if some anti-segregationist whites reluctantly expressed pro-segregationist opinions. However, in a conversation among a few whites, the fact that anti-segregationists reluctantly express pro-segregationists views will more often lead to pluralistic ignorance in the group, since it is more likely that the group is composed of mostly anti-segregationists.

When the context is uninformative, we know by Result 1 that the first individuals to express their opinion will reveal their private view. Further, this may lead individuals to pool on whichever view was expressed frequently by those first movers. But when the group is large, this means that the opinion expressed by everyone in the group may be determined by whichever preferences the first movers happened to have. This will often not reflect the majority view, leading to pluralistic ignorance. However, when the group is small, the 'first movers' are a large proportion of the group – when the group is of size two, the first mover is half of the group. An uninformative context would allows whites in small groups to reveal their type, and thus avoid mistakenly acting according to what they think others want. However, this may lead to pluralistic ignorance in forming public opinion among all whites in the U.S., as whites rely on the opinions of a few other whites to learn about what most whites want.

3.2.3 Pluralistic Ignorance and Failures of Collective Action

I now turn to discussing the inefficiency of pluralistic ignorance, and how it may lead to a perverse failure of collective action. In order to do so, modify the utility function to allow for a *cooperative action* $a_i = 1$.

$$-(\phi_i - a_i)^2 + \beta \mathbb{E}\left(\mathcal{J}_{j,i} \mid \hat{\phi}_{-i}\right) + \gamma \phi_i \sum_{k \in I} a_k \tag{2}$$

If (2) has $\gamma > 0$, I say it is a *cooperatitve utility function.* To fix ideas, focus on the case of whites' support for segregation policies. $\gamma > 0$ means that support for segregationist housing policies ($a_i = 1$) is a public good for pro-segregation whites, perhaps because public opinion helps shape policy. If (2) has $\gamma = 0$, the utility function is as originally formulated and I say it is the *original utility function.*

I can now ask whether an equilibrium can sustain the following perverse failure of collective action: individuals reluctantly take the action that yields positive externalities, think that this action is socially optimal, and are mistaken about its social optimality.

Definition 4. *There is* ***cooperative pluralistic ignorance*** *when (a) agents have cooperative utility functions, (b) most agents reluctantly choose action 1,* $P(\alpha_i^*(0, h_i) = 1 \mid \overline{\phi}) > 1/2$*, and (c) most believe most others are not acting reluctantly:* $\mathbb{E}\big(P_i(0 \neq a_j^*(0, h_j) \mid \phi_i, h_i) \mid \overline{\phi}\big) < 1/2$*.*

I say the cooperative pluralistic ignorance is ***inefficient*** *if and only if*

$$\sum_{i \in I} -(\phi_i - a_i^*)^2 + \gamma \phi_i \sum_{k \in I} a_k^* < 0$$

To illustrate, if there is cooperative pluralistic ignorance most anti-segregationist whites (individuals of type 0) believe they are contributing to a public good by making a pro-segregationist announcement ($a_i = 1$) since they mistakenly believe most in the group prefer segregationist policies. If the benefit pro-segregationists get from anti-segregationists' reluctant cooperation is lower than the cost to anti-segregationists' utility, I say this is inefficient.

Result 3. *Individuals will be in cooperative pluralistic ignorance with a positive probability if (a)* $P(\theta = 1) > 1/2$*, the context is uninformative, and* $I \geq 5$*, or (b) the context is strongly 1. It will be inefficient for some* γ *sufficiently small.*

Pro-segregationists are more incentivized to reveal their type as γ increases, since they may influence others to also make a pro-segregation announcement. Anti-segregationists face the same incentives as with the original utility function, so for $P(\theta = 1)$ sufficiently large there is a positive probability that anti-segregationists pool on making a pro-segregationist announcement.

Result 3 shows that there can be a situation in which individuals are pooling on an action they believe is socially efficient – which would be the case if most players were pro-segregationist as whites believe in equilibrium – although it is in fact inefficient.[3] In the next section I will argue that this situation captures an important aspect of pluralistic ignorance.

4 Discussion of Basic Results

Sections 2 and 3 focused on the motivating example of white segregationist preferences for ease of exposition. I now argue that second order conformity can be used to explain a much wider range of phenomena. Therefore, this section will be somewhat eclectic, arguing about how the model should be thought of, how it relates to other models, and ending with a series of empirical examples that are illuminated by a second order conformity approach.

The roadmap for this section is as follows. In subsection 4.1 I argue that the model can be used to capture important concepts traditionally used outside of game theory such as appropriate behaviors and context dependent preferences, and also provides a novel definition of norms. I then argue in subsection 4.2 that the features of pluralistic ignorance captured by the model have not been captured by past formalizations. I also discuss how the notion of norm associated to alternative definitions of pluralistic ignorance differs from the one that arises with second order conformity. Subsection 4.3 provides two ways of thinking about the population θ in the model: as the empirical distribution of a large collection of groups, or as a shared subjective belief over the probability with which group members have a given preference. This distinction is useful for thinking through a range of empirical examples presented in subsection 4.4. I will discuss how the model applies to public opinion formation, but I will also consider perhaps surprising applications such as to norm heterogeneity, interstate crisis bargaining, and examples outside of politics such as dating, gift giving and drinking. Readers may skip any of these examples without loss of continuity.

4.1 Appropriateness, Norms and Context-Dependent Preferences

Here I argue that the second order conformity model can capture some key concepts that are typically found outside the game theoretic literature, such as appropriateness and context dependent preferences. I will also provide a definition of norms which I will argue below differs from past formalizations. In fact, I will argue that the novelty in this conception of

[3] Another way to make the same point without changing the analysis is to add to i's original utility function a term which increases in the actions equal to i's ideal point by *past* players $j < i$. Then either action provides a public good for others, although players' optimal choice is unaffected and therefore I can use Result 1 to characterize the equilibrium. If instead i's utility were affected by *all* players' actions, i would have incentives to influence others. I conjecture that the results will be qualitatively similar, but the analysis is beyond the scope of the paper.

norms allows me to capture features of pluralistic ignorance which have not been adequately captured.

Second order conformity naturally gives rise to thinking about the logic of appropriateness, a logic of behavior in which 'actors seek to fulfill the obligations encapsulated in a role, an identity, a membership in a political community or group, and the ethos, practices and expectations of its institutions' (March and Olsen, 2004). Recalling that $\mathcal{J}_{j,i}$ is equal to one if j believes preferences match, and α_i is i's pure strategy, consider the following:

Definition 5. *Judge j of type ϕ_j believes i's behavior $a \in \{0,1\}$ after history h_i is* ***appropriate*** *if it makes j believe preferences match in equilibrium: $\mathcal{J}^*_{j,i}(h_i, \alpha^*_i = a, \phi_j) = 1$.*

Individual i is motivated to choose an action that makes members of her group believe preferences match, and thus fulfill the obligations encapsulated in the group's expectations. In a group of whites, j will consider i's announcement over segregationist policy appropriate if what the announcement reflects about i's preference matches what j thinks most whites prefer. Appropriate behavior will depend on the group with whom one interacts, so the same behavior in different situations will have a different 'social meaning' – the distribution of judgments about the behavior. If through their sequential decisions individuals agree on what the group's majority preference is, then all group members believe the same behavior is appropriate, and I call that behavior the norm.[4]

Definition 6. *If at history h_i all group members of either type believe $a \in \{0,1\}$ is appropriate, then a is a* ***norm****. If t is the first period in which a is a norm, I say the norm* ***emerged*** *at period t.*

By Result 1, if and only if individuals pool on an action in equilibrium, that action is a norm. Moreover, a norm emerges the first period an individual pools on an action. Notice that endogenous norm emergence allows us to endogenize preferences in a particular sense: an individuals' preferred behavior at the beginning of the game may differ from preferred behavior at the end, after the group has interacted. Preferences are context-dependent in this sense. Several authors have claimed that formal theories have not captured preferences that are endogenous to social interaction, which is key to analyzing certain social situations (Sunstein, 1999, Wendt, 1992, Finnemore and Sikkink, 1998, March and Olsen, 2004). The paper will provide several applications where preference complementarities can capture some of the insights of this literature.

[4] Of course, this definition could be modified so that a behavior is a norm if a certain fraction of whites believe it is appropriate.

4.2 Alternative Approaches to Pluralistic Ignorance

I now present a more detailed review of the literature on pluralistic ignorance. I will argue that past formalizations do not capture the features of pluralistic ignorance that the second order conformity model does, and discuss where an alternative formalization that has not been considered in the literature fall short.

4.2.1 Past Approaches

'Pluralistic ignorance' has received several definitions in both empirical and theoretical work. In this paper, I have focused on the original use of the phrase, as captured by Katz and Allport (1931) and O'Gorman (1975): 'a majority of group members privately reject a norm, but incorrectly assume that most others accept it, and therefore go along with it.' I have argued that inefficient cooperative pluralistic ignorance captures three features of this definition of pluralistic ignorance: many individuals are acting reluctantly, believe that most are not acting reluctantly, and they believe their reluctant action is providing a public good. I now argue that alternative formulations of pluralistic ignorance do not capture these features. These formulations have an associated definition of norm, which I discuss.

One approach to pluralistic ignorance is to suppose individuals want to take a certain action a as long as enough others also take that action (so called 'strategic complementarities', or coordination problems), but are misinformed about the distribution of action thresholds (e.g. Chwe, 2000, Kuran, 1997). In a strategic complementarities model, a norm is sometimes defined as the action individuals are coordinating on in equilibrium (e.g. Young, 1993). Models with strategic complementarities have not captured individuals mistakenly believing that most group members are not acting reluctantly. Some of these models feature equilibrium certainty over others' preferences, such as Kuran (1997), so there is no misunderstanding about what others prefer. Chwe (2000) presents a model where coordination fails because some individuals don't know what others' action threshold is, which does not provide a natural sense in which individuals hold mistaken beliefs over others' preferences. An important class of models, called 'global games' (Carlsson and Van Damme, 1993), assume incomplete information over the action threshold needed to successfully coordinate (e.g. Edmond, 2013).Although they may reach mistaken conclusions about the threshold needed for successful coordination, individuals typically know they would all prefer to coordinate if the action threshold is low enough.

A second definition of pluralistic ignorance comes from models where individuals are motivated to signal that their type matches what others consider a desirable preference (e.g. Bernheim, 1994, Bénabou and Tirole, 2011, Benabou and Tirole, 2012, Ellingsen et al., 2008).[5] Unlike with second order conformity, there is an exogenously given 'appropriate' behavior which provides reputational benefits and possibly positive externalities – norms in these models are therefore exogenous. Since there is common knowledge about what the norm and appropriate action is, individuals will not incorrectly believe others are acting reluctantly.[6] Benabou and Tirole (2012) give a different definition of pluralistic ignorance within their model of exogenous norms, which is that there is equilibrium misinformation about the intensity over the desirable preference. Pluralistic ignorance in their setup is assumed, not derived as an equilibrium outcome.

4.2.2 An Alternative Setup: Altruists Uncertain Over Group Preferences

An alternative way to set up the model would have been to assume that individuals were pure altruists who wanted to choose the action that most group members wanted, but were uncertain over the population from which the group was drawn. This approach would have avoided the complications that arise from having individuals care about how they are judged by others, and I claim would have yielded Results 1, 2 and 3. However, evidence from the lab has shown that individuals' pro-social motivation is better explained by social expectations than by altruism (Krupka and Weber, 2013, Dana et al., 2007, 2006).

For example, consider the experiment by Dana et al. (2006). They show that after playing an anonymous dictator game where a 'dictator' split $10 dollars with a recipient, many dictators subsequently preferred to take $9 and 'exit' the game without the second player knowing that a dictator game had been played. An altruist would not have exited the game, since she could have allocated $1 to the other player and been strictly better off. Further, when a different group of dictators was told that their recipient would not know where the money assigned to them came from, they did not exit the game. The interpretation is that when the recipient knew a dictator game was played, the dictator did not want to be judged harshly and therefore gave reluctantly. However, when the recipient would not judge the dictator because she did not know a dictator game was played, she did not act reluctantly and therefore did not prefer to exit the game. The definition of reluctance captures this tension between being judged and wanting to follow an ideal point.

[5] Like in my model, Sliwka (2007) assumes that conformists do whatever the majority preference is. However, the model has only an informed principal and an agent, and the results focus on whether the principal reveals her information truthfully.

[6] Although Bernheim (1994) allows for a more flexible reputational function, it is also exogenously given and commonly known. In my model, *which* action yields reputational benefit is endogenously determined.

4.3 A Frequentist and a Subjectivist Interpretation of the Model

In this subsection I highlight two ways of thinking about the model that will be helpful when considering the examples I discuss below.

In a ‘frequentist’ interpretation, the population can be thought of as capturing a large number of groups, each member of a group only interacting with other group members. For example, it could be many groups of whites, each of whom is having a conversation about segregationist policies. The population here can be thought of as the distribution of preferences across the groups, with the understanding that an individual believes that its group was drawn from this population. An alternative, ‘subjectivist’ interpretation is that there is one group whose members are deciding how to act with other group members in a given situation, and the population provides the shared expectations for that situation. For example, a group of whites may meet and start a conversation in a specific situation. The population can then be thought of as representing the type of people that are likely to appear in that specific situation. To repeat an earlier example, it is more likely that the group of whites will be more pro-segregation at a white supremacist rally than at a college campus. Although these two ways of thinking about what the model is capturing can be complementary, keeping the distinction in mind will be helpful, and I will sometimes insist on one of them.

4.4 Examples of Second Order Conformity

In this subsection I provide empirical examples where second order conformity can yield insights. In section 4.4.1 I provide further examples of public opinion formation to which the model can be applied. I present some evidence in favor of the large group prediction of the model, and provide an experimental design to test the small group prediction. In section 4.4.2 I argue that the model can be used to study norm variation across groups. In section 4.4.3 I argue inadvertent wars are the result of states acting according to what they think other states want, with an explicit model provided in the appendix. I argue that second order conformity provides a bridge between rationalist and constructivist approaches to interstate crisis bargaining. In section 4.4.4 I apply the model to examples outside of politics such as dating, gift giving and drinking. The interested reader may focus on just some of these examples without loss of continuity.

4.4.1 Public Opinion Formation: Examples and an Experimental Design

I have argued that the model can be thought of as a stylized model public opinion formation, in either small or large groups. The opening example regards whites’ views on segregation, but the model applies to any topic which is discussed and where subjects feel like their announcements must match those of some group. For example, Mildenberger and Tingley (2016) studies pluralistic ignorance in the context of climate change, and Van Boven (2000) studies it in the context of political correctness. In both of these examples, 'small' groups, such as a group of friends, have conversations in which they express their opinions and may be judged by their answers. However, these topics are also important public opinion topics, in which the whole society forms a 'large' group.

Result 2 indicates that we should expect the distribution of pluralistic ignorance to differ in large and small groups. I will now report some evidence by Shamir and Shamir (2000) in favor of the large group approach to public opinion formation, and propose an experimental design for testing the implications of the small group approach.

Pluralistic Ignorance Across Public Opinion Topics. I have argued that the model as applied to large groups captures public opinion formation. Result 2 implies that, in large groups, we should expect more pluralistic ignorance when there is certainty about the population from which group members are drawn. Some evidence in favor of this prediction is provided by Shamir and Shamir (2000), who look at 24 public opinion issues in Israel. They find that the more visible the issue – a range of measures which capture certainty over θ – the less likely there will be a misperception over the distribution of preferences. For instance, they show that public opinion issues which may be obscure – in the sense of not receiving much media coverage – may nonetheless display low pluralistic ignorance if individuals can use 'proxy' distributions. Unknown public opinion issues where the distribution of preferences can be approximated by the distribution of political affiliation, for example, may lead to precise estimates of the actual distribution. These proxy distributions are a way of capturing what the population refers to in the model.

Experimental Design. Although we do not have much evidence for the implication of Result 1 on small groups, it can be tested in an experimental setting. Here I sketch the experimental design. Individuals can be invited to discuss a topic online where social expectations typically matter, for example climate change. In a baseline, private interview, subjects are asked about their views on climate change. Some time later (to avoid bias), subjects would go online to chat with each other. The chats will have around 4 people. Before the chat, they will be given information about the distribution of preferences from which the subjects were chosen. This distribution will be the proxy for the population. The chat protocol would allow each person to take turns giving their own opinion on climate change. After each person gives an opinion, others privately write down their reaction only to that opinion. However, they cannot publicly comment on what others have said. Once everyone has given their opinions, they are asked to estimate the distribution of opinions in the group. Afterwards, and in order to increase the weight on how others judge their behavior, subjects are shown what others wrote about their opinion. With this protocol we can test whether varying the distribution of preferences in the population affects whether individuals' statements reflect their baseline opinions, and when pluralistic ignorance appears most frequently.

4.4.2 Using the Model to Study Norm Heterogeneity

We can think of behavior in the model as any action where individuals may be motivated to act partly based on what others expect of them – littering (Reno et al., 1993), voting (Gerber and Rogers, 2009), tax compliance (Wenzel, 2005). Therefore, the model can be thought of a stylized approach to societal norm emergence. I can then use Results 1 and 2 to describe the conditions under which we should expect different norm to form, and when pluralistic ignorance will be more likely.

Given the the frequentist way of thinking about the model (as discussed in subsection 4.3), a corollary of Result 1 is that the distribution of norms across groups will vary much more when there the context is very uninformative than when the context is very informative, or strongly $\phi \in \{0, 1\}$ as defined in section 3.1.1. Thus, we may expect that if there is a strong belief that most in society are racist, racist behaviors will be judged benevolently and be commonplace among groups of friends – racism is the norm in all groups. Similarly, racist behavior will be judged harshly and be uncommon when there is a strong belief most in the population are not racist – racism is inappropriate in all groups. However, when there is uncertainty about the population distribution of racism, then there will be much higher variation in the appropriateness of racism across groups.

If we think of groups in the model as 'societies' or 'tribes', Result 1 may explain why some customs are common across groups, while other customs vary wildly. Indeed, Brown (1991) has compiled a list of 'human universals' or features of society for which there is no known exception. On the other hand, the wide variation in customs across societies has also been documented (e.g. Pinker, 2003, Edgerton, 1992, Henrich et al., 2001). The model then offers an explanation of why human tendencies which are readily apparent (high $P(\theta)$) lead to emergence of similar norms, while weaker ones lead to high variation. It further predicts that inefficient norms are more common among those norms that vary more widely across societies. Although outside the scope of this paper, if we allowed groups to compete, evolutionary pressures may privilege some societies over others because of their norms (Fehr and Fischbacher, 2003).

4.4.3 Second Order Conformity in Interstate Conflict Bargaining

At the core of the celebrated 'security dilemma' (Jervis, 1978) is a concern for second order conformity: some states prefer to increase their defense capabilities only if they think the other wants to attack it. Consider a setting with two states. A 'non-expansionary' state wants to increase their defenses only if the other wants to attack it, while 'expansionary' states want to attack. This setup differs from that of section 2 in that expansionary states are assumed to not care about the other state's preference. However, the dynamics are similar given that expansionary states do care about doing what they think the other state wants. In an appendix the model is suitably modified to capture this situation, and here I provide a summary.

States are drawn either from a 'dangerous' population of mostly expansionary states, or from a 'peaceful' population of mostly non-expansionary states. In this setting, context is the commonly known probability that states are draws from a dangerous population. If states are very sure that they are drawn from a dangerous population, they will both increase their defense capabilities and end up going to war – analogous to how in Result 1 individuals pool on ϕ when the context is strongly ϕ. Thus, non-expansionary states may end up going to war even though neither wanted to in the first place – analogous to pluralistic ignorance as in Result 2. These types of wars, product of misunderstandings, are often called 'inadvertent wars' (Fearon, 1995, George, 1991).

These types of equilibrium misunderstandings are difficult to capture without second order conformity. Indeed, in his seminal treatment on the subject, Fearon (1995) noted that the theoretical logic of these misunderstandings had not been worked out.[7] In contrast, when there is uncertainty over the population from which states are drawn (an uninformative context), two non-expansionary states will avoid increasing their defenses to signal their type, and thus avoid a war. Therefore different norms can emerge to regulate state behavior, and states may 'gesture' towards each other with their defense decisions to establish certain norms – analogous to Result 1 in which the first movers reveal their type when the context is uninformative. These claims of norm emergence echo an argument made by 'constructivists' in international relations scholarship (e.g Wendt, 1992), who have sometimes sought to distinguish their reasoning from game-theoretic analysis (but see Fearon and Wendt, 2002). This observation underscores that second order conformity captures concepts traditionally discussed outside of game theory.

[7] From Fearon (1995): "Presumably because of the strongly zero-sum aspect of military engagements, a state that has superior knowledge of an adversary's war plans may do better in war and thus in prewar bargaining – hence, states rarely publicize war plans. While the theoretical logic has not been worked out, it seems plausible that states' incentives to conceal information about capabilities and strategy could help explain some disagreements about relative power."

4.4.4 Examples Outside of Politics

In this section I present some examples of social phenomena which can be fruitfully analyzed through the lens of second order conformity.

Dating, gift giving, drinking. Consider two strangers who may have a romantic interest in each other. Each may want to ask the other out, but only if the other wants to go on a date. That is, each wants the other to believe that their preference for going on a date matches the other's. They have uncertainty as to whether the other is most likely interested (that is, is drawn from a population where most want to go on a date) or most likely not (drawn from a population where most don't want to go on a date). The common knowledge over the probability they are most likely interested $P(\theta)$ captures the shared expectations of the situation – it is more likely that the other is interested at a singles bar than at a funeral. As a second example, groups of friends or family members may want to engage in gift-exchange depending on what they think others want, possibly resulting in several holidays going by where gifts are reluctantly exchanged (Waldfogel, 2009). To give a third example, teenagers may want to drink heavily if that's what they think their friends want (Prentice and Miller, 1993).

Norm heterogeneity in dating practices. In small groups, Result 2 states that there should be most pluralistic ignorance when there is high certainty over the population distribution of preferences. For example, when two people are going on a date, some of their behavior will have strong social expectations – the man in a heterosexual couple pays for dinner in the first date in many societies. However, there are much weaker social expectations regarding the cuisine of the restaurant they go to. The result therefore predicts that there will be more first dates where the man pays for the meal even if both parties preferred a more equitable split, but that the choice of restaurant will be closer to what at least one of the two wanted. Note that this result does not assume that the couple cares about social expectations when considering who pays the bill but does not care when considering the restaurant. Instead, only in the first instance can they use their priors to determine what is appropriate, so in the second instance one of them will find it optimal to reveal their type.

Consider a second implication to dating given a frequentist way of thinking about the model. I argued that the likelihood of a pair of individuals dating varied with the shared beliefs over the distribution of preferences in a population. This can help explain the change in household composition across time. The gender of partners living in a household in the U.S. in the 80's were a male-female pair in a larger proportion than in the twenty-first century, even before same-sex marriage was legalized. This has been explained in the public opinion literature by a change in the proportion of people who accept homosexual relationships (Baunach, 2012). As the beliefs over the proportion of homosexuals in society changed, so did the amount of homosexual dating.

Deferential Language. A corollary of Result 1 is that a higher weight on social expectations (higher β) makes it more likely that individuals' actions do not represent their preferences. Some evidence for this is provided by the finding that an individual is more deferential in their language towards someone with more authority (Lee and Pinker, 2010). Indeed, individuals are more interested in saying what they believe a person with higher authority wants to hear.

5 The Impact of Principals and Policies on Social Change

I have presented the basic model in sections 2 and 3, and argued for its wide applicability in section 4. In this section, I extend the model to allow for a range of principals or policies that may affect equilibrium behavior, or the 'norm that emerges' as defined in subsection 4.1. The overarching questions of this section are: how can a principal or a policy affect social change? When can they help a group avoid pluralistic ignorance, and when can they increase the probability that it arises?

In subsection 5.1, I consider an informed and trustworthy principal who makes an announcement with the objective of minimizing pluralistic ignorance. A classic example of this is a teacher who would like the class to engage in a lively discussion on a controversial topic, avoiding students from all taking what they think is the majority perspective. I show that principals cannot always achieve this without knowing the actual distribution of preferences in the group, so in subsection 5.2 I consider a social information campaign where a principal surveys preferences in the group and discloses the aggregate results to the group. Second order conformity concerns may lead those surveyed to misstate their preferences, so a successful social information campaign has to extract truthful answers and have the group know that it has done so. Subsection 5.3 considers 'social meaning regulators', principals or policies that affect what a group considers appropriate behaviors ('appropriateness' and 'social meaning' are defined in section 4.1). Social meaning regulation has received attention in legal scholarship (Lessig, 1995) and in non-game theoretic political science (Finnemore, 1996), but to my knowledge has not received attention by game theoretic scholars. Subsection 5.4 shows how an uninformed, obscure and politically inactive 'everyman' can start a wave of protests, and argue that this logic can help us understand the beginning of the Arab Spring. In subsection 5.5 I argue that the weight on social expectations can sometimes be directly manipulated in order to impact behavior, and I offer an unusual policy by the mayor of Botogá's of putting mimes on the street as an example. Although the first five subsections can be read somewhat independently, the discussion sections discuss the relationships between the principals and policies. Subsection 5.6 provides some closing comments for the section.

In all of the examples, I will consider the timeline depicted in Figure 2. The game is as in the basic model, except for a new period 'P' which occurs after nature has made its moves but before group members act. A principal or policy $\mathcal{P}$ takes an action that I will specify per subsection, and I will study its impact on behavior.

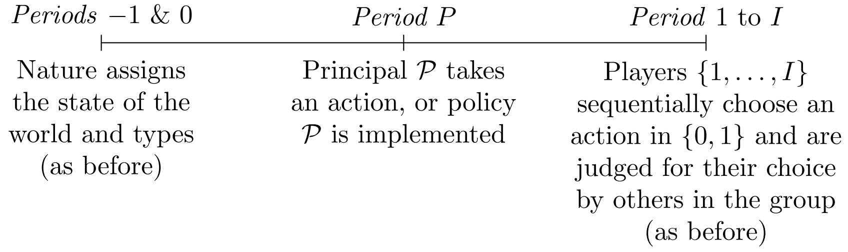

Periods -1 & 0
Period P
Period 1 to I
Nature assigns the state of the world and types (as before)
Principal $\mathcal{P}$ takes an action, or policy $\mathcal{P}$ is implemented
Players $\{1, \ldots, I\}$ sequentially choose an action in $\{0, 1\}$ and are judged for their choice by others in the group (as before)

Figure 2: Timeline with principal or policy $\mathcal{P}$

All of the subsections on principals and policies will be organized in the same way. Each will open by motivating the type of principal or policy $\mathcal{P}$ I am considering, set up the actions available to $\mathcal{P}$, show the result and then discuss it. Throughout this section, I will draw freely from a wide variety of motivating examples, taking advantage of the broad applicability of the model discussed in Section 4. Readers more interested in the theoretical arguments may wish to focus less attention on the discussion sections.

5.1 Informed principal

Here I consider a principal who is informed of the population preferences and trustworthy, and has the opportunity to make a statement before group members make their decision.

5.1.1 Motivation

Harvard College holds a non-credit seminar in which a group of 12 students get together to discuss how to live wisely (Light, 2015). A theme the seminar organizers ask the students to explore in an open discussion is where they stand on the trade-off between leading a relaxed or a hard-working life. The seminar organizers would like to achieve a lively discussion, and would therefore like to avoid students from hiding their opinions by saying what they think others think is the standard opinion. How should the seminar organizers address the group before starting the discussion? This question is relevant beyond the example of the Harvard seminar. Getting individuals in a group dynamic to express their views may not only help them reach better decisions that affect themselves, it may also result in more creative or better informed solutions to a problem the group is working on. Relatedly, lawmakers worry about whether laws may make individuals behave in a way they think others want, diminishing the diversity of behavior in society Sunstein (1999).

5.1.2 Setup

Suppose that $\mathcal{P}$ is a principal who knows more about the populations than others, and has incentives to reveal her private information truthfully – such as the seminar organizers in the Harvard course. For concreteness and simplicity, suppose it is common knowledge that $\mathcal{P}$ observes θ perfectly, and would like to minimize pluralistic ignorance.[8] Note, importantly, that $\mathcal{P}$ observes the population distribution of preferences, but not the group distribution of preferences. Perhaps the principal has private information about groups on average, but not about a specific group it is addressing. In the Harvard College example, the seminar organizers have observed past seminar students, but do not know the preferences of the current group.

Principal $\mathcal{P}$ observes the state of the world, and chooses an action $a_{\mathcal{P}}$. Decision maker i uses the information given by $\mathcal{P}$'s action to learn about how others in the group will judge her. Then if $\mathcal{P}$ reveals the state of the world through her action ($a^*_{\mathcal{P}} = \theta$), individuals will know the population distribution from observing $\mathcal{P}$'s actions.

I would like to consider the strategy that $\mathcal{P}$ would follow if she did not face a commitment problem. I'll therefore say $\mathcal{P}$ *can commit* if she can announce a strategy and not deviate from it.

5.1.3 Result

Result 4. *Suppose a principal $\mathcal{P}$ observes the population θ and would like to minimize pluralistic ignorance by making an announcement $a_{\mathcal{P}} \in \{0, 1\}$. Her optimal strategy depends on the size of the group. Consider the thresholds $\tau_1(G)$ and $\tau_2(G)$ from Result 2.*

- *Suppose the group size is small, or $I \leq \tau_1(G)$, and the context is strongly $\phi \in \{0, 1\}$. Then if $\mathcal{P}$ can commit, she chooses a mixed strategy such that with one message $a \in \{0, 1\}$ group members believe the context is uninformative, and with the other message $1 - a$ group members believe the context is strongly ϕ. If $\mathcal{P}$ cannot commit, no message sent by $\mathcal{P}$ leads group members to believe the context is uninformative.*

- *Suppose the group size is small, or $I \leq \tau_1(G)$, and the context is uninformative. Then $\mathcal{P}$ does not reveal information about the population with his action.*

[8] Although natural alternatives would be for $\mathcal{P}$ to maximize $\sum_{i \in I} -(a_i - \phi_i)^2$ or the sum of utilities, these specifications lead to technical complications that are unnecessary to make the main points.

- *Suppose the group size is large, $I \geq \tau_2(G)$. Then $\mathcal{P}$ reveals the population with her action: $a^*_{\mathcal{P}} = \theta$.*

The probability of pluralistic ignorance is weakly reduced by $\mathcal{P}$, but is positive for any finite I.

Result 4 is mostly a corollary of Result 2, although it is worth giving intuition for why the principal has a commitment problem when the context is strongly ϕ. If the principal wishes to increase uncertainty about the population in this context, her message must be informative about the population. Therefore, if one message 'a' makes it more likely that the context is uninformative, the other message '$1 - a$' must make it more likely the context is strongly ϕ. But since the principal always does best off when the context is uninformative, she would want to always choose a. Harvard undergrads would 'see through' an attempt by the seminar organizers if they thought there was an obvious answer to the question of how they were supposed to want to live their life.

5.1.4 Discussion

Result 4 underscores three points. First, when group size is small, a principal who is tying to minimize pluralistic ignorance will want to increase uncertainty about the population distribution of preferences. For example, seminar organizers in the Harvard course on living wisely present the trade-off between leading a relaxed or a hard-working life as an open-ended question which has several valid answers. If instead the trade-off were presented as one where there is an answer most obviously agree with, seminar organizers would lead all to agree with that answer despite their private views. Relatedly, Sunstein (1999) discusses that legislators face a trade-off when deciding whether to write a law that establishes some action as the society's norm. In doing so, he argues, he may make it harder for 'norm communities' to flourish – different groups with different norms.

Second, when the group size is large, the principal will want to increase certainty about the population distribution. Indeed, with large groups, the population distribution of preferences approximates the group distribution of preferences. Third, perfect information about the population distribution of preferences may nonetheless lead to pluralistic ignorance. Indeed, even when the group is 'large', the group distribution may differ from the population distribution. We have studies from several communities in the U.S. that teenagers overestimate the amount of alcohol that other teenagers consider ideal to drink at a party (Kenny et al., 2011). This misperception is thought to lead to over-drinking. We can think of this information as the population level information over the distribution of preferences for drinking across communities. A principal can publicize this information to affect the norms of a community that has not been part of a study on over-drinking. However, the principal may not want to do so since over-drinking may not be a problem with the teenagers in that community.[9] A different rational for not publicizing information from other communities is that individuals may not be responsive to information about groups different than their own (e.g. Goldstein et al., 2008).

The challenges faced by an informed trustworthy principal may be assuaged by a 'social information campaign', in which group members' preferences are recorded and then publicized. To this I now turn.

5.2 Social Information Campaign

I now consider a "social information campaign', in which a principal surveys group members, aggregates the preferences and reports the aggregate results back to the group.

5.2.1 Motivation

Social information campaigns have been growing in popularity as a policy tool since Cialdini et al. (1990), although it has had mixed success (Kenny et al., 2011). Part of the challenge may be in getting respondents to answer truthfully instead of saying what they think the surveyor wants to hear, the so-called 'surveyor demand effect'. Some successful social information campaigns have been able to avoid subjects interacting with a surveyor, such as with energy consumption (Allcott, 2011), charitable donations (Frey and Meier, 2004), or use of linens in hotels (Goldstein et al., 2008), but see Beshears et al. (2015) for an unsuccessful campaign where behavior was measured directly. Other behavior is much harder to measure without surveys, such as alcohol consumption (Perkins et al., 2010) or political attitudes (Van Boven, 2000). In these cases, principals must make sure that survey respondents respond truthfully, and as importantly that they are perceived by the intended audience to do so.

5.2.2 Setup

To capture a social information campaign, I modify the model so that at period P, a principal surveys each player privately about their ideal point, and then reports the average responses $\mathcal{A} \in [0,1]$ to the group. I further modify the model so that the principal has a preference

[9] This is a stylized discussion of this issue, as in fact studies typically measure the perceived and actual amount of desired drinking in a community. In the model, agents have a binary choice, which would map on to this situation as 'over-drinking' and 'not over-drinking'. If we take the more continuous nature of the problem into account, the concern is that information about desired amount of drinking from other communities may over or under-estimate the desired amount of drinking in the community.

of her own, $\phi_{\mathcal{P}} \in \{0,1\}$, and it is common knowledge that her preference is $\phi_{\mathcal{P}} = 1$ with probability $P(\phi_{\mathcal{P}} = 1)$. Thus, period P is subdivided into three. First, $\mathcal{P}$'s type is drawn by Nature. Second, $\mathcal{P}$ surveys the group. Third, $\mathcal{P}$ reports the average responsed $\mathcal{A}$ to the group.

When the principal surveys a player in period P, the player's response $\tilde{a}_i \in \{0,1\}$ maximizes a utility function that weighs responding according to her ideal point with responding what she think the principal wants to hear.

$$-(\phi_i - \tilde{a}_i)^2 + \tilde{\beta}\mathbb{E}(\mathcal{J}_{\mathcal{P},i} \mid \hat{\phi}_{\mathcal{P}}) \tag{3}$$

That is, i's utility at the survey response stage has the same terms as before, but now principal is the only judge of i's action. Note that I am assuming that, when responding the survey, players completely discount their utility from the sequential decision stage.

5.2.3 Result

Result 5. *Suppose the principal $\mathcal{P}$ has preference $\phi_{\mathcal{P}} \in \{0,1\}$ determined by Nature with commonly known probability. $\mathcal{P}$ surveys group members about their preferences, and reports the average response $\mathcal{A}$ back to the group at period P. Group members' survey response $\tilde{a}_i$ takes into account how $\mathcal{P}$ will judge them, as represented by* (3). *Then*

- *If group members have strong beliefs about $\mathcal{P}$'s type and place a lot of weight on how the principal will judge them ($|P(\phi_{\mathcal{P}} = 1) - .5|$ and $\tilde{\beta}$ large enough), group members will not update their beliefs with the social information campaign: $P(\phi_k \mid h_i, \phi_j, \mathcal{A}) = P(\phi_k \mid h_i, \phi_j)$ for any k, j, i. The probability of pluralistic ignorance is unaffected.*

- *If group members are uncertain about $\mathcal{P}$'s type or place little weight on how the principal will judge them ($\tilde{\beta}$ or $|P(\phi_{\mathcal{P}} = 1) - .5|$ low enough), there is no pluralistic ignorance after period P.*

If i is certain about the principal's preferences and she places a lot of weight on how the principal will judge her, then i will answer what she thinks the principal wants to hear. This captures the surveyor demand effect. If furthermore group members believe surveyor demand effects were widespread, then the social information campaign will be ineffective. However, if either there is uncertainty about the principal's preferences or individuals place little weight on how they'll be judged, they will respond truthfully. The average reported by the principal will be accurate, and if group members know this, they will pool on the correct majority view.

5.2.4 Discussion

In order for a social information campaign to be successful, it must be the case that principals reveal their preferences truthfully to the principal, and that group members believe that preferences were truthfully revealed. Survey designers who are trying to avoid survey participants from hiding their opinions, often preface sensitive questions by saying that there is a 'diversity of opinions' on a certain topic. This is an attempt at increasing the survey respondent's belief that the population is not too biased towards any particular opinion, and thus avoiding the surveyor from answering what she thinks the interviewer wants to hear. Surveyors are also trained in not appearing judgmental in their reaction to survey responses, presumably to lower survey respondent's weight on being judged ($\tilde{\beta}$). Surveyor demand effects may explain the mixed success of social information campaigns (Kenny et al., 2011), and in particular why social information campaigns which measure behavior objectively seem to be more effective.

There are two further challenges to social information campaigns that use a surveyor, which I will only sketch out. The first is that social information campaigns may suffer from a selection problem. If the objective of a social information campaign is to change norms, then only principals who are not happy with the current norm will deploy a social information campaign. But then group members will infer the principal's preference ($\phi_{\mathcal{P}}$) from the existence of the campaign, making the campaign uninformative.

The second concern is that the results of the campaign itself may be informative about the principal's preference. Suppose group members are already in pluralistic ignorance when the principal deploys a social information campaign. For example, it may be that group members act sequentially in a first stage, which leads them to pluralistic ignorance, then the principal deploys the social information campaign, then they act sequentially again.[10] Further suppose that group members have incomplete information about the impact of the surveyor demand effect on survey response. If the social information campaign reveals that the distribution of preferences is very different from what people believed, they may explain this by a large surveyor demand effect, and therefore not be swayed by the results.

[10] To avoid incentives to influence, I can further assume that players are myopic and only care about the present period. Notice that in this case, the analysis of multiple rounds of sequential play is straightforward. In the first round, group members act like if there was only one round. In further rounds, those who have already separated have no incentives to follow social expectation since their types are known, so they choose their ideal point. Those who pooled have the same information as player I at the end of the first round, so continue to withhold.

5.3 Regulating Social Meaning

I now turn to a type of 'norm entrepreneurship' (Finnemore, 1996) which has not received much attention in formal models: that of an individual or policy who changes a behavior's social meaning (as defined in Section 4.1).

5.3.1 Motivation

A principal or a policy $\mathcal{P}$ may change behavior by adding to its social meaning. For example, when Nancy Reagan asked teenagers to 'Just Say No To Drugs', teenagers may have reacted by believing that since nobody in their group wants to do what Nancy Reagan says, doing drugs signals that their preferences match those of the group (Sunstein, 1999). Gandhi led Indians to believe that non-violent protest was a signal of being a true Indian. Other social movements, such as Otpor! in Serbia, followed this strategy (Popovic and Miller, 2015).

Policies may achieve a similar purpose. There was a norm for dueling among the elite in the U.S., whose social inefficiency has been argued by Lessig (1995), Schwartz et al. (1984). As Lessig (1995) argues, many attempts at banning it were ineffective. One policy that was particularly effective was prohibiting those who duel from holding public office. The reasoning is that, unlike other policies, the prohibition pitted two norms of the elite against each other. A member of the elite could get out of a duel by claiming that the elite's responsibility to hold public office was higher than the responsibility to duel. Thus, by linking the act of dueling to a second norm, the elite was able to reject a duel and avoid signaling that their preferences did not match those of the group.

5.3.2 Setup

To capture this formally, assume that individuals have a pair of preferences (ϕ_i, ψ_i). The first preference is, as before, the ideal point over a behavior $a \in \{0, 1\}$ that players will choose sequentially. The second preference is over some seemingly unrelated issue. For example, $\psi_i \in \{0, 1\}$ can be i's views on Nancy Reagan, Gandhi or serving in public office, with $\psi_i = 1$ indicating a favorable view. ψ_i is uninformative about the state of the world θ.

Before group members make their sequential decisions (to respectively do drugs, protest non-violently, or propose a duel), $\mathcal{P}$ can establish a 'link' to behavior a. This could be Nancy Reagan's prescription to 'Say No To Drugs', Gandhi's prescription for non-violent resistance, or the law banning duelers from holding public office. I will model this in a simple way.

$\mathcal{P}$ may establish a link by truthfully reporting her approval for action a – an endorsement in the case of a principal like Nancy Reagan, a law in the case of a policy that states that only those who do not duel can run for office. I write $\psi_{\mathcal{P}} = a$ if $\mathcal{P}$ 'approves' action a.[11] If $\mathcal{P}$ establishes a link to the action a she approves of, she sets $l = 1$, and $l = 0$ otherwise. $\mathcal{P}$ is interested in maximizing the amount of group members who choose the action she approves of, $\psi_{\mathcal{P}}$.

Group members' utility function is a sum of two components. The first component is the original utility function. The second component is operative only if $\mathcal{P}$ establishes a link, and captures supporters' motivation to act according to what $\mathcal{P}$ favors, as well as group members' motivation to do what they think others think about $\mathcal{P}$. Expected utility is then:

$$-(\phi_i - a_i)^2 + \beta \mathbb{E}_{\hat{\phi}_{-i}}(\mathbb{E}(\mathcal{J}_{j,i} \mid \hat{\phi}_{-i})) + l\delta\big[-\psi_i(\psi_{\mathcal{P}} - a_i)^2 + \beta \mathbb{E}_{\hat{\psi}_{-i}}(\mathbb{E}(\mathcal{K}_{j,i} \mid \hat{\psi}_{-i}))\big] \qquad (4)$$

$\mathbb{E}_{\hat{\psi}_{-i}}$ is the expectation over $\hat{\psi}_{-i}$, the possible distribution of views regarding $\mathcal{P}$. $\mathcal{K}_{j,i} \in [0,1]$ is a judgment function equal to one if j believes i's preference over $\mathcal{P}$, ψ_i, matches the group without i's majority preference, ψ^m_{-i}:

$$\mathcal{K}_{j,i} \begin{cases} = 1 & \text{if} \quad P(\psi_i = x \mid h_i, a_i, \phi_j) > 1/2 \ \& \ P(\psi^m_{-i} = x \mid h_i, a_i, \phi_j) > 1/2, \ x \in \{0,1\} \\ = 0 & \text{if} \quad P(\psi_i = x \mid h_i, a_i, \phi_j) > 1/2 \ \& \ P(\psi^m_{-i} = x \mid h_i, a_i, \phi_j) < 1/2, \ x \in \{0,1\} \\ \in [0,1] & \text{if} \quad P(\psi_i = x \mid h_i, a_i, \phi_j) = 1/2 \text{ or } P(\psi^m_{-i} = x \mid h_i, a_i, \phi_j) = 1/2, \ x \in \{0,1\} \end{cases}$$

If $\mathcal{P}$ establishes a link, player i of type $\psi_i = 1$ has a bias towards choosing the action $\mathcal{P}$ prefers (the $-(\psi_{\mathcal{P}} - a_i)$ term), and all players want to signal whether their preferences with respect to $\mathcal{P}$ match the majority in the group (the $\mathbb{E}(\mathcal{K}_{j,i} \mid \hat{\psi}_{-i})$ term).

In the result below, I will consider the probability of pluralistic ignorance over the vector of ideal points $\overline{\phi}$.

5.3.3 Result

Result 6. *Suppose $\mathcal{P}$ decides whether to establish a link $l \in \{0,1\}$ between a behavior and group members' views regarding $\mathcal{P}$. If $\mathcal{P}$ does establish a link ($l = 1$), group members who favor $\mathcal{P}$ will be concerned about acting according to her preference over the behavior ($\psi_{\mathcal{P}}$) and group members will want others to think their preference over $\mathcal{P}$ match those of the group majority, as captured in* (4)*. Then:*

- *Suppose most in the group favor $\mathcal{P}$, or $\sum_{j \in -i} \psi_j/(I-1) > .5$, and group members place a lot of weight on the link established by $\mathcal{P}$ (δ large enough). Then $\mathcal{P}$ establishes a link and all players pool on $\psi_{\mathcal{P}}$. The probability of pluralistic ignorance diminishes*

[11] The verb 'approve' is abused in the example of a law proscribing public office for duelers, where the logic is that the law establishes that holding public office is inconsistent with dueling. The law is not an agent, but for simplicity of exposition I will treat it as such.

if $\mathcal{P}$ favors the majority group preference ($\psi_{\mathcal{P}} = \phi^m_{-i}$ for all i) and the context is uninformative. The probability of pluralistic ignorance increases if $\mathcal{P}$ favors the minority group preference ($\psi_{\mathcal{P}} \neq \phi^m_{-i}$ for all i).

- *Suppose most in the group do not favor $\mathcal{P}$, or $\sum_{j \in -i} \psi_j/(I-1) < .5$, and group members place a lot of weight on the link established by $\mathcal{P}$ (δ large enough). Then $\mathcal{P}$ does not establish a link. If $\mathcal{P}$ did establish a link, all players would pool on $1 - \psi_{\mathcal{P}}$.*

Suppose the context is uninformative. Without $\mathcal{P}$, the first movers in a group would choose their ideal point and either action could be established as a norm (by Result 1). If most in the group favor $\mathcal{P}$ and she establishes a link, then group members will believe there are added social reasons to follow $\psi_{\mathcal{P}}$ – either because they want to do what $\mathcal{P}$ considers appropriate or because they want to be considered to be part of the group majority who cares about $\mathcal{P}$.[12]

Suppose instead that context is strongly ϕ. Then without $\mathcal{P}$, individuals pool on ϕ and judges would believe that preferences do not match for those who choose $1 - \phi$ (by Result 1). If most in the group favor $\mathcal{P}$ and she establishes a link to $\psi_{\mathcal{P}} = 1 - \phi$, then it is no longer the case that all group members choose ϕ: those who place weight on doing what $\mathcal{P}$ thinks is appropriate ($\psi_i = 1$) will choose $\psi_{\mathcal{P}}$ for δ large enough. The social meaning of choosing $\psi_{\mathcal{P}} = 1 - \phi$ is thus 'ambiguated' (Lessig, 1995), since now some types whose preferences match the group would choose $1 - \phi$, which makes it less socially costly for those with $\psi_i = 0$ to choose $\psi_{\mathcal{P}}$.

5.3.4 Discussion

By making an action a signal of whether a player values $\mathcal{P}$, the 'social meaning' of the action changes (see subsection 4.1 for a discussion of social meaning). Group members may come to believe that some will do what $\mathcal{P}$ prefers, affecting what the behavior signals. $\mathcal{P}$ will therefore be able to impact behavior not by informing the group about the group's preferences, but by informing the group about what $\mathcal{P}$ approves of. However, enough group members must put a sufficiently large amount of weight on acting according to what $\mathcal{P}$ approves of. Thus, many laws that were implemented with the objective to ban dueling failed in making the elite give enough weight to them so as to ignore their social responsibility to duel. If an elite was not convinced that someone who 'truly identified' as an elite would avoid dueling to follow a certain law, not dueling continued to signal not being an elite.

[12] This is what Lessig (1995) refers to as 'tying'. "In these cases, the social meaning architect attempts to transform the social meaning of one act by tying it to, or associating it with, another social meaning that conforms to the meaning that the architect wishes the managed act to have."

Finnemore (1996) provides several examples of norm entrepreneurs whose main role in bringing about change was, she argues, to redefine what states 'should' do in a normative sense. In his push to establish the Red Cross, Henry Dunant framed his appeal about humanitarian norms in war not in terms of 'interests and advantage', but 'in terms of the responsibilities of Christian gentlemen and civilized nations.' If Dunant is a supported member of 'Christian gentlemen and civilized nations', then when group members learn about his preferences about how other should act, they are learning about the behavior approved by Dunant and therefore by the group. Gandhi's concept of non-violent protest as a way to define what it meant to be Indian is a second example.[13]

Contrast this logic to that of a trustworthy principal with private information such as in section 5.1, and a common assumption in the theoretic literature (Canes-Wrone et al., 2001, Maskin and Tirole, 2004, Lupia and McCubbins, 1998, Hermalin, 1998, e.g.). If the principal had private information unrelated to her preferences, having strong preferences about how others should act would make her statements uninformative. Put another way, if the principal influences behavior due to private information then her statement will be uninformative about her preference, but if the principal influences behavior by regulating social meaning her statements will be highly informative of her preferences.

Social meaning regulation ties in to several literatures. Social psychologists and sociologists have developed theories of leadership based on the idea that a leader who establishes large support among a group can redefine its norms (e.g. Hogg, 2001, Burns, 1978, Hollander, 1958). Establishing support, referred to by these authors as 'intragroup prototypicality', 'transformational leadership' or 'idiosyncracy credit', requires establishing oneself as being highly valued by the group. There is a growing body of work on the expressive function of the law. Benabou and Tirole (2012) shows how the law can provide information about how *intensely* a population values an exogenously given appropriate behavior. In my setup, in contrast, the law is providing information about whether littering is considered good or bad. As Sunstein (1999) argues, 'perhaps most people are happy that littering is not stigmatized.' McAdams (2015) and Acemoglu and Jackson (2017) consider the coordinating role of the law in a setting where individuals have strategic complementarities (a behavioral motivation discussed in section 4.2). In these models, what drives the coordinating power of the law is that it provides a public signal that lets individuals know what others do. In contrast, the what's essential for a social meaning regulator to successfully affect behavior is whether the group approves of her. A related literature considers how a leader can coordinate followers

[13] This example is perhaps more controversial, as it can be argued that Gandhi had private information about the action that would bring down the British colonial government. That is, Gandhi believed that if Indians did not cooperate with the British colony, it would not be able to govern. This alternative explanation is incomplete, however, since it does not address individuals' incentives to free-ride.

through sending a public signal (Angeletos et al., 2006, Acemoglu and Jackson, 2014).

5.4 Uninformed Catalysts

In this section I consider how an uninformed, obscure and politically inactive 'everyman' can start a wave of protests, and in so doing provide a novel logic of abrupt social change.

5.4.1 Motivation

Mohamed Bouazizi was 28 years old on December 17th, 2010. He sold fruit from his stand, was educated in a one room country school and was not politically engaged. According to his aunt, his main aspiration was to buy a truck to improve his fruit sales. In the morning, a police officer confiscated his cart under dubious allegations. Upset, Mohamed went to complain to the municipal government, but he was ignored. In response, he reportedly told the officials that if they didn't see him, he would burn himself. Shortly later, he doused himself with gasoline and set himself on fire.

On December 18th, a small group of protesters gathered. This was recorded on a phone and uploaded to Facebook. In contrast to other protests which had been silenced by the Tunisian regime, this one was widely spread partly due to the sudden increase in the popularity of the social networking site. Less than month later President Zine El Abidine Ben Ali resigned. Less than two years later 5 rulers were forced out of power, while other protests spread across the region.[14]

In order to explain Bouazizi's impact, I will draw upon one of the oldest and best known models of social movements: Hans Christian Andersen's 1837 fairytale 'The Emperor's New Clothes'. Although make-believe, past scholars have considered the story a useful starting point for thinking about social movements (e.g. Bicchieri, 2005, Centola et al., 2005), and have interpreted the story in a way I will contest. As with any other model, its ultimate usefulness is in shedding light on empirical phenomena, and my aim in challenging the typical interpretation is to shed better understand the political impact of an everyman's protest.

In the fairytale, an emperor is tricked by thieves into thinking that they sold him a magical robe that can only be seen by those who are deserving of their rank. In fact, he is not sold anything. The emperor goes out to the plaza, where no one wants to admit that they see a naked emperor out of fear of revealing their undeservingness. The tale classically ends with a poor child exclaiming 'The emperor has no clothes!', and all citizens laughing.

The typical interpretation is that the child's public statement creates common knowledge that allows everyone to laugh at the emperor. Formally, this is typically modeled as a coor-

[14] A good review of the Arab Uprisings is given by (Gelvin, 2015).

dination game in which citizens use the child's public signal to coordinate on an equilibrium. However, this interpretation leaves out a crucial component of the story: the child was special. Because he was a child, people knew he was acting *brashly* - without consideration of the social consequences of his actions. Because he was a *poor child*, people also knew that he was at the lowest social rank. There was no doubt that what he saw was not due to his not deserving of his rank. Everyone knew he had perfect private information about the state of the world. Furthermore, given his brashness and deservingness, whether or not he knew he had these traits was irrelevant for his message to affect others' beliefs about the state of the world. In fact, a very reasonable interpretation of that story is that he was an innocent child, who did not understand the consequences of his brashness. This means that the child was an *uninformed catalyst*.

Mohammed Bouazizi was also an uninformed catalyst. He had a brash reaction to an extremely upsetting event. He had not strategically planned for it, and, given his obscurity, had little realistic expectations that it would produce a small wave of protests, let alone the Arab Spring. Furthermore, he was very much an everyman that citizens could relate to. This is suggested by his working class background, his devotion to his family and his altruism towards children. In statements about Mohammed, a common interpretation is that his reaction was a reflection of what Tunisians 'really' felt. In a Times article written shortly after Mohammed's death, a young man in the city where Mohammed self immolated said 'We are all Bouazizis if our hopes are dashed.' A neighbor of Mohammed said 'We were silent before but Mohammed showed us that we must react.' (Abouzeid, 2011)

5.4.2 Setup

We then reinterpret the model. Individuals now are deciding whether to protest. An individual's decision to protest only depends on trading off her views on the regime with protesting only if she thinks others think the regime is bad. This setup allows us to not have to modify the model, although it abstracts from the common assumption that the number of individuals that protest impacts the payoff from protesting (e.g. Kuran, 1997). This abstraction helps us focus on the novel results from the second order conformity approach.

To model a principal who is an uninformed catalyst, I introduce the concept of 'benignity'. To illustrate what I mean by benignity, consider the left hand matrix of Table 1. A benign person is someone whose preference match those of the majority in the group.[15] Thus, if the regime is 'good' – that is, most citizens support it – then a benign principal favors the regime ($\phi_{\mathcal{P}} = 1$). Formally, the regime is 'good' if $\phi_{-i}^{m} = 1$ for all i. Otherwise, the benign

[15] I could have alternatively defined a benign person as someone whose preference matches the population. This presentation is for succinctness, and I will not draw any conclusion from this decision.

	Good regime ($\phi^m_{-i} = 1$)	Bad regime ($\phi^m_{-i} = 0$)
Benign	Favors regime ($\phi_{\mathcal{P}} = 1$)	Disfavors regime ($\phi_{\mathcal{P}} = 0$)
Not benign	Favors regime with prob $P(\phi \mid \theta)$	Favors regime with prob $P(\phi \mid \theta)$

	Has clothes ($\phi^m_{-i} = 1$)	Has no clothes ($\phi^m_{-i} = 0$)
Deserving	Sees clothes ($\phi_{\mathcal{P}} = 1$)	Sees no clothes ($\phi_{\mathcal{P}} = 0$)
Undeserving	Sees no clothes ($\phi_{\mathcal{P}} = 1$)	Sees no clothes ($\phi_{\mathcal{P}} = 1$)

Table 1: Majority group preference and catalyst characteristic in my model (left) and in The Emperor's New Clothes (right)

principal disfavors the regime. We assume that the principal is drawn from a population with an arbitrarily small percentage of benign individuals. The preferences of principals who are not benign is distributed in the same way as before: a population θ is drawn with a commonly known distribution, and individuals' preference are randomly drawn from this population.

Note the similarity to the Emperor's New Clothes, represented on the righthand side matrix in Table 1. Deservingness in the fairytale is analogous to benignity in the model, while the emperor having clothes is analogous to the regime being socially good. In the fairytale, only those who are deserving when the emperor has clothes see the emperor with clothes.

A second characteristic I introduce is brashness. If the principal is brash, then she will act according to her ideal point. $\mathcal{P}$ has utility function

$$-(\phi_{\mathcal{P}} - a_{\mathcal{P}})^2 + \beta \mathbb{E}(\mathcal{J}_{j,\mathcal{P}} \mid \hat{\phi}_{-i})$$

and is brash if $\beta \in [0, 1)$. (If not brash, $\beta \in (1, 2)$.) I interpret the child as too innocent about social norms, too unaware about the alleged magical properties of the emperor's robes, or too surprised by the emperor's nakedness to not shout out that he was naked. Mohammed Bouazizi was too upset by how he was treated to shrug off the police abuse.

The third characteristic I introduce is uninformedness. A principal is uninformed if she has no private information about benignity. Neither the child in the fairytale nor Bouazizi had any reasonable expectations about the reaction their action would provoke.

5.4.3 Result

Result 7. *Suppose the principal is uninformed, and makes a decision $a_{\mathcal{P}}$ in period P. Subjects will all pool on $\phi_{\mathcal{P}}$ with certainty for any $P(\theta)$ and there will be no pluralistic ignorance if and only if group members know that the principal is benign and brash.*

If $\mathcal{P}$ is not brash, she will not reveal her type. If she is not benign, individuals will not be able to learn about the population with certainty from $\mathcal{P}$'s actions. If $\mathcal{P}$ is benign and brash, she reveals her type, which informs individuals about the group's majority preference with certainty.

5.4.4 Discussion

Past models of 'behavioral cascades' have been used to explain social change, including the Arab Spring and the quick fall of the communist regimes in Eastern Europe (Lohmann, 1993, Kuran, 1997). They are appealing because they fit the basic facts of an initial demonstration which spread quickly. Like my model, they have two components: the conditions or context that leads people to be acting suboptimally and the characteristics a first mover must possess to start a cascade. Although many of the observational implications are identical, all make a stark prediction about the nature of the first mover that does not characterize an uninformed catalyst. Indeed, the models predict that the first mover is knowledgeable and trustworthy (Canes-Wrone et al., 2001, Cukierman and Tommasi, 1998, Maskin and Tirole, 2004, Majumdar and Mukand, 2008, Bicchieri, 2005, Hermalin, 1998, Lupia and McCubbins, 1998), has extreme preferences (Granovetter, 1978, Kuran, 1997), or is a prominent cooperator who is showing others how to act (Acemoglu and Jackson, 2014). My model explains how the first mover can be uninformed, obscure and politically inactive. Lohmann (1994) presents the closest argument. She assumes that individuals take turns jointly deciding whether to protest, and it is the cumulative private signals of many moderates' discontent with the regime that encourages others to act. What is missing from her account is an explanation of why sometimes an individual seen as an 'everyman' or 'everywoman' can serve as a powerful motivator of many people.

It is worth contrasting an uninformed catalyst from a social meaning regulator. The uninformed catalyst's behavior provides knowledge about a group's true distribution of preferences because her brash actions are known to reflect what everyone wishes they were doing. In contrast, a social meaning regulator is able to define what everyone should do. An important implication of this difference is that the impact of the uninformed catalyst will be that the group avoids pluralistic ignorance, while the social meaning regulator may lead all to pool on $\phi_{\mathcal{P}}$ whatever the distribution of ϕ_i. This implies that only social meaning regulators will be able to change behavior when the context is strongly ϕ and the majority preference in the group is ϕ. We can further illustrate this distinction by considering other examples of uninformed catalysts.

Rosa Parks dropped out of secondary school and worked as a secretary, but her refusal to give her seat to a white person in a segregated bus catalyzed the Civil Rights movement (Theoharis, 2015). Unlike Bouazizi, she was politically engaged before her protest. However, when Civil Rights leaders such as Martin Luther King, Jr. decided to use Rosa Parks' case to call a bus boycott and take her case to court, the reason was that Rosa Parks was seen as an upstanding, humble Christian (Taylor, 2015). Others before Parks had refused to give up their seat – such as Claudette Colvin or Pauli Murray – but did not receive the backing of the NAACP because they were not perceived to be as benign as Rosa Parks (Branch, 1988). Uninformed catalysts, therefore, may be seized by politically savvy entrepreneurs who recognize the impact of a brash action taken by a benign individual. However, note that if the political entrepreneur influenced a benign principal's actions, the catalyst's actions would no longer be brash – they would instead reflect the political entrepreneur's preference. Therefore, the model predicts that if group members perceive a non-benign political entrepreneur's influence in an everyman's action, the impact of an otherwise uninformed catalyst will diminish.

To give another example, the Stonewall riots were a spontaneous demonstration by members of the LGBT community (Duberman, 2013). The demonstrations began after the New York police raided the Stonewall Inn, a bar that served as one of the few meeting places where LGBT members could openly express their sexual preferences. Speaking to the brashness of the action, one participant said that 'We all had a collective feeling like we'd have enough of this kind of shit.' Speaking to the fact that they were benign, the bar served the poorest and most marginalized people in the gay community. Before the Stonewall riots, gay protests portrayed the message that homosexuals were as 'normal' as heterosexuals but had different sexual preferences. The riots catalyzed the gay pride movement, in which homosexual protesters became open about their gayness as a distinct lifestyle. As with Rosa Parks, gay activists seized on the Stonewall riots to mobilize their support. A similar riot that had broken out earlier, at a cafeteria in Compton, San Francisco in 1966, did not attract the same sort of following. As opposed to the Stonewall riots, the Compton riot was seen as an act of anti-transgender discrimination, a group that was seen at the time as having fringe preferences with respect to the gay community. Indeed, the gay movement catalyzed by the Stonewall riots initially distanced themselves from transgender people.

5.5 Increasing Weight on Social Expectations

A group or society may know what the appropriate behavior is, but not be motivated to act appropriately. Here I briefly consider policies to increase concerns for social expectation.

Relax the assumption that individuals put a lot of weight on social expectations ($\beta \in (1, 2]$), and allow them to place relatively more weight on following their private views: now $\beta \in [0, 2]$. A policymaker may want to change the weight put on social expectations if, for example, it is leading individuals to act in a way that is not civic.

It is a simple corollary of Result 1 that in a context that is strongly ϕ, increasing β from below 1 to above 1 will weakly increase the proportion of group members who choose ϕ.

Antanas Mockus provides an intriguing instance of how this can be done, as related by Fisman and Miguel (2010). The major of Bogotá placed mimes on the streets of the city to ridicule jaywalkers. Although jaywalking was widespread, it was considered irresponsible. The role of mimes was to make it more embarrassing for street walkers to act in ways that revealed their preference does not match that of a responsible citizen. The impact of the policy was to reduce jaywalking dramatically.

5.6 General Discussion

I close this section with some general remarks on how principals and policies impact social change in a second order conformity model.

For a principal or a policy to impact behavior, it does not have to be the case that there is common knowledge about the principal's actions or the policy. In a second order conformity model, only second order beliefs are required, i.e., beliefs about beliefs. For judges to update their beliefs about what is appropriate, it is necessary for them to observe the action by $\mathcal{P}$, not for it to be common knowledge. Therefore, decision makers only need to know judges knows what $\mathcal{P}$ did. This contrasts to strategic complementarities models, in which higher order beliefs are necessary to impact equilibrium beliefs, at times requiring common knowledge (Morris and Shin, 2002). Higher order knowledge can be built into a second order conformity model. For example, if judges worried about how they will be judged for how they judge a decision maker, then a decision maker must worry about third order beliefs. This could iterate indefinitely. However, I conjecture that the empirically relevant results involve second order reasoning. Furthermore, we have some evidence that individuals do reason at about two orders of belief (Camerer et al., 2004).

Hyde (2011) develops a theory of norm diffusion which does not fit neatly into the setup of section 2. To explain the spread of election monitoring, she argues that democratic countries gradually recognized that it served as a good signal of a new democracy's type. Once countries learned that democratic countries considered electoral monitoring a good signal, non-established democracies and non-democracies adopted the practice. This story does not fall neatly in my model because it seems that the established democracies are judging whether non established democracy's type matches theirs. Therefore, to capture this dynamic properly, we would have to allow for a group of judges and a group of those who are judged. Of course, this is a natural extension, and the desire of a decision maker to have others believe they share the same type is the essence of second order conformity.

6 Conclusion

The paper introduced a model of second order conformity, in which individuals are motivated to do what they think the majority of their group prefers. This behavioral assumption allowed me to capture how pluralistic ignorance can lead to a perverse failure of collective action, and provides a novel result regarding the impact of context and group size on the probability of pluralistic ignorance. I then use the framework to discuss how different principals and policies can affect social change. The analysis allows me to shed novel light on social information campaigns and the expressive function of the law, and to capture logics of leadership that has received less focus in the game theoretic literature such as social meaning regulators and uninformed catalysts.

A running theme in the paper is that second order conformity provides formal tools for thinking about phenomena studied by social scientists who do not emphasize rational choice explanations. I argued that the model naturally captures some concepts traditionally outside of game theory, such as appropriate behaviors and context-dependent preferences, and provides a new way of capturing concepts like norm emergence. I apply these concepts to a variety of political and non-political examples, including interstate crisis bargaining, opinions on climate change and gift-giving practices.

It is worth highlighting that several situations in the paper have already been analyzed insightfully with game theoretic tools. A lot has been learned by these analyses, and a formal theorist may find it jarring to not see strategic complementarities in a model of protests (say). The claim made in this paper is that part of what motivates players in these situations is to do what they think others want, which cannot be reduced to more standard motivations. Much of the strength of this argument rests on whether there are novel insights that can be obtained from a second order conformity approach. To highlight these insights, the model abstracted from other motivations more common in the literature.

An interesting question for future work is considering the impact of combining second order conformity with other behavioral motivations. Relatedly, experimental evidence is needed to provide evidence for the behavioral impact of social expectations (as we do in Fernández-Duque and Hiscox, 2017), and to test for the novel predictions of the model (as proposed in section 4.4.1). Finally, future work would benefit from considering a continuous version to avoid dealing with mixed strategies. On the one hand, allowing for mixed strategies in the current model will generally not affect the conclusions much: although individuals learn more slowly, they are still motivated to choose the action that is widely considered the group majority. Further, as I show in the appendix, mixed strategies are rare under the assumptions I considered. On the other hand, mixed strategies quickly complicate the model and make it hard to consider dynamics when individuals are neither very certain nor very uncertain about which population the group's preferences were drawn from.

References

Rania Abouzeid. Bouazizi: The man who set himself and tunisia on fire. *TIME*, 2011.

Daron Acemoglu and Matthew O Jackson. History, expectations, and leadership in the evolution of social norms. *The Review of Economic Studies*, 82(2):423–456, 2014.

Daron Acemoglu and Matthew O Jackson. Social norms and the enforcement of laws. *Journal of the European Economic Association*, 15(2):245–295, 2017.

Daron Acemoglu, Munther A Dahleh, Ilan Lobel, and Asuman Ozdaglar. Bayesian learning in social networks. *The Review of Economic Studies*, 78(4):1201–1236, 2011.

S Nageeb Ali and Navin Kartik. Herding with collective preferences. *Economic Theory*, 51 (3):601–626, 2012.

Hunt Allcott. Social norms and energy conservation. *Journal of public Economics*, 95(9): 1082–1095, 2011.

Francisco Alpizar, Fredrik Carlsson, and Olof Johansson-Stenman. Anonymity, reciprocity, and conformity: Evidence from voluntary contributions to a national park in costa rica. *Journal of Public Economics*, 92(5):1047–1060, 2008.

George-Marios Angeletos, Christian Hellwig, and Alessandro Pavan. Signaling in a global game: Coordination and policy traps. *Journal of Political economy*, 114(3):452–484, 2006.

Kenneth Arrow. Political and economic evaluation of social effects and externalities. In *The analysis of public output*, pages 1–30. NBER, 1970.

Abhijit V Banerjee. A simple model of herd behavior. *The Quarterly Journal of Economics*, 107(3):797–817, 1992.

Robert H Bates, Rui JP de Figueiredo Jr, and Barry R Weingast. The politics of interpretation: rationality, culture, and transition. *Politics & Society*, 26(4):603–642, 1998.

Dawn Michelle Baunach. Changing same-sex marriage attitudes in america from 1988 through 2010. *Public Opinion Quarterly*, 76(2):364–378, 2012.

Roland Bénabou and Jean Tirole. Identity, morals, and taboos: Beliefs as assets. *The Quarterly Journal of Economics*, 126(2):805–855, 2011.

Roland Benabou and Jean Tirole. Laws and norms. Technical report, National Bureau of Economic Research, 2012.

B Douglas Bernheim. A theory of conformity. *Journal of political Economy*, 102(5):841–877, 1994.

John Beshears, James J Choi, David Laibson, Brigitte C Madrian, and Katherine L Milkman. The effect of providing peer information on retirement savings decisions. *The Journal of finance*, 70(3):1161–1201, 2015.

Cristina Bicchieri. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press, 2005.

Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy*, 100(5):992–1026, 1992.

Taylor Branch. *Parting the waters: Martin Luther King and the Civil rights movement, 1954-63*. Macmillan, 1988.

Donald E Brown. *Human universals*. McGraw-Hill New York, 1991.

James M Burns. leadership. ny, 1978.

Colin F Camerer, Teck-Hua Ho, and Juin-Kuan Chong. A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3):861–898, 2004.

Brandice Canes-Wrone, Michael C Herron, and Kenneth W Shotts. Leadership and pandering: A theory of executive policymaking. *American Journal of Political Science*, pages 532–550, 2001.

Hans Carlsson and Eric Van Damme. Global games and equilibrium selection. *Econometrica: Journal of the Econometric Society*, pages 989–1018, 1993.

Damon Centola, Robb Willer, and Michael Macy. The emperors dilemma: A computational model of self-enforcing norms. *American Journal of Sociology*, 110(4):1009–1040, 2005.

Michael Suk-Young Chwe. Communication and coordination in social networks. *The Review of Economic Studies*, 67(1):1–16, 2000.

Michael Suk-Young Chwe. *Rational ritual: Culture, coordination, and common knowledge.* Princeton University Press, 2013.

Robert B Cialdini, Raymond R Reno, and Carl A Kallgren. A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of personality and social psychology*, 58(6):1015, 1990.

J Coleman. Foundations of social theory. cambridge, mass.: Belknap press of harvard university press. 1990.

Alex Cukierman and Mariano Tommasi. When does it take a nixon to go to china? *American Economic Review*, pages 180–197, 1998.

Jason Dana, Daylian M Cain, and Robyn M Dawes. What you dont know wont hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and human decision Processes*, 100(2):193–201, 2006.

Jason Dana, Roberto A Weber, and Jason Xi Kuang. Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80, 2007.

Stefano DellaVigna, John A List, and Ulrike Malmendier. Testing for altruism and social pressure in charitable giving. *The quarterly journal of economics*, 127(1):1–56, 2012.

Martin Duberman. *Stonewall.* Open Road Media, 2013.

Robert B Edgerton. *Sick societies.* Simon and Schuster, 1992.

Chris Edmond. Information manipulation, coordination, and regime change. *Review of Economic Studies*, 80(4):1422–1458, 2013.

Tore Ellingsen, Magnus Johannesson, et al. Pride and prejudice: The human side of incentive theory. *American Economic Review*, 98(3):990–1008, 2008.

Jon Elster. *The cement of society: A survey of social order.* Cambridge University Press, 1989.

Joan Esteban. Collective action and the group size paradox. *American political science review*, 95(3):663–672, 2001.

Erik Eyster and Matthew Rabin. Naive herding in rich-information settings. *American economic journal: microeconomics*, 2(4):221–243, 2010.

Hanming Fang. Social culture and economic performance. *American Economic Review*, pages 924–937, 2001.

James Fearon and Alexander Wendt. Rationalism v. constructivism: a skeptical view. *Handbook of international relations*, pages 52–72, 2002.

James D Fearon. Rationalist explanations for war. *International organization*, 49(03):379–414, 1995.

Ernst Fehr and Urs Fischbacher. The nature of human altruism. *Nature*, 425(6960):785, 2003.

Mauricio Fernández-Duque and Michael Hiscox. Leadership and social expectations. *unpublished manuscript*, 2017.

Martha Finnemore. National interests in international society. 1996.

Martha Finnemore and Kathryn Sikkink. International norm dynamics and political change. *International organization*, 52(04):887–917, 1998.

Ray Fisman and Edward Miguel. *Economic gangsters: corruption, violence, and the poverty of nations*. Princeton University Press, 2010.

Bruno S Frey and Stephan Meier. Social comparisons and pro-social behavior: Testing" conditional cooperation" in a field experiment. *The American Economic Review*, 94(5): 1717–1722, 2004.

Drew Fudenberg and Jean Tirole. Game theory, 1991. *Cambridge, Massachusetts*, 393:12, 1991.

James L Gelvin. *The Arab uprisings: what everyone needs to know*. Oxford University Press, USA, 2015.

Alexander L George. *Avoiding war: Problems of crisis management*. Westview Pr, 1991.

Alan S Gerber and Todd Rogers. Descriptive social norms and motivation to vote: Everybody's voting and so should you. *The Journal of Politics*, 71(1):178–191, 2009.

Noah J Goldstein, Robert B Cialdini, and Vladas Griskevicius. A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of consumer Research*, 35(3):472–482, 2008.

Benjamin Golub and Matthew O Jackson. Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–149, 2010.

Ezequiel Gonzalez-Ocantos, Chad Kiewiet De Jonge, Carlos Meléndez, Javier Osorio, and David W Nickerson. Vote buying and social desirability bias: Experimental evidence from nicaragua. *American Journal of Political Science*, 56(1):202–217, 2012.

Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83(6):1420–1443, 1978.

Ramachandra Guha. *India after Gandhi: The history of the world's largest democracy*. Pan Macmillan, 2017.

Russell Hardin. *Collective Action*. Resources for the Future, 1982.

Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath. In search of homo economicus: behavioral experiments in 15 small-scale societies. *The American Economic Review*, 91(2):73–78, 2001.

Benjamin E Hermalin. Toward an economic theory of leadership: Leading by example. *American Economic Review*, pages 1188–1206, 1998.

Michael A Hogg. A social identity theory of leadership. *Personality and social psychology review*, 5(3):184–200, 2001.

Edwin P Hollander. Conformity, status, and idiosyncrasy credit. *Psychological review*, 65 (2):117, 1958.

Susan D Hyde. *The pseudo-democrat's dilemma: why election observation became an international norm*. Cornell University Press, 2011.

Robert Jervis. Cooperation under the security dilemma. *World politics*, 30(2):167–214, 1978.

Jeffrey A Karp and David Brockington. Social desirability and response validity: A comparative analysis of overreporting voter turnout in five countries. *Journal of Politics*, 67 (3):825–840, 2005.

Daniel Katz and Floyd H Allport. Students attitudes. *Syracuse, NY: Craftsman*, page 152, 1931.

Patrick Kenny, Gerard Hastings, Hastings, G., Angus, K., Bryant, and C. Understanding social norms: Upstream and downstream applications for social marketers. *The SAGE handbook of social marketing*, pages 61–79, 2011.

James A Kitts. Egocentric bias or information management? selective disclosure and the social roots of norm misperception. *Social Psychology Quarterly*, pages 222–237, 2003.

Erin L Krupka and Roberto A Weber. Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3):495–524, 2013.

Timur Kuran. *Private truths, public lies: The social consequences of preference falsification.* Harvard University Press, 1997.

Kypros Kypri and Brett Maclennan. Commentary on melson et al.(2011): Pluralistic ignorance is probably real but important questions remain about its relation to drinking and role in intervention. *Addiction*, 106(6):1085–1086, 2011.

James J Lee and Steven Pinker. Rationales for indirect speech: the theory of the strategic speaker. *Psychological review*, 117(3):785, 2010.

Lawrence Lessig. The regulation of social meaning. *The University of Chicago Law Review*, 62(3):943–1045, 1995.

Richard J. Light. How to live wisely. *The New York Times*, 2015.

Susanne Lohmann. A signaling model of informative and manipulative political action. *American Political Science Review*, 87(2):319–333, 1993.

Susanne Lohmann. The dynamics of informational cascades: The monday demonstrations in leipzig, east germany, 1989–91. *World politics*, 47(1):42–101, 1994.

Arthur Lupia and Mathew D McCubbins. *The democratic dilemma: Can citizens learn what they need to know?* Cambridge University Press, 1998.

Sumon Majumdar and Sharun Mukand. The leader as catalyst-on leadership and the mechanics of institutional change. 2008.

James G March and Johan P Olsen. The logic of appropriateness. *The Oxford Handbook of Public Policy*, 2004.

Eric Maskin and Jean Tirole. The politician and the judge: Accountability in government. *The American Economic Review*, 94(4):1034, 2004.

Doug McAdam, John D McCarthy, and Mayer N Zald. *Comparative perspectives on social movements: Political opportunities, mobilizing structures, and cultural framings*. Cambridge University Press, 1996.

Richard H McAdams. *The expressive powers of law: Theories and limits*. Harvard University Press, 2015.

Ambrose John Melson, John B Davies, and Theresa Martinus. Overestimation of peer drinking: error of judgement or methodological artefact? *Addiction*, 106(6):1078–1084, 2011.

Matto Mildenberger and Dustin Tingley. Beliefs about climate beliefs: The problem of second-order climate opinions in climate policymaking. *unpublished manuscript*, 2016.

Stephen Morris and Hyun Song Shin. Social value of public information. *The American Economic Review*, 92(5):1521–1534, 2002.

Hubert J O'Gorman. Pluralistic ignorance and white estimates of white support for racial segregation. *Public Opinion Quarterly*, 39(3):313–330, 1975.

Mancur Olson. *Logic of Collective Action: Public Goods and the Theory of Groups (Harvard economic studies. v. 124)*. Harvard University Press, 1965.

H Wesley Perkins, Jeffrey W Linkenbach, Melissa A Lewis, and Clayton Neighbors. Effectiveness of social norms media marketing in reducing drinking and driving: A statewide campaign. *Addictive behaviors*, 35(10):866–874, 2010.

Steven Pinker. *The blank slate: The modern denial of human nature*. Penguin, 2003.

Srdja Popovic and Matthew Miller. *Blueprint for Revolution: How to Use Rice Pudding, Lego Men, and Other Nonviolent Techniques to Galvanize Communities, Overthrow Dictators, or Simply Change the World*. Spiegel & Grau, 2015.

Deborah A Prentice and Dale T Miller. Pluralistic ignorance and alcohol use on campus: some consequences of misperceiving the social norm. *Journal of personality and social psychology*, 64(2):243, 1993.

Raymond R Reno, Robert B Cialdini, and Carl A Kallgren. The transsituational influence of social norms. *Journal of personality and social psychology*, 64(1):104, 1993.

Andrew Schotter et al. The economic theory of social institutions. *Cambridge Books*, 1981.

Warren F Schwartz, Keith Baxter, and David Ryan. The duel: can these gentlemen be acting efficiently? *The Journal of Legal Studies*, 13(2):321–355, 1984.

Jacob Shamir and Michal Shamir. *The anatomy of public opinion.* University of Michigan Press, 2000.

David A Siegel. Social networks and collective action. *American Journal of Political Science*, 53(1):122–138, 2009.

Dirk Sliwka. Trust as a signal of a social norm and the hidden costs of incentive schemes. *American Economic Review*, 97(3):999–1012, 2007.

Robert Sugden. Spontaneous order. *The Journal of Economic Perspectives*, 3(4):85–97, 1989.

Cass R Sunstein. *Free markets and social justice.* Oxford University Press, 1999.

Sidney G Tarrow. *Power in movement: Social movements and contentious politics.* Cambridge University Press, 2011.

Justin Taylor. 5 myths about rosa parks, the woman who had almost a 'biblical quality'. *The Washington Post*, 2015.

Jeanne Theoharis. How history got the rosa parks story wrong. *The Washington Post*, 2015.

Leaf Van Boven. Pluralistic ignorance and political correctness: The case of affirmative action. *Political Psychology*, 21(2):267–276, 2000.

Joel Waldfogel. *Scroogenomics: Why you shouldn't buy presents for the holidays.* Princeton University Press, 2009.

Alexander Wendt. Anarchy is what states make of it: the social construction of power politics. *International organization*, 46(02):391–425, 1992.

Michael Wenzel. Misperceptions of social norms about tax compliance: From theory to intervention. *Journal of Economic Psychology*, 26(6):862–883, 2005.

H Peyton Young. The evolution of conventions. *Econometrica: Journal of the Econometric Society*, pages 57–84, 1993.

A Derivation of Judgment Function

Suppose judge j is trying to determine whether the probability that i's type matches the majority preference of the group without i is more than 1/2. Then $-(\mathcal{J}_{j,i} - \mathbb{1}\{\phi^m_{-i} = \phi_i\})^2$ is the payoff to j from judging i. j chooses $\mathcal{J}_{j,i}$ to maximize his expected utility: $\mathcal{J}_{j,i} \in [0, 1]$ maximizes

$$-P(\phi_i = \phi^m_{-i} \mid h_i, a_i, \phi_j)(J_{j,i} - 1)^2 - P(\phi_i \neq \phi^m_{-i} \mid h_i, a_i, \phi_j)(J_{j,i} - 0)^2$$

that is,

Lemma 1. *$\mathcal{J}^*_{j,i} = 1$ if and only if $P(\phi^m_{-i} = x \mid h_i, a_i, \phi_j) > 1/2$ and $P(\phi_i = x \mid h_i, a_i, \phi_j) > 1/2$ for some $x \in \{0, 1\}$. If one of the probabilities is equal to 1/2, then $\mathcal{J}^*_{j,i} \in [0, 1]$.*

Proof. Use the law of iterated expectations to write $P(\phi_i = \phi^m_{-i} \mid h_i, a_i, \phi_k) =$

$$\sum_{x,y\in\{0,1\}} P(\theta = x, \phi_i = y \mid h_i, a_i, \phi_k)P(\phi^m_{-i} = y \mid h_i, a_i, \phi_k, \phi_i = y, \theta = x)$$

By Bayes' rule, $P(\phi_i, \theta \mid h_i, a_i, \phi_k) = P(\phi \mid \theta, h_i, a_i, \phi_k)P(\theta \mid h_i, a_i, \phi_k)$.

Further, $P(\phi^m_{-i} \mid h_i, a_i, \phi_k, \phi_i, \theta) = P(\phi^m_{-i} \mid h_i, a_i, \phi_k, \theta)$, since once we condition on θ, private information becomes irrelevant for determining the state of the world.

We can then rewrite $P(\phi_i = \phi^m_{-i} \mid h_i, a_i, \phi_k) =$ as

$$\sum_{x\in\{0,1\}} P(\theta = x \mid h_i, a_i, \phi_k) \times \sum_{y\in\{0,1\}} P(\phi_i{=}y \mid \theta = x, h_i, a_i, \phi_k)P(\phi^m_{-i} = y \mid h_i, a_i, \phi_k, \theta = x)$$

which, by the law of iterated expectations, is equal to

$$P(\phi_i = 1 \mid h_i, a_i, \phi_k)P(\phi^m_{-i} = 1 \mid h_i, a_i, \phi_k) + P(\phi_i = 0 \mid h_i, a_i, \phi_k)P(\phi^m_{-i} = 0 \mid h_i, a_i, \phi_k)$$

We can therefore conclude that the optimal probability $\mathcal{J}_{j,i}$ with which j judges i to have preferences that match the rest of the group is equal to one, zero and any element of $[0, 1]$ if and only if

$$\sum_{x\in\{0,1\}} P(\phi_i = x \mid h_i, a_i, \phi_j)P(\phi^m_{-i} = x \mid h_i, a_i, \phi_k)$$

is respectively greater, less than, or equal to 1/2. The result follows from the fact that the objective function is the summation of two probabilities and their complements.□

B A Generalization of Results

In this section I provide a generalization of the main results in the paper. In order to do so, I need to deal with semi-separating strategies, which will sometimes be the only type of equilibrium strategies. This requires us to expand the notation, which I do in subsection B.1. We then present preliminary Lemmas before turning to the main results in subsection B.2. Proofs are in separate appendices.

B.1 Mixed Strategies, Classes of Types and Full Types

In this subsection I generalize some of the definitions from the body of the text.

Let $\sigma_k(a_k \mid h_k, \phi_k)$ be the probability k of type ϕ_k chooses a_k after history h_k. We say that i *semi-separates on* a if $\sigma^*_i(a \mid \phi_i = x, h_i) = 1 > \sigma^*_i(a \mid \phi_i \neq x, h_i) > 0$. If an equilibrium is *non-reversing*, it must be the case that $x = a$.

Individuals who separate or semi-pool will continue to judge others, but what is known about their private preference is different from before they made their decision. Therefore, decision maker i will hold different beliefs about how they'll be judged by j after j separates or semi-pools. It is therefore useful to define a *class of types* according to the pair of probabilities with which each player's types chose an observed action. Index this pair of probabilities with the set $[-1, 1]$. Let $\phi^{obs}_j : H \to [-1, 1]$, which maps history h_i to this set, indicate the class of type of j after history h_i. I will often omit the argument of ϕ^{obs}_j for simplicity and because it will not lead to ambiguity. Since I will show that there are no optimal strategies where both players mix, the following assignment fully describes equilibrium strategies:

Assign $\phi^{obs}_j = 0$ to withholders. Let $\phi^{obs}_j > 0$ denote when i semi-separates on 1 with $\phi_j = 0$ choosing 0 probability ϕ^{obs}_j. Let $\phi^{obs}_j < 0$ denote when i semi-separates on 0 with $\phi_j = 1$ choosing 1 with probability $-\phi^{obs}_j$.

Note that $\phi^{obs}_j = -1$ and $\phi^{obs}_j = 1$ are, respectively, revealers of types 1 and 0. We will refer to individuals with class of type $\phi^{obs} \in (-1, 0) \cup (0, 1)$ as semi-withholders. The probability that j is of type 1 increases in ϕ^{obs}_j. Let d_{-i} be the set of players at history h_i who have a class of type distinct from all players who choose an action before them.

The tuple of types and class of types (ϕ_j, ϕ^{obs}_j) is the *full type* of player j of type ϕ_j. If judge k of full type (ϕ_j, ϕ^{obs}_j) sets $\mathcal{J}_{j,i} = 1$, I say '(ϕ_j, ϕ^{obs}_j) judges believe i's preference matches those of the group'. If it is clear from context, I simply say '(ϕ_j, ϕ^{obs}_j) believes preferences match'. If i semi-separates, then generically it will make some full type judge j indifferent between $\mathcal{J}_{j,i} = 1$ and $\mathcal{J}_{j,i} = 0$ in equilibrium. We set $\mathcal{J}_{j,i} \in [0, 1]$ such that i is willing to semi-separate. Existence of a semi-separating strategy when pooling or separating are not equilibrium strategies are guaranteed by standard arguments.

B.2 Belief Formation And Social Expectations

By using Bayes' rule and the law of iterated expectation, we can rewrite the decision-maker's choice (1) as follows: $a_i^*(\phi_i, h_i)$ is equal to 1, an element of $[0,1]$ or equal to 0 if and only if

$$\sum_{j\in d_{-i}} P(\phi_{-i}^{obs} = \phi_j^{obs} \mid h_i) \left\{ \sum_{x\in\{0,1\}} P(\phi_j = x \mid h_i, a_i, \phi_i) \left[\mathcal{J}_{j,i}^*(h_i, a_i{=}1, \phi_j) - \mathcal{J}_{j,i}^*(h_i, a_i{=}0, \phi_j)\right] \right\} \tag{5}$$

is respectively less than, equal or greater than $(1-2\phi_i)/\beta$, where $P(\phi_{-i}^{obs} \mid h_i)$ is the probability mass function of classes of types other than i given h_i. i considers the difference in the expected judgment of class of type ϕ_j^{obs} from choosing 1 versus choosing 0 (the term in curly brackets), and takes the weighted sum of this difference for all classes of types at i. Note that i's private information only enters the term by affecting the probability i assigns to judge j being of a certain type (the probability in the curly brackets).

For judge j's choice we only need to focus on $P(\phi_i \mid h_i, a_i, \phi_j) \geq 1/2$ and on $P(\phi_{-i}^m = x \mid h_i, a_i, \phi_j) \geq 1/2$. Begin by unpacking the probability that the majority of a group is of type x, or

$$P(\phi_{-i}^m = x \mid h_i, a_i, \phi_j) = \sum_{k\in d_{-i}} P(\phi_{-i}^{obs} = \phi_k^{obs} \mid h_i, \phi_j) P(\phi_k = x \mid h_i, a_i, \phi_j) \tag{6}$$

where $P(\phi_{-i}^{obs} = \phi_k^{obs} \mid h_i, \phi_j)$ is the probability that classes of types other than i are equal to ϕ_k^{obs} given h_i and ϕ_j. Note how it differs from the first multiplicand of the decision maker's choice in equation (5). Both equations (5) and (6) are sums weighted by the probability judges of i are of a certain class of types. The decision maker i is not part of the judges of i, so she uses public information to form beliefs about others' classes of types. In contrast, j is part of i's judges. She uses public information to form beliefs about other judges, but she has private information about her own preferences. Further note that i's action does not reveal information about classes of types that chose their action before i.

Lemma 2. *$P(\phi_k = 1 \mid h_i, a_i, \phi_j) > z$ if and only if one of two mutually exclusive conditions hold:*

1. *$z \geq P(\phi_k = 1 \mid \theta = 0, h_i, a_i)$ and*

$$\prod_{l\in\{1,\ldots,i\}\setminus\{j\}} \left(\frac{P(a_l \mid h_l, \theta{=}1)}{P(a_l \mid h_l, \theta{=}0)}\right) \frac{P(\phi_j \mid \theta{=}1)}{P(\phi_j \mid \theta{=}0)} \left[\frac{P(\phi_k{=}1 \mid \theta{=}1, h_i, a_i) - z}{z - P(\phi_k{=}1 \mid \theta{=}0, h_i, a_i)}\right] > \frac{P(\theta{=}0)}{P(\theta{=}1)} \tag{7}$$

2. *$z < P(\phi_k = 1 \mid \theta = 0, h_i, a_i)$*

$P(\phi_{-i}^m = 1 \mid h_i, a_i, \phi_j) > z'$ if and only if one of two mutually exclusive conditions hold:

1. $z^{'} > P(\phi^m_{-i} = 1 \mid h_i, a_i, \phi_j, \theta = 0)$ *and*

$$\prod_{l \in \{1,\ldots,i\} \setminus \{j\}} \left(\frac{P(a_l \mid h_l, \theta{=}1)}{P(a_l \mid h_l, \theta{=}0)} \right) \frac{P(\phi_j \mid \theta{=}1)}{P(\phi_j \mid \theta{=}0)} \left\{ \frac{P(\phi^m_{-i}{=}1 \mid h_i, a_i, \phi_j, \theta{=}1) - z^{'}}{z^{'} - P(\phi^m_{-i}{=}1 \mid h_i, a_i, \phi_j, \theta{=}0))} \right\} > \frac{P(\theta{=}0)}{P(\theta{=}1)} \tag{8}$$

2. $z^{'} < P(\phi^m_{-i} = 1 \mid h_i, a_i, \phi_j, \theta = 0)$

Lemma 2 rewrites the probabilities of interest in terms of likelihood ratios. Notice that the first and second terms are identical in the left hand sides of (7) and (8). The first and second terms capture what j knows about the state of the world through the signals she's observed – others' and her own, respectively. j's private information only appears in the left hand side of (7) in the second term. A judge j of type ϕ_j will have stronger beliefs than j of type $1 - \phi_j$ that withholders or semi-revealers are of type ϕ_j – the left hand side of (7) is weakly higher for $\phi_j = 1$ than for $\phi_j = 0$.

Judge j's private information also enters the left hand side of (8) in the third term. The third term in the left hand sides of (7) and (8) are the only terms that differ, even with $z = z^{'}$. The terms respectively isolate what can be learned from k given her strategy and what can be learned from players other than i given their strategy if the state of the world is known.

Let $\tilde{f}(h_i, i, \phi^{obs}_i, \phi^{obs}_j, \phi_j, z; I)$ equal the left hand side of (7), and $\tilde{g}(h_i, i, \phi^{obs}_i, \phi^{obs}_j, \phi_j, z; I)$ equal the left hand side of (8). Based on Lemma 2, it is worth establishing the following:

Definition 7.

$$f(h_i, i, \phi^{obs}_i, \phi^{obs}_j, \phi_j, z; I) \equiv \begin{cases} \tilde{f} & \textit{if } z > P(\phi_i = 1 \mid h_i, a_i, \theta = 0) \\ \infty & \textit{if } z = P(\phi_i = 1 \mid h_i, a_i, \theta = 0) \\ \infty^+ & \textit{if } z < P(\phi_i = 1 \mid h_i, a_i, \theta = 0) \end{cases}$$

$$g(h_i, i, \phi^{obs}_i, \phi^{obs}_j, \phi_j, z'; I) \equiv \begin{cases} \tilde{g} & \textit{if } z^{'} > P(\phi^m_{-i} = 1 \mid h_i, a_i, \phi_j, \theta = 0) \\ \infty & \textit{if } z^{'} = P(\phi^m_{-i} = 1 \mid h_i, a_i, \phi_j, \theta = 0) \\ \infty^+ & \textit{if } z^{'} < P(\phi^m_{-i} = 1 \mid h_i, a_i, \phi_j, \theta = 0) \end{cases}$$

with $\infty < \infty^+$.

By Lemma 2, we know that if f is equal to ∞^+, then the probability that k is of type 1 is greater than z. Similarly, if g is equal to ∞^+, then the probability that the majority of preferences in the group without i is of type 1 is greater than $z^{'}$. The term ∞^+ allows us to consider the left hand sides of (7) and (8) in one dimension: $(-\infty, \infty] \cup \infty^+$.

Since j's preference affects the average preference in the group, the third term of the left hand side of (8) is higher for j of type 1. But then the second and third term of the left hand side of (8) bias judge j of type ϕ_j in a similar way – the left hand side of (8) for $z^{'} = 1/2$ is weakly higher for $\phi_j = 1$ than for $\phi_j = 0$. That is, judge j of type ϕ_j will have relatively stronger beliefs that the average preferences of individuals in the group without i will be of type ϕ_j. Therefore, the more likely a judge j is of type x, a favorable judgment ($\mathcal{J}_{j,i} = 1$) is easier to obtain if i is more likely to be of type x. Judges are biased in their judgment towards their type.

Similarly, since $P(\phi_k = 1 \mid h_i, a_i, \phi_i)$ is the only term in which i's private information appears in the decision-maker i's problem (5), i of type ϕ_i has relatively stronger beliefs that judges who have not fully revealed their type are of type ϕ_i. These mutual biases allow us to state the following:

Lemma 3. *1. In equilibrium, there is no individual whose types are both randomizing.*

2. There exists a unique $\tilde{y}(\phi_k^{obs}, \theta) \in [-1, 1]$, increasing in ϕ_k^{obs}, with the following characteristics

- *Suppose k semi-separates on one. Then $\tilde{y}(\phi_k^{obs}, \theta) \in [0, 1]$ such that*

$$P(\phi_k = 1 \mid h_i, a_i, \theta) = (1 - \tilde{y}(\phi_k^{obs}, \theta))P(\phi_k = 1 \mid \theta) + \tilde{y}(\phi_k^{obs}, \theta),$$

with $\tilde{y}(\phi_k^{obs}, \theta) = 1$ if and only if $\phi_k^{obs} = 1$, $\tilde{y}(\phi_k^{obs}, \theta) = 0$ if and only if $\phi_k^{obs} = 0$.

- *Suppose k semi-separates on zero. Then $\tilde{y}(\phi_k^{obs}, \theta) \in [-1, 0]$ such that*

$$P(\phi_k = 1 \mid h_i, a_i, \theta) = (1 + \tilde{y})P(\phi_k = 1 \mid \theta),$$

with $\tilde{y} = -1$ if and only if $\phi_k^{obs} = -1$.

3. There exists a unique $\tilde{x}(\phi_i^{obs}) \in [-1, 1]$, increasing in ϕ_i^{obs}, such that,

$$\frac{P(a_i \mid h_i, \theta{=}1)}{P(a_i \mid h_i, \theta{=}0)} = \left(\frac{P(\phi_i = 1 \mid h_i, \theta = 1)}{P(\phi_i = 1 \mid h_i, \theta = 0)}\right)^{\tilde{x}} \tag{9}$$

$\tilde{x}(-1) = -1$, $\tilde{x}(0) = 0$ and $\tilde{x}(1) = 1$.

Lemma 3 shows that the class of types ϕ_k^{obs} I defined characterize the equilibrium strategies, although note that separating leads to two possible values of ϕ_i^{obs}, -1 and 1. Further, ϕ_k^{obs} summarizes the information given by k's action. If $\phi_k^{obs} = 0$, k does not reveal any information about his own type or about the state of the world. A larger positive value of

ϕ_k^{obs} increases the probability that his own type and the state of the world are equal to one, while a more negative value increases the probability that they are equal to zero.

The proof of Lemma 3 also shows why semi-separating equilibria complicate the analysis: $\tilde{x}(\cdot) \neq \tilde{y}(\cdot)$ are non-linear functions that make $f(\cdot,\cdot,\cdot,\cdot)$ and $g(\cdot,\cdot,\cdot,\cdot)$ hard to keep track of when i semi-separates. Notice we *cannot* redefine the action lead for action 1 at period h_i as $\Delta(h_i) \equiv \sum_{k<i} \phi_k^{obs}$ – the sum does uniquely identify a sum of $\sum_{k<i} \tilde{x}(\phi_k^{obs})$ or $\sum_{k<i} \tilde{y}(\phi_k^{obs})$ if players have semi-separated. The results will therefore mostly focus on analyzing histories of play such that past players have separated or pooled.

The class of type of the decision maker i (the player making a decision at the current period) does not affect her beliefs about the majority preference of the group without i. Therefore, I will write $g_i(h_i, \phi_i, z'; I)$ as the value of g for decision maker i. The following lemma will allow us to describe how f and g depend on the history of play.

Lemma 4. 1. *f and g are higher for type 1 individuals. For an individual k of type 0 (1), the more (less) likely others think k is of type 0 (1), the higher are f and g.*

$f(h_i, i, \phi_i^{obs}, \phi_j^{obs}, 0, z; I) \leq f(h_i, i, \phi_i^{obs}, \widehat{\phi}_k^{obs}, 1, z; I)$ for any ϕ_j^{obs}, $\widehat{\phi}_k^{obs}$, with a strict inequality if and only if $f(h_i, i, \phi_i^{obs}, \phi_j^{obs}, 0, z; I) \in (-\infty, \infty)$; $f(h_i, i, \phi_i^{obs}, -1, 0, z; I) = f(h_i, i, \phi_i^{obs}, 1, 1, z; I)$; and $\partial f / \partial \phi_j^{obs} \leq 0$, with a strict inequality if and only if $f \in (-\infty, \infty)$. The same results hold replacing f with g.

2. *f is higher the more likely $k \leq i$ is of type 1.. f becomes arbitrarily large or small depending on i's strategy.*

If $k \leq i$, $\partial f / \partial \phi_k^{obs} > 0$ if $f(h_i, i, \phi_i^{obs}, \phi_j^{obs}, \phi_j, z; I) \in (0, \infty)$, $\partial f / \partial \phi_k^{obs} = 0$ if $z < P(\phi_i = 1 \mid h_i, a_i, \theta = 0)$.

$f(h_i, i, 0, \phi_j^{obs}, \phi_j; I) \in (0, \infty)$, $f(h_i, i, 1, \phi_j^{obs}, \phi_j; I) = \infty^+$, $f(h_i, i, -1, \phi_j^{obs}, \phi_j; I) < 0$.

3. *g is higher the more likely $k \leq i$ is of type 1. The value g takes as a function of i's strategy depends on the range of values g is in when i pools.*

For $k \leq i$, $\partial g / \partial \phi_k^{obs} > 0$ if $g(h_i, i, 0, \phi_j^{obs}, \phi_j, z'; I) \in (0, \infty)$, $\partial g / \partial \phi_k^{obs} = 0$ if $z' < P(\phi_{-i}^m = 1 \mid h_i, a_i, \phi_j, \theta = 0)$.

If $g(h_i, i, 0, \phi_j^{obs}, \phi_j, z'; I) \in X$, then $g(h_i, i, \phi_i^{obs}, \phi_j^{obs}, \phi_j, z'; I) \in X$ for all ϕ_i^{obs}, $X \in \{(-\infty, 0], (0, \infty), \infty, \infty^+\}$.

Suppose at history h_i, all players $k < i$ have separated or pooled.

4. *The values of f and g at history h_i when $\phi_i^{obs} = 0$ and $z = z' = 1/2$ are bounded by a value which depends on the action lead of action 1.*

If $\Delta(h_i) + \mathbb{1}\{\phi_j^{obs}(h_i) = 0\}(2\phi_j - 1)$ *is less than, equal or greater than 0, then* $f(h_i, i, 0, \phi_j^{obs}, \phi_j, 1/2; I)$ *is respectively less than, equal or greater than 1, and* $g(h_i, i, 0, \phi_j^{obs}, \phi_j, 1/2; I)$ *is respectively less than, equal or greater than* $f(h_i, i, 0, \phi_j^{obs}, \phi_j, 1/2; I)$*.*

5. *The difference between* f *and* g *at* h_i *before* i *makes a decision decreases with* I *and increases in* i *when the action lead for action 1 is different than 0 and* $z = z' = 1/2$*.*

   $|f(h, i, 0, \phi_j^{obs}, \phi_j, 1/2; I) - g(h, i, 0, \phi_j^{obs}, \phi_j, 1/2; I)|$ *decreases in* I *and increases in* i *for* $\Delta(h) \neq 0$*.*

6. *The value of g is the same for decision maker* i *and for a revealer judge who observes* i*'s signal.*

   $g_i(h_i, \phi_i, z'; I) = g(h_i, i, 2\phi_i - 1, \phi_j^{obs}, \phi_j, z'; I)$ *for* $\phi_j^{obs}, \phi_j \in \{(-1, 0), (1, 1)\}$

7. *The value of f for judge* j *at* h_{i+1} *given that* i *separates can be determined by the value of* f *of judges at history* h_i*.*

   $f(h_i, i, 0, \phi_j^{obs}, \phi_j, z; I) = f(h_{i+1}, i + 1, 0, \phi_k^{obs}, \phi_k, z; I)$ *if for* $\phi_i^{obs} \in \{-1, 1\}$, $\Delta(h_i) + \mathbb{1}\{\phi_j^{obs}(h_i) = 0\}(2\phi_j - 1) = \Delta(h_i) + \mathbb{1}\{\phi_k^{obs}(h_i) = 0\}(2\phi_k - 1) + \phi_i^{obs}$

8. *The value of g for judge* j *at* h_i *given that* i *separates or semi-separates can be bounded by the value of* f *of judges at history* h_i*.*

   Suppose $\phi_i^{obs} \in \{-1, 1\}$*. Then* $g(h_i, i, \phi_i^{obs}, \phi_k^{obs}, \phi_k, z'; I)$ *is greater, equal or less than* $f(h_i, i, 0, \phi_j^{obs}, \phi_j, z'; I)$ *if (a)* $\Delta(h_i) + \mathbb{1}\{\phi_j^{obs}(h_i) = 0\}(2\phi_j - 1) = \Delta(h_i) + \mathbb{1}\{\phi_k^{obs}(h_i) = 0\}(2\phi_k - 1) - \phi_i^{obs}$ *and (b)* $g(h_i, i, 0, \phi_k^{obs}, \phi_k, z'; I)$ *is respectively greater, equal or less than* $f(h_i, i, 0, \phi_k^{obs}, \phi_k, z'; I)$*.*

Lemma 4 allows us to describe the payoffs to decision maker i. If at some history h_i all players $k < i$ have revealed or pooled, it will help us to write $\hat{e}(\Delta(h_i), i, \phi_i^{obs}, \phi_j^{obs}, \phi_j; I) \equiv e(h_i, i, \phi_i^{obs}, \phi_j^{obs}, \phi_j, 1/2; I)$ for $e \in \{f, g\}$, and I will sometimes omit I. Then, at the beginning of the game,

$$\hat{g}(0, 1, 0, 0, 0) < \hat{f}(0, 1, 0, 0, 0) < \hat{g}(0, 1, 0, \phi_j^{obs}, \phi_j) =$$

$$1 = \hat{f}(0, 1, 0, \phi_j^{obs}, \phi_j) < \hat{f}(0, 1, 0, 0, 1) < \hat{g}(0, 1, 0, 0, 1)$$

for $(\phi_j^{obs}, \phi_j) \in \{(-1, 0), (1, 1)\}$, where the inequalities and equalities follow from point 4 of Lemma 4. Note that although there are no revealer judges in the first period, I can still describe their values of $\hat{f}$ and $\hat{g}$.

Which judges believe preferences match will depend on the value of $P(\theta = 0)/P(\theta = 1) \equiv \mathcal{H}(\theta)$. According to Lemma 2, If both the values of $\hat{f}$ and $\hat{g}$ are lower (higher) than $\mathcal{H}$ for some player j, then j thinks preferences match since the majority preference without i and i are most likely to be of type 1 (0). If either $\hat{f}$ or $\hat{g}$ is higher than $\mathcal{H}$, and the other is lower, j thinks preferences do not match. We can apply this to figure out which judges believe preferences match for all values of $\mathcal{H}$. If $\mathcal{H}(\theta) \in (1, \hat{f}(0,1,0,0,1))$, then if i pools all judges believe preferences match: withholder judges of type $\phi \in \{0,1\}$ will believe both player 1 and the majority preference of the group without 1 is more likely to be of type ϕ. If $\mathcal{H}(\theta) > \hat{g}(0,1,0,0,1)$, all judges again believe preferences match if i pools: now all judges believe player 1 and the majority preference of the group without 1 is more likely to be of type 0. However, note that if $\mathcal{H}(\theta) \in (\hat{f}(0,1,0,0,1), \hat{g}(0,1,0,0,1))$, not all judges believe preferences match. Whereas withholders of type 0 believe player 1 and the majority preference of the group without 1 is more likely to be of type 0, withholders of type 1 believe player 1 is more likely to be of type 0, but the majority preference of the group is more likely to be of type 1. This shows that there is a discontinuity in $\mathcal{H}(\theta)$ over player 1's social expectation from pooling. By symmetry, a similar analysis can be made for $\mathcal{H}(\theta) < 1$.

Judges' beliefs if a player separates or semi-separates can also be derived from Lemma 4. An increase in ϕ_1^{obs} increases $\hat{f}(0,1,\phi_1^{obs},\phi_j^{obs},\phi_j)$, from a negative value to ∞^+. That is, what player 1 reveals about herself will make any judge believe she is most likely of type 0 or most likely of type 1 (point 2 of Lemma 4). Although ϕ_1^{obs} also increases $\hat{g}(0,1,\phi_1^{obs},\phi_j^{obs},\phi_j)$, the increase is limited (point 3 of Lemma 4). For any value of ϕ_1^{obs}, f and g will be ordered according to the judges' full type (point 1 of Lemma 4). If player 1 reveals to be of type 1, then

$$\hat{g}(0,1,0,-1,0) < \hat{g}(0,1,1,-1,0) < 1 < \hat{g}(0,1,1,\phi_j^{obs},\phi_j) = \hat{f}(0,1,0,0,1) < \hat{g}(0,1,1,0,1)$$

for $(\phi_j^{obs},\phi_j) \in \{(-1,0),(1,1)\}$ by point 8 of Lemma 4. We then have upper bounds for the values of $\hat{g}$ for any ϕ_1^{obs}, and can similarly derive lower bounds. This allows us to get feasible judgments of player 1 for any equilibrium strategy. For instance, suppose player 1 reveals her type is 1 ($\phi_1^{obs} = 1$) and $\mathcal{H}(\theta) \in (1, \hat{g}(0,1,1,\phi_j^{obs},\phi_j))$. Withholder judges of type $\phi \in \{0,1\}$ will believe the majority preference of the group without 1 is more likely of type 0, so only type 1 judges will believe preferences match.

In order to aggregate these judgments into player 1's social expectations, we need to know how much weight player 1 gives to each type of judge. Point 7 of Lemma 4 tells us that player 1 of type ϕ_1's belief over the majority preference of the group without herself is equal to the belief of a revealer judge who observes ϕ_1. Like player 1, revealer judges have no private information about anybody's preference in the group without 1. Player 1's private information only updates the group about the state of the world, which informs group members about the distribution of preferences they don't observe. If revealers observe signal ϕ_1 about the state of the world, they then have the same information than player 1.

As long as $\mathcal{H}(\theta) < \hat{g}(0, 1, 1, 0, 1)$, we can conclude that player 1 of type 1 believes most players without 1 are of type 1. But then player 1 of type 1 would not deviate from a separating strategy: she would believe at least half of the judges believe preferences match. Since $\beta \in (1, 2)$, this is sufficient to ensure she will choose 1 if doing so reveals she is of type 1. With a similar analysis we can verify whether player 1 of type 0 would deviate from a separating strategy. It is useful to define an individual i's *incentives to separate* as the minimum difference in social expectations from i of type $\phi \in \{0, 1\}$ revealing her own type versus revealing the other type. It is straightforward that separating is an equilibrium strategy for any incentives to separate larger than a given threshold.

Now consider the following

Definition 8. *Let $M(h_i)$ be the majority type at history h_i, with*

$$M(h_i) = \begin{cases} 1 & \text{if } \hat{g}(\Delta(h_i), i, 0, -1, 0) \geq \mathcal{H}(\theta) \\ 0 & \text{if } \hat{g}(\Delta(h_i), i, 0, -1, 0) \leq \mathcal{H}(\theta) \end{cases}$$

If $M(h_i) = \phi$, i of type ϕ would choose ϕ if doing so revealed her type, since she would expect at least half of judges to believe preferences match. We will exploit this observation when using the D1 criterion to refine out-of-equilibrium beliefs.

If at history h_i players $k < i$ have separated or pooled, then we can continue to apply Lemma 4 to describe social expectations. Points 4, 5, 6, and 8 of Lemma 4 provide values or bounds for f and g at period $i + 1$ given the values of f and g past periods. Proceeding in this way, I can state the following:

Result 8. 1. *Incentives to separate increase with β and decrease with $P(\theta = M(h_i))$.*

Suppose at history h_i all players $k < i$ have separated or pooled, and either (a) $\Delta(h_i) \neq 0$ and the context is uninformative; or (b) $\Delta(h_i) \notin \{0, 1 - 2M(h_i)\}$, and the context is neither uninformative nor strongly ϕ. Then incentives to separate weakly decrease with I and weakly increase with i.

2. *Suppose $a_i = M(h_i)$ and decision maker i semi-separates on $M(h_i)$ or separates. Then $i + 1$ semi-separates on $M(h_i)$ or pools. Further, incentives to separate are lower for $i + 1$ than they were for i. If instead $a_i = 1 - M(h_i)$, either incentives to separate are higher for $i + 1$ than they were for i, $i + 1$ semi-separates on $1 - M(h_i)$ or $i + 1$ pools.*

Suppose i pools. Then all players $k > i$ pool.

3. *Suppose at history h_i all players $k < i$ have either separated or pooled. Then the equilibrium strategies depend on $\Delta(h_i)$ and $\mathcal{H}(\theta)$ as in Table 2 for $M(h_1) = 1$ – an analogous Table can be derived for $M(h_1) = 0$.*

   If the context is strongly ϕ with yielding private views, player 1 pools on $M(h_1)$.

   If the context is strongly ϕ with swaying private views, no player pools on $1 - M(h_1)$. For any history h_i such that all players $k < i$ have either separated or pooled, i pools on $M(h_i)$ if $\Delta(h_i) = 2M(h_i) - 1$.

	Context is …	
	Uninformative	Neither strongly ϕ nor uninformative
$\Delta(h_i) = 2$	Pool on $M(h_i) = 1$	Doesn't arise in equilibrium
$\Delta(h_i) = 1$	Pool on $M(h_i) = 1$ or separate	Pool on $M(h_i) = 1$
$\Delta(h_i) = 0$	Separate	Separate, pool on $M(h_i) = 1$ or semi-separate on $M(h_i) = 1$
$\Delta(h_i) = -1$	Pool on $M(h_i) = 0$ or separate	Separate, pool on $M(h_i)$ or semi-separate on $M(h_i)$
$\Delta(h_i) = -2$	Pool on $M(h_i) = 0$ or separate	Separate, pool on $M(h_i) = 0$ or semi-separate on $M(h_i) = 0$
$\Delta(h_i) = -3$	Pool on $M(h_i) = 0$	Separate, pool on $M(h_i) = 0$ or semi-separate on $M(h_i) = 0$
$\Delta(h_i) \geq -4$	Doesn't arise in equilibrium	Separate, pool on $M(h_i) = 0$ or semi-separate $M(h_i) = 0$ (pools for $\Delta(h_i)$ large enough)

Table 2: Equilibrium strategies according to the public signals and the value of the halfway mark

With Result 8 in hand, I can generalize Result 3.

Result 9. *Suppose that if players semi-separate on $M(h_1)$ in the first period, all player $k > 1$ pools on $M(h_1)$.*

Individuals will be in cooperative pluralistic ignorance with a positive probability for $P(\theta = 1) > p \in (0, 1/2)$ if $I \geq 5$. It will be inefficient for some γ sufficiently small.

The assumption stated in the result allows us to avoid the complications of dealing with histories of play in which players semi-separate.

C Proof of Lemma 2

Proof. Write i's posterior over k's type at period j using the law of iterated expectations:

$$P(\phi_k = b \mid h_j, a_j, \phi_i) = \sum_{\theta \in \{0,1\}} P(\theta \mid h_j, a_j, \phi_i) P(\phi_k = b \mid \theta, h_j, a_j)$$

$P(\phi_k = b \mid \theta, h_j, a_j, \phi_i) = P(\phi_k = b \mid \theta, h_j, a_j)$, since once we condition on the state of the world θ, the only reason to condition on history (h_j, a_j) is to determine i's type-dependent strategy, which is publicly known in equilibrium. We can apply Bayes' rule to obtain

$$\frac{\sum_{\theta \in \{0,1\}} P(\theta, h_j, a_j, \phi_i) P(\phi_k = b \mid \theta, h_j, a_j)}{P(h_j, a_j, \phi_i)}$$

By Bayes' rule again, $P(\theta, h_j, a_j, \phi_i) = P(\theta) P(h_j, a_j, \phi_i \mid \theta)$. By the law of iterated expectations, $P(h_j, a_j, \phi_i) = \sum_{\theta \in \{0,1\}} P(\theta) P(h_j, a_j, \phi_i \mid \theta)$. Since draws are independent, $P(h_j, a_j, \phi_i \mid \theta) = P(h_j, a_j \mid \theta) P(\phi_i \mid \theta)$. This equality again uses the fact that, conditional on θ, private information is irrelevant for calculating the probability of a given history (h_j, a_j). Further, the action of the first player does not depend on others' actions, although others' actions depend on the first players' actions: $P(h_j, a_j \mid \theta) = P(a_1 \mid \theta) P(a_2, ..., a_j \mid a_1, \theta)$. But conditional on θ and a_1, player 2's action does not depend on others' actions. Iterating this argument, we get:

$$\frac{\sum_{\theta \in \{0,1\}} P(\theta) \prod_{l \in \{1,\ldots,j\} \setminus \{i\}} [P(a_l \mid h_l, \theta)] P(\phi_i \mid \theta) P(\phi_k = b \mid \theta, h_j, a_j)}{\sum_{\theta \in \{0,1\}} P(\theta) \prod_{l \in \{1,\ldots,j\} \setminus \{i\}} [P(a_l \mid h_l, \theta)] P(\phi_i \mid \theta)} \tag{10}$$

Now I can write $P(\phi_k = 1 \mid h_j, a_j, \phi_i) > x$ in terms of likelihood ratios:

$$\prod_{l \in \{1,\ldots,j\} \setminus \{i\}} \left(\frac{P(a_l \mid h_l, \theta{=}1)}{P(a_l \mid h_l, \theta{=}0)} \right) \frac{P(\phi_i \mid \theta{=}1)}{P(\phi_i \mid \theta{=}0)} \left[\frac{P(\phi_k{=}1 \mid \theta{=}1, h_j, a_j) - x}{x - P(\phi_k{=}1 \mid \theta{=}0, h_j, a_j)} \right] \lesseqgtr \frac{P(\theta{=}0)}{P(\theta{=}1)} \tag{11}$$

Where $\lesseqgtr$ is greater or equal to ($\geq$) if $x > P(\phi_k{=}b \mid \theta{=}0, h_j, a_j)$, and is less than ($<$) if $x < P(\phi_k{=}b \mid \theta{=}0, h_j, a_j)$.

We will now show that if the numerator or the denominator of the term in brackets is negative, the other term is positive. The second condition of the Lemma implies that the denominator of the term in square brackets is negative. If this implies that the numerator is positive, the left hand side is negative and (11) is satisfied.

Notice that $P(\phi_k = 1 \mid \theta, h_j, a_j)$ is equal to $P(\phi_k = 1 \mid \theta)$ if $k > j$ and equal to

$P(\phi_k = 1 \mid \theta, h_k, a_k)$ otherwise, again using the fact that once we condition on θ, the history of play is only relevant to determine the optimal type-dependent strategy of player k, $\sigma_k^*(a_k \mid \phi_k, h_k)$. Further, use Bayes' rule to set

$$P(\phi_k = 1, h_j, a_j, \theta) = P(\theta)P(\phi_k = 1, h_j, a_j \mid \theta)$$

which is equal to $P(\theta)P(\phi_k = 1 \mid \theta)\sigma_k^*(a_k \mid \phi_k = 1, h_k)$. We can therefore set

$$P(\phi_k = 1 \mid \theta, h_j, a_j) = \frac{P(\phi_k = 1 \mid \theta)\sigma_k^*(a_k \mid \phi_k = 1, h_k)}{P(\phi_k = 1 \mid \theta)\sigma_k^*(a_k \mid \phi_k = 1, h_k) + P(\phi_k = 0 \mid \theta)\sigma_k^*(a_k \mid \phi_k = 0, h_k)} \tag{12}$$

Now consider the term

$$\left[\frac{P(\phi_k = 1 \mid \theta = 1, h_j, a_j) - x}{x - P(\phi_k = 1 \mid \theta = 0, h_j, a_j)}\right] \tag{13}$$

The numerator is weakly negative if and only if

$$P(\phi_k = 1 \mid \theta = 1)\frac{(1-x)\sigma(a_k \mid \phi_k = 1, h_j)}{x\sigma(a_k \mid \phi_k = 0, h_k)} \leq P(\phi_k = 0 \mid \theta = 1) \Rightarrow \frac{\sigma(a_k \mid \phi_k = 1, h_k)}{\sigma(a_k \mid \phi_k = 0, h_k)} < \frac{x}{1-x}$$

The denominator is weakly negative if and only if

$$P(\phi_k = 0 \mid \theta = 0)\frac{x\sigma(a_k \mid \phi_k = 0, h_k)}{(1-x)\sigma(a_k \mid \phi_k = 1, h_k)} \leq P(\phi_k = 1 \mid \theta = 0) \Rightarrow \frac{\sigma(a_k \mid \phi_k = 1, h_k)}{\sigma(a_k \mid \phi_k = 0, h_k)} > \frac{x}{1-x} \tag{14}$$

Therefore, by symmetry of signal informativeness, if the numerator or the denominator is negative, the other term is positive.

Another way to reach this conclusion is by noting that

$$\frac{\partial P(\phi_k \mid \theta, h_j, a_j)}{\partial P(\phi_k \mid \theta)} = \frac{\sigma_k^*(a_k \mid \phi_k = 1, h_k)\sigma_k^*(a_k \mid \phi_k = 0, h_k)}{[P(\phi_k = 1 \mid \theta)\sigma_k^*(a_k \mid \phi_k = 1, h_k) + P(\phi_k = 0 \mid \theta)\sigma_k^*(a_k \mid \phi_k = 0, h_k)]^2} > 0$$

Since $P(\phi_k = \theta \mid \theta) > P(\phi_k \neq \theta \mid \theta)$,

$$P(\phi_k = \theta \mid \theta, h_j, a_j) > P(\phi_k \neq \theta \mid \theta, h_j, a_j) \tag{15}$$

Therefore, if $x > P(\phi_k = \theta \mid \theta, h_j, a_j)$, the numerator of (13) is negative and the denominator is positive; if $x \in (P(\phi_k = \theta \mid \theta, h_j, a_j), P(\phi_k \neq \theta \mid \theta, h_j, a_j))$, the numerator and denominator of are positive; if $P(\phi_k \neq \theta \mid \theta, h_j, a_j) < x$, the numerator is positive and the denominator is negative.

Noting that $P(\phi^m_{-i} = x \mid h_i, a_i, \phi_j, \theta) = \sum_{k \in d_{-i}} P(\phi^{obs}_{-i} = \phi^{obs}_k \mid h_i, \phi_j) P(\phi_k = x \mid h_i, \phi_j, \theta)$, I can use (10) to write $P(\phi^m_{-i} = x \mid h_i, a_i, \phi_j)$ as

$$\frac{\sum_{\theta \in \{0,1\}} P(\theta) \prod_{l \in \{1,\ldots,i\} \setminus \{j\}} [P(a_l \mid h_l, \theta)] P(\phi_j \mid \theta) P(\phi^m_{-i} = x \mid h_i, a_i, \phi_j, \theta)}{\sum_{\theta \in \{0,1\}} P(\theta) \prod_{l \in \{1,\ldots,i\} \setminus \{j\}} [P(a_l \mid h_l, \theta)] P(\phi_j \mid \theta)}$$

Let us now write $P(\phi^m_{-i} = 1 \mid h_i, a_i, \phi_j, \theta) > x$ in terms of likelihood ratios

$$\prod_{l \in \{1,\ldots,i\} \setminus \{j\}} \left(\frac{P(a_l \mid h_l, \theta{=}1)}{P(a_l \mid h_l, \theta{=}0)} \right) \frac{P(\phi_j \mid \theta{=}1)}{P(\phi_j \mid \theta{=}0)} \left\{ \frac{P(\phi^m_{-i} = 1 \mid h_i, a_i, \phi_j, \theta = 1) - x}{x - P(\phi^m_{-i} = 1 \mid h_i, a_i, \phi_j, \theta = 0))} \right\} \lesseqgtr \frac{P(\theta{=}0)}{P(\theta{=}1)}$$

Where $\lesseqgtr$ is greater or equal to ('$\geq$') if $x > P(\phi^m_{-i} = 1 \mid h_i, a_i, \phi_j, \theta = 0)$, and is less than ('$<$') if $x < P(\phi^m_{-i} = 1 \mid h_i, a_i, \phi_j, \theta = 0)$

By (15), it must be that

$$P(\phi^m_{-i} = 1 \mid h_i, a_i, \phi_j, \theta = 1) > P(\phi^m_{-i} = 1 \mid h_i, a_i, \phi_j, \theta = 0) \tag{16}$$

Therefore, if $x > P(\phi^m_{-i} = 1 \mid h_i, a_i, \theta = 1)$, the numerator of the term in curly brackets is negative and the denominator is positive. Therefore, $\lesseqgtr$ in (16) is greater or equal to ('$\geq$') and the inequality does not hold.

If $x \in (P(\phi^m_{-i} = 1 \mid h_i, a_i, \phi_j, \theta = 1), P(\phi^m_{-i} = 0 \mid h_i, a_i, \phi_j, \theta = 0))$, the numerator and denominator of are positive. Therefore, $\lesseqgtr$ is greater or equal to ('$\geq$'), and the inequality may or may not hold.

If $P(\phi^m_{-i} = 0 \mid h_i, a_i, \phi_j, \theta = 0) < x$, the numerator is positive and the denominator is negative. Therefore, $\lesseqgtr$ is less than ('$<$'), and the inequality holds. □

D Proof of Lemma 3

Proof.

1. Since $\beta > 1$, for ϕ_i to be indifferent between actions, it must be that

$$\mathbb{E}\left(\mathbb{E}\left(\mathcal{J}_{j,i}(h_i, a_i{=}\phi_i, \phi_j) \mid \hat{\phi}_{-i}\right) \mid h_i, \phi_i\right) < \mathbb{E}\left(\mathbb{E}\left(\mathcal{J}_{j,i}(h_i, a_i{=}1 - \phi_i, \phi_j) \mid \hat{\phi}_{-i}\right) \mid h_i, \phi_i\right)$$

   But we know that ϕ_i thinks the expected judgment from choosing $a_i = \phi_i$ is weakly greater than does $1 - \phi_i$

$$\mathbb{E}\left(\mathbb{E}\left(\mathcal{J}_{j,i}(h_i, a_i{=}\phi_i, \phi_j) \mid \hat{\phi}_{-i}\right) \mid h_i, \phi_i\right) \geq \mathbb{E}\left(\mathbb{E}\left(\mathcal{J}_{j,i}(h_i, a_i{=}\phi_i, \phi_j) \mid \hat{\phi}_{-i}\right) \mid h_i, 1 - \phi_i\right)$$

This follows from the fact that, as shown by Lemma 2, judges are relatively biased towards their type, and decision makers think there are relatively more judges of the same type. But this leads to contradiction:

$$\mathbb{E}\left(\mathbb{E}\left(\mathcal{J}_{j,i}(h_i, a_i{=}1-\phi_i, \phi_j) \mid \hat{\phi}_{-i}\right) \mid h_i, \phi_i\right) > \mathbb{E}\left(\mathbb{E}\left(\mathcal{J}_{j,i}(h_i, a_i{=}\phi_i, \phi_j) \mid \hat{\phi}_{-i}\right) \mid h_i, \phi_i\right) \geq$$

$$\mathbb{E}\left(\mathbb{E}\left(\mathcal{J}_{j,i}(h_i, a_i{=}\phi_i, \phi_j) \mid \hat{\phi}_{-i}\right) \mid h_i, 1-\phi_i\right) > \mathbb{E}\left(\mathbb{E}\left(\mathcal{J}_{j,i}(h_i, a_i{=}1-\phi_i, \phi_j) \mid \hat{\phi}_{-i}\right) \mid h_i, 1-\phi_i\right) \geq$$

$$\mathbb{E}\left(\mathbb{E}\left(\mathcal{J}_{j,i}(h_i, a_i{=}1-\phi_i, \phi_j) \mid \hat{\phi}_{-i}\right) \mid h_i, \phi_i\right) \square$$

2. We will now show that $\tilde{y}(\phi_k^{obs})$ exists as in the statement of the Lemma. Consider $P(\phi_k = 1 \mid h_i, a_i, \theta)$ as it is written out in (12) in the proof of Lemma 2. If k semi-separates on one, this term strictly increases in $\phi_k^{obs} = \sigma_k^*(a_k \mid \phi_k = 1, h_k) \in (0,1)$ and takes a value between $P(\phi_k = 1 \mid \theta)$ and 1. If k semi-separates on zero, the term strictly increases in $\phi_k^{obs} = -\sigma_k^*(a_k \mid \phi_k = 0, h_k) \in (-1,0)$ and takes a value between 0 and $P(\phi_k = 1 \mid \theta)$. If ϕ_k^{obs} is equal to -1, 0 and 1, $P(\phi_k = 1 \mid h_i, a_i, \theta)$ is respectively equal to 0, $P(\theta_k = 1 \mid \theta)$ and 1. Therefore, a function $\tilde{y}(\phi_i^{obs})$ exists as specified in the Lemma.□

3. We now show that in equilibrium there exists a unique $\tilde{x}(\phi_i^{obs}) \in [-1,1]$ such that,

$$\frac{P(a_i \mid h_i, \theta{=}1)}{P(a_i \mid h_i, \theta{=}0)} = \left(\frac{P(\phi_i = 1 \mid h_i, \theta = 1)}{P(\phi_i = 1 \mid h_i, \theta = 0)}\right)^{\tilde{x}}$$

We can use the law of iterated expectations to expand the coefficients:

$$\frac{P(a_i \mid h_i, \theta{=}1)}{P(a_i \mid h_i, \theta{=}0)} = \sum_{\phi_i \in \{0,1\}} \frac{\sigma(a_i \mid h_i, \phi_i) P(\phi_i \mid h_i, \theta = 1)}{\sigma(a_i \mid h_i, \phi_i) P(\phi_i \mid h_i, \theta = 0)}$$

which uses the fact that $\sigma(a_i \mid h_i, \phi_i, \theta) = \sigma(a_i \mid h_i, \phi_i)$.

It is easy to check that the result holds for $\phi_i^{obs} \in \{-1, 0, 1\}$ if I set $\tilde{x}(-1) = -1$, $\tilde{x}(0) = 0$ and $\tilde{x}(1) = 1$.

If $\sigma(a_i \mid h_i, \phi_i) = 1$ and $\sigma(a_i \mid h_i, \phi_i) = \phi_i^{obs} \in [0,1]$, the expression reduces to

$$\frac{\phi_i^{obs} P(\phi_i = 1 \mid h_i, \theta = 1) + P(\phi_i = 0 \mid h_i, \theta = 1)}{\phi_i^{obs} P(\phi_i = 1 \mid h_i, \theta = 0) + P(\phi_i = 0 \mid h_i, \theta = 0)}$$

This expression is strictly increasing in ϕ_i^{obs} and has range $[0,1]$, so we can find a $\tilde{x}(\phi_i^{obs}) \in [0,1]$ such that the result holds. The result for $\tilde{x}(\phi_i^{obs}) \in [-1,0]$ follows from a similar reasoning by the symmetry of the problem.□

E Proof of Lemma 4

1. If players k and k' separate, they have no private information about the state of the world or group members' realized preferences. Therefore, they have the same information about f and g – the public information about types given the history of play and priors. If l is of type 0 (1) and does not separate, her private information biases her beliefs over f and g downwards (upwards) with respect to k and k'. As $\phi_l^{obs} \in [-1,1]$ increases, players other than l assign a lower probability to l being of type 0, assigning a probability 1 if $\phi_l^{obs} = -1$ and probability 0 if $\phi_l^{obs} = 1$; this follows from Lemma 3. Therefore, if $\phi_l^{obs} > \phi_{l'}^{obs}$, what $-l$ learned from observing ϕ_l^{obs} led them to update f and g upwards more than what $-l'$ learned from observing $\phi_{l'}^{obs}$. Therefore, $e \in \{f, g\}$ will be strictly lower for l than for l' if $e(h_i, i, \phi_i^{obs}, \phi_l^{obs}, \phi_l, z; I) \in (-\infty, \infty)$, and otherwise l and l' have e equal to ∞ or ∞^+.

2. $\partial P(\phi_i = 1 \mid h_i, a_i, \phi_j)/\partial \phi_i^{obs} > 0$ and $\partial \tilde{x}(\phi_k^{obs})/\partial \phi_k^{obs} > 0$, as shown in Lemma 3.

   Therefore, f increases with ϕ_k^{obs} unless $f = \infty^+$, in which case f does not change. If $\phi_i^{obs} = 1$, $P(\phi_i = 1 \mid h_i, a_i, \theta) = 1$, which implies $f = \infty^+$. If $\phi_i^{obs} = -1$, $P(\phi_i = 1 \mid h_i, a_i, \theta) = 0$, which implies $f < 0$.

3. Since $\partial \tilde{x}(\phi_i^{obs})/\partial \phi_k^{obs} > 0$, as ϕ_k^{obs} increases, the probability that players other than i who have not separated are of type 1 increases. Further, $P(\phi_{-i}^m = 1 \mid h_i, \phi_j, \theta) = \sum_{k \in d_{-i}} P(\phi_{-i}^{obs} = \phi_k^{obs} \mid h_i, \phi_j) P(\phi_k = 1 \mid h_i, \phi_j, \theta) =$

   $$\frac{1}{I} \sum_{k \neq i | \phi_k^{obs} \geq 0} \left[(1 - \tilde{y}(\phi_k^{obs}, \theta)) P(\phi_k = 1 \mid \theta) + \tilde{y}(\phi_k^{obs}, \theta)\right] + \frac{1}{I} \sum_{k \neq i | \phi_k^{obs} < 0} \left[(1 + \tilde{y}(\phi_k^{obs}, \theta)) P(\phi_k = 1 \mid \theta)\right]$$

   Take some individual $l < i$ with class of type ϕ_l^{obs}. Consider an increase in ϕ_l^{obs}. If $\phi_l^{obs} > 0$, then the derivative of $P(\phi_{-i}^m = 1 \mid h_i, \phi_j, \theta)$ with respect to ϕ_l^{obs} is $\tilde{y}'(\phi_l^{obs}, \theta)(1 - P(\phi_l \mid \theta))/I > 0$. If $\phi_l^{obs} < 0$, the derivative is $\tilde{y}'(\phi_l^{obs}, \theta)/I > 0$. At $\phi_l^{obs} = 0$, the left derivative of $P(\phi_{-i}^m = 1 \mid h_i, \phi_j, \theta)$ with respect to ϕ_l^{obs} is $\tilde{y}'(\phi_l^{obs}, \theta)(1 - P(\phi_l \mid \theta))/I > 0$, and the right derivative is $\tilde{y}'(\phi_l^{obs}, \theta)/I > 0$. An increase in $P(\phi_{-i}^m = 1 \mid h_i, \phi_j, \theta)$ increases the third term in $\tilde{g}$. Therefore, the majority preference of players other than i is more likely to be of type 1.

   Since ϕ_i^{obs} affects beliefs about others' preference through $\tilde{x}(\phi_i^{obs})$, the third term in $\tilde{g}$ is unaffected by ϕ_i^{obs}. Therefore, if the third term term is in the set $X \in$

$\{(-\infty, 0], (0, \infty), \infty, \infty^+\}$, a change in $\tilde{x}(\phi_i^{obs})$ will yield a value of g in X.

4. $\Delta(h_i) + \mathbb{1}\{\phi_j^{obs}(h_i) = 0\}(2\phi_j - 1)$ captures the sum of signals judge j observes, with $\Delta(h_i) = \sum_{k \leq i} \tilde{x}(\phi_k^{obs}) = \sum_{k \leq i} \tilde{y}(\phi_k^{obs}, \theta)$ since players have either revealed or withheld at history h_i. If $\Delta(h_i) + \mathbb{1}\{\phi_j^{obs}(h_i) = 0\}(2\phi_j - 1) = 0$, then j observes the same number of 0 and 1 signals. But then f is equal to one: the first two terms of $\tilde{f}$ are equal to one, as is the third term (since $P(\phi_i \mid h_i, a_i\theta) = P(\phi_i \mid \theta)$, $f(h_i, i, \phi_i^{obs}, \phi_j^{obs}, \phi_j, 1/2; I)$ is equal to $\tilde{f}$). g is also equal to one: we know the first two terms of $\tilde{g}$ are equal to one, so I only need to show that the third term is also equal to one.

   Let $\sum_{k \neq i | \phi_k^{obs} > 0} \tilde{y}(\phi_k^{obs}, \theta) \equiv \tilde{y}_1$ be the sum of one-signals and $\sum_{k \neq i | \phi_k^{obs} < 0} \tilde{y}(\phi_k^{obs}, \theta) \equiv \tilde{y}_0$ be the sum of zero-signals. Rewrite $P(\phi_{-i}^m = 1 \mid h_i, a_i, \phi_j, \theta)$ as

$$\left[1 - \frac{\tilde{y}_1}{I} - \frac{\tilde{y}_0}{I}\right] P(\phi_k = 1 \mid \theta) + \frac{\tilde{y}_1}{I} \tag{17}$$

   Define $N \equiv \tilde{y}_1 + \tilde{y}_0 - |\tilde{y}_1 - \tilde{y}_0|$ as the mass of balanced signals. Let

$$y_1 \equiv \begin{cases} \tilde{y}_1 - \tilde{y}_0 & \text{if } \tilde{y}_1 > \tilde{y}_0 \\ 0 & \text{otherwise} \end{cases}$$

$$y_0 \equiv \begin{cases} \tilde{y}_0 - \tilde{y}_1 & \text{if } \tilde{y}_0 > \tilde{y}_1 \\ 0 & \text{otherwise} \end{cases}$$

   y_1 and y_0 capture the excess signals. If the sum of public one-signals is greater than the sum of public zero-signals, $y_1 > 0$ and $y_0 = 0$. If the sum of public zero-signals is greater, $y_1 = 0$ and $y_0 > 0$. Then I can rewrite (17) as

$$\frac{I - N - y_1 - y_0}{I} P(\phi_k = 1 \mid \theta) + \frac{N}{2I} + \frac{y_1}{I} \tag{18}$$

   If at period i, j observes $N/2$ zero- and one-signals,

$$\frac{I - N}{I} P(\phi_k = 1 \mid \theta = 1) + \frac{N}{2I} - \frac{1}{2} =$$

$$\frac{2(I - N)}{2I} - \frac{(I - N)}{I} P(\phi_k = 1 \mid \theta = 0) + \frac{N}{2I} - \frac{I}{2I} =$$

$$\frac{1}{2} - \frac{N}{2I} - \frac{I - N}{I} P(\phi_k = 1 \mid \theta = 0) \Rightarrow$$

$$\left\{ \frac{\frac{I-N}{I}P(\phi = 1 \mid \theta = 1) + \frac{N}{2I} - \frac{1}{2}}{\frac{1}{2} - \frac{N}{2I} - \frac{I-N}{I}P(\phi = \mid \theta = 0)} \right\} = 1$$

The rest of the claim follows from the following facts which I have shown:

(a) The first term of $\tilde{f}$ and $\tilde{g}$ increase in $\Delta(h_i) = \sum_{k \leq i} \tilde{x}(\phi_k^{obs})$ (Lemma 3).

(b) The third term of $\tilde{f}$ is equal to one when $\phi_i^{obs} = 0$ (proof of point 4 of this Lemma)

(c) The third term of $\tilde{g}$ is greater, equal or less than one if $\Delta(h_i) = \sum_{k<i} \tilde{y}(\phi_k^{obs})$ is greater, equal or less than zero (point 3 of this Lemma)

5. For a fixed $\Delta(h)$, i and I do not affect $f(h_i, i, 0, \phi_j^{obs}, \phi_j, 1/2; I)$, which only depends on h through $\sum_{k<i} \tilde{x}(\phi_k^{obs})$. Therefore, I only need to look at how $g(h_i, i, 0, \phi_j^{obs}, \phi_j, 1/2; I)$ changes with I and i. Since the first two terms on $\tilde{g}$ are also not affected by i and I (for the same reason), we may focus on the third term.

   First notice that, for a fixed $\Delta(h)$, an increase in i increases N in (18). Now use (18) to rewrite $\tilde{g}$ with $z' = 1/2$ as

$$\frac{(I - N - y_1)(P(\phi = 1 \mid \theta = 1) - P(\phi = 1 \mid \theta = 0)) + y_1 - 2P(\phi = 1 \mid \theta = 1)y_0}{(I - N - y_1)(P(\phi = 1 \mid \theta = 1) - P(\phi = 1 \mid \theta = 0)) + 2P(\phi = 1 \mid \theta = 0)y_0 - y_1}$$

   The derivative with respect to $I - N$ is

$$\frac{2[P(\phi_k = 1 \mid \theta = 1) - P(\phi_k = 1 \mid \theta = 0)][y_0 - y_1]}{[(I - N - y_1)(P(\phi = 1 \mid \theta = 1) - P(\phi = 1 \mid \theta = 0)) + 2P(\phi = 1 \mid \theta = 0)y_0 - y_1]^2}$$

   It is easy to see that the derivative of this term with respect to y_1 is negative, and the derivative of this term with respect to y_0 is positive.□

6. Neither i nor revealer judges have any private information about anybody's preference in the group without 1. When i reveals her type, revealer judges have the same information about the state of the world as i.

7. If $\phi_i^{obs} \in \{-1, 1\}$, then $\sum_{k<i} \tilde{x}(\phi_k^{obs}) = \sum_{l<i+1} \tilde{x}(\phi_l^{obs}) + \phi_i^{obs}$. $f(h_i, i, 0, \phi_j^{obs}, \phi_j, z; I)$ and $f(h_{i+1}, i+1, 0, \phi_k^{obs}, \phi_k, z; I)$ are respectively equal to $\tilde{f}(h_i, i, 0, \phi_j^{obs}, \phi_j, z; I)$ and $\tilde{f}(h_{i+1}, i+1, 0, \phi_k^{obs}, \phi_k, z; I)$ (proof of point 4 of this Lemma) and only depend on history through the first terms. But then $f(h_i, i, 0, \phi_j^{obs}, \phi_j, z; I) = f(h_{i+1}, i+1, 0, \phi_k^{obs}, \phi_k, z; I)$ if and only if j and k observe the same sum of signals.

8. By (a), the first and second terms of $\tilde{f}(h_i, i, 0, \phi_j^{obs}\phi_j, z'; I)$ are equal to the first and second terms of $\tilde{g}(h_i, i, \phi_i^{obs}, \phi_k^{obs}, \phi_k, z'; I)$ – j and k have the same information about

the state of the world. $f(h_i, i, 0, \phi_l^{obs}\phi_l, z'; I)$, $l \in \{j, k\}$ is equal to $\tilde{f}(h_i, i, 0, \phi_l^{obs}\phi_l, z'; I)$ with the third term equal to one (point 4 of this Lemma). If $g(h_i, i, 0, \phi_k^{obs}, \phi_k, z'; I)$ is greater, equal or less than $f(h_i, i, 0, \phi_k^{obs}, \phi_k, z'; I)$, it must be that the third term of $\tilde{g}(h_i, i, 0, +\phi_i^{obs}, \phi_k^{obs}, \phi_k, z'; I)$ is respectively greater, equal or less than one. □

F Proof of Result 8

Begin with the following:

Lemma 5. *1. i of type ϕ chooses ϕ after history h_i ($\alpha_i^*(\phi, h_i) = \phi$) if she believes that by doing so, at least half the judges would believe preferences match:*

$$\mathbb{E}\left(\mathbb{E}\left(\mathcal{J}_{j,i}(h_i, a_i{=}\phi, \phi_j) \mid \hat{\phi}_{-i}\right) \mid h_i, \phi_i{=}\phi\right) \geq \frac{1}{2}$$

2. i pools on ϕ ($\alpha_i^*(\phi, h_i) = \alpha_i^*(1-\phi, h_i) = 1-\phi$) if when she chooses ϕ no judges believe preferences match, but when she chooses $1-\phi$ all judges do:

$$\mathbb{E}\left(\mathbb{E}\left(\mathcal{J}_{j,i}(h_i, a_i{=}\phi, \phi_j) \mid \hat{\phi}_{-i}\right) \mid h_i, \phi_i\right) = 1, \mathbb{E}\left(\mathbb{E}\left(\mathcal{J}_{j,i}(h_i, a_i{=}1-\phi, \phi_j) \mid \hat{\phi}_{-i}\right) \mid h_i, \phi_i\right) = 0$$

3. Semi-separating on $1 - M(h_i)$ is not an equilibrium strategy.

4. If revealer judges believe preferences match when i reveals $1 - M(h_i)$, separating is an equilibrium strategy. If furthermore all judges believe preferences match if i pools, separating is the unique strategy that could be sustained in a Perfect Bayesian Equilibrium with a D1 refinement.

5. Suppose that at history h_i, all players $k < i$ have either separated or pooled. If revealer judges believe preferences do not match when i reveals $1 - M(h_i)$, then pooling on $1 - M(h_i)$ is not an equilibrium strategy. Suppose further that all judges believe preferences match when i reveals $M(h_i)$. Then if i separates, is the unique strategy that could be sustained in a Perfect Bayesian Equilibrium with a D1 refinement.

6. Suppose that if i pools at history h_i, all judges believe preferences match. Then if i of type $1 - M(h_i)$ deviates from a separating strategy, i pools on $M(h_i)$ in equilibrium ($\alpha_i^*(M(h_i), h_i) = \alpha_i^*(1 - M(h_i), h_i) = M(h_i)$). If i of type $1 - M(h_i)$ deviates from a pooling strategy, i separates in equilibrium ($\alpha_i^*(\phi, h_i) = \phi$).

Proof.

1. Individual i of type ϕ chooses $a_i = \phi$ if and only if

   $$\mathbb{E}\left(\mathbb{E}\left(\mathcal{J}_{j,i}(h_i, a_i{=}1-\phi, \phi_j) \mid \hat{\phi}_{-i}\right) \mid h_i, \phi_i{=}\phi\right) - \mathbb{E}\left(\mathbb{E}\left(\mathcal{J}_{j,i}(h_i, a_i{=}\phi, \phi_j) \mid \hat{\phi}_{-i}\right) \mid h_i, \phi_i{=}\phi\right) < \frac{1}{\beta} \tag{19}$$

   with $1/\beta \in (1/2, 1)$. But if $\mathbb{E}\left(\mathbb{E}\left(\mathcal{J}_{j,i}(h_i, a_i{=}\phi, \phi_j) \mid \hat{\phi}_{-i}\right) \mid h_i, \phi_i{=}\phi\right) \geq 1/2$, the left hand side of (19) is less than $1/2$, so the right hand side is greater than the left hand side.

2. If $\mathbb{E}\left(\mathbb{E}\left(\mathcal{J}_{j,i}(h_i, a_i, \phi_j) \mid \hat{\phi}_{-i}\right) \mid h_i, \phi_i{=}\phi\right)$ is 1 if $a_i = 1 - \phi$ and 0 if $a_i = \phi$, the left hand side of (19) is equal to one, which is greater than the right hand side. □

3. i of type $M(h_i)$ would deviate from a strategy of semi-separating on $1 - M(h_i)$. Indeed, if i chooses $M(h_i)$, type $M(h_i)$ believes at least half of judges believe preferences match (point 6 of Lemma 4 and point 1 of this Lemma).

4. If revealer judges believe preferences match when i reveals $1 - M(h_i)$, then we can use point 1 of this Lemma and point 6 of Lemma 4 to conclude that there is an equilibrium separating strategy.

   If furthermore all judges believe preferences match if i pools on ϕ, then i of type ϕ will not deviate from ϕ for any out-of-equilibrium beliefs. But then we can use D1 to conclude that judges will believe $1 - \phi$ chose the out-of-equilibrium action $1 - \phi$. Therefore, if i pools or semi-separates on ϕ, i of type $1 - \phi$ would reveal $1 - \phi$ if she chose $1 - \phi$. We can then use point 6 of Lemma 4 and point 1 of this Lemma again to conclude that there will always be a deviation from a non-separating strategy.

5. If revealer judges believe preferences do not match when i reveals $1 - M(h_i)$, then both types of player i believe that most judges are of type 1 (point 6 of Lemma 4).

   Suppose i pools on $1 - M(h_i)$. If i of type $1 - M(h_i)$ would not deviate from pooling, we can use D1 to conclude that judges put full weight on $M(h_i)$ types deviating, which would lead type $M(h_i)$ to deviate. Therefore, suppose that it is not the case that both revealer and $(0, M(h_i))$ judges believe preferences match when i pools. i of type $1 - M(h_i)$ would not deviate to $M(h_i)$ if only judges of type $1 - M(h_i)$ believed preferences match: $1 > \beta P(\phi_{-i} = 1 - M(h_i) \mid h_i, \phi_i = 1 - M(h_i))$ for $\beta \in (1, 2)$. Therefore, for i of type $1 - M(h_i)$ to deviate, type $M(h_i)$ or revealer judges must believe preferences match.

   We know by point 1 of Lemma 4 that $g(h_i, i, \phi_i^{obs}, \phi_j^{obs}, \phi_j, 1/2; I)$ is greater than $\mathcal{H}(\theta)$ if $M(h_i) = 1$ and less than if $M(h_i) = 0$ for all ϕ_i^{obs} if j is a revealer or type $(0, M(h_i))$

judge. Further, we know that $f(h_i, i, \phi_i^{obs}, 0, 0, 1/2; I) < f(h_i, i, \phi_i^{obs}, \phi_k^{obs}, \phi_k, 1/2; I) < f(h_i, i, \phi_i^{obs}, 0, 1, 1/2; I)$ for any ϕ_i^{obs} for a revealer judge k. But then it follows that for any out-of-equilibrium beliefs in which type $(0, M(h_i))$ judges believe preferences don't match, revealer judges believe preferences don't match. Conversely, if revealer judges believe preferences match, $(0, M(h_i))$ judges believe preferences match. Further, for pooling on $1 - M(h_i)$ to be an equilibrium strategy, it must be that type $M(h_i)$ believes more than half of judges will believe preferences match. But then it must be that $(0, M(h_i))$ judges believe preferences match, and that revealer judges believe preferences do not match.

For type $1 - M(h_i)$ to want to deviate, the out-of-equilibrium beliefs must have type $(0, M(h_i))$ judges believe preferences match (if they do not, then revealer judges will not believe preferences match). Therefore, i of type $M(h_i)$ will have weakly higher social expectations for all out-of-equilibrium beliefs in which type $1 - M(h_i)$ deviates. But then we can then use D1 to rule out type $1 - M(h_i)$ from deviating if i pools on $1 - M(h_i)$. But then with refined beliefs, i of type $M(h_i)$ would always deviate from $1 - M(h_i)$ to choose $M(h_i)$ and reveal $M(h_i)$ (point 1 of this Lemma).

If we suppose further that all judges believe preferences match when i reveals $M(h_i)$, then if i separates, it is the unique equilibrium strategy. If i pools on $M(h_i)$, we can use D1 to conclude judges believe type $1 - M(h_i)$ would choose $1 - M(h_i)$ (point 1 of this Lemma). Then if i of type $1 - M(h_i)$ does not deviate from revealing, then she would choose $1 - M(h_i)$ given any social expectations of choosing $M(h_i)$, and therefore would never deviate.

6. Suppose i pools on ϕ. Then type ϕ would never want to deviate. But then we can apply D1 to rule ϕ out from choosing the out-of-equilibrium action $1 - \phi$. By point 1 of Lemma 5, i of type $M(h_i)$ would then deviate from a strategy of pooling on $1 - M(h_i)$. Indeed, if i chooses $M(h_i)$, type $M(h_i)$ believes at least half of judges believe preferences match (point 6 of Lemma 4 and point 1 of this Lemma).

   Whether i is separating or pooling on $M(h_i)$, the social expectation i of type $1-M(h_i)$ gets from choosing $1 - M(h_i)$ is the same – in both cases type $1 - M(h_i)$ is revealed. The social expectation i of type $1 - M(h_i)$ gets from choosing $M(h_i)$ and revealing $M(h_i)$ is weakly lower than from pooling on $M(h_i)$ and having all judges believing preferences match. Therefore, if i of type $1 - M(h_i)$ deviates from separating, she will not deviate from pooling on $M(h_i)$. Conversely, if i of type $1 - M(h_i)$ deviates from pooling on $M(h_i)$, she will not deviate from separating. □

I now show each point from Result 8.

1. Incentives to separate increase with β since the right hand side of (19) increases as β decreases, while the left hand side is unaffected.

   As $P(\theta = M(h_i))$ increases, $\mathcal{H}(\theta)$ increases if $M(h_i) = 0$ and decreases if $M(h_i) = 1$, while the values of f and g are unaffected for all players. Therefore, there are weakly more full type judges who believe preferences match, and type $1-M(h_i)$ judges believe judges are more likely to be of type $M(h_i)$ (point 6 of Lemma 4). This is sufficient to establish that incentives to separate decrease, since type $M(h_i)$ would not deviate from $M(h_i)$ (points 3, 4 and 5 of Lemma 5).

   I will show the impact of I and i on incentives to separate below, as I analyze social expectations for different values of the action lead $\Delta(h_i)$.

2. If i semi-separates on ϕ, it must be that not all judges would have believed preferences match if i had pooled (point 6 of Lemma 5), and that revealer judges believe preferences do not match when i reveals $1 - M(h_i)$ (point 4 of Lemma 5). By point 3 of Lemma 5, we know that if i semi-separates, she must semi-separate on $M(h_i)$. Without loss of generality, suppose $M(h_i) = 1$.

   When i semi-separates on 1 and chooses 1, or if i separates and chooses 1, the values of $f(h_k, k, \phi_k^{obs}, \phi_j^{obs}, \phi_j, z; I)$ and $g(h_k, k, \phi_k^{obs}, \phi_j^{obs}, \phi_j, z'; I)$ are larger for $k = i+1$ than for $k = i$ for all judges other than i, and for all values of ϕ_k^{obs} (points 2 and 3 of Lemma 4). Then if $i+1$ reveals $1 - M(h_i)$, weakly fewer full type judges (who are all of type $1 - M(h_i)$) will believe preferences match, and $i+1$ of type $1 - M(h_i)$ believes judges are less likely to be of type $1 - M(h_i)$ (point 6 of Lemma 4). Conversely, if $i+1$ reveals $M(h_i)$, weakly more full type judges believe preferences match (which includes all revealer and type $M(h_i)$ judges), and $i+1$ of type $1 - M(h_i)$ believes judges are more likely to be of type $M(h_i)$. Therefore, incentives to separate decrease. Since i did not separate in equilibrium, $i+1$ will not separate in equilibrium. We know by point 3 of Lemma 5 that if separating is not an equilibrium strategy for $i+1$, there will be no semi-separating on $1 - M(h_i)$ at period $i+1$.

   If instead i chooses 0, $M(h_{i+1}) \in \{0, 1\}$. If $M(h_{i+1}) = 1$, the impact on incentives to separate is the opposite as above, so incentives to separate increase. If $M(h_{i+1}) = 0$, then either separating is an equilibrium strategy, or only semi-separating on 0 or pooling are equilibrium strategies (point 1 of Lemma 5).

   Suppose i finds it optimal to pool on ϕ. Then $i+1$ will have the same information as i did, so will face the same incentives. Since pooling must be the unique equilibrium pure strategy for i to pool, it will be the equilibrium pure strategy for $i+1$.

3. We now analyze social expectations for different values of $\Delta(h_i)$. We will use points 4, 7 and 8 of Lemma 4 to bound the values of $\hat{f}$ and $\hat{g}$, point 6 of Lemma 4 to calculate i's social expectation, and point 5 of Lemma 4 to consider comparative statics of i and I. The analysis will focus on the case where $M(h_1) = 0$, or $\mathcal{H}(\theta) > 1$. By symmetry, a similar analysis can be made for $\mathcal{H}(\theta) < 1$.

- $\Delta(h_i) = 0$. If i pools,

  $$\hat{g}(0, i, 0, 0, 0) < \hat{f}(0, i, 0, 0, 0) < \hat{g}(0, i, 0, 1, 1) =$$

  $$1 = \hat{f}(0, i, 0, 1, 1) < \hat{f}(0, i, 0, 0, 1) < \hat{g}(0, i, 0, 0, 1)$$

  Recall that the values of $\hat{f}$ and $\hat{g}$ are the same for revealers of type 1 $(1, 1)$ and revealers of type 0 $(-1, 0)$. If player i reveals to be of type 1, then $\hat{f}(0, i, 1, \phi_j^{obs}, \phi_j) = \infty^+$ for all j, and

  $$\hat{g}(0, i, 0, 0, 0) < \hat{g}(0, i, 1, 0, 0) < 1 < \hat{g}(0, i, 1, 1, 1) = \hat{f}(0, i, 0, 0, 1) < \hat{g}(0, i, 1, 0, 1)$$

  If player i reveals to be of type 0, then $\hat{f}(0, i, -1, \phi_j^{obs}, \phi_j) < 0$ for all j, and

  $$\hat{g}(0, i, -1, 0, 0) < \hat{g}(0, i, -1, 1, 1) = \hat{f}(0, i, 0, 0, 0) < 1 < \hat{g}(0, i, -1, 0, 1)$$

  If $\mathcal{H}(\theta) \in (1, \hat{f}(0, i, 0, 0, 1))$, then if i pools, all judges believe preferences match; if i reveals $1 - M(h_i) = 1$, revealer judges and $(0, M(h_i))$ judges believe preferences match. Therefore, separating is the unique PBE strategy with a D1 refinement (point 4 of Lemma 5).

  Suppose $\mathcal{H}(\theta) > \hat{g}(0, i, 1, 0, 1)$, which only holds when private information yields. All judges believe preferences match if i pools or i reveals 0, but no judges believe preferences match if i reveals 1. Therefore, pooling on 1 is the unique equilibrium strategy (point 2 of Lemma 5). We can then conclude that if $\mathcal{H}(\theta) > \hat{g}(0, 0, 1, 0, 1)$, all players pool on $M(h_i) = 0$ (as stated earlier in this proof).

  If $\mathcal{H}(\theta) \in (\hat{f}(0, i, 0, 0, 1), \hat{g}(0, i, 1, 0, 1))$, revealer and $(0, 0)$ judges believe preferences match if i pools on $M(h_i)$, with out-of-equilibrium beliefs putting full weight on $1 - M(h_i)$ (point 1 of Lemma 5). If i reveals $1 - M(h_i)$, $(0, 1)$ judges believe preferences match. If i reveals $M(h_i)$, $(0, 0)$, revealer and possibly $(0, 1)$ judges believe preferences match. Suppose that when i reveals $M(h_i)$, only $(0, 0)$ and revealer judges believe preferences match. Then i of type $1 - M(h_i)$ gets the same social expectation from either action under a separating strategy or a pooling strategy. Therefore, a pure strategy exists. Now suppose all judges believe preferences match if i reveals $M(h_i)$. Then semi-separating may be the unique equilibrium strategy, if i of type $1 - M(h_i)$ prefers $M(h_i)$ under a separating strategy and $1 - M(h_i)$ under a pooling strategy. As I decreases or i increases, the proportion of revealers in the group increases, but since $\hat{g}(0, i, \phi_i^{obs}, 0, 1)$ also increases (point 5 of Lemma 4), $(0, 1)$ judges will believe preferences don't match if i reveals $M(h_i)$ for a value of I low enough or i high enough. The impact of I and i on the incentives to reveal is therefore non-monotonic. However, for all values of I and i such that $(0, 1)$ judges believe preferences don't match if i reveals $M(h_i)$, increasing I and decreasing i increases incentives to separate.

- $\Delta(h_i) = 1$. If i pools,

  $$\hat{g}(1, i, 0, 0, 0) = \hat{f}(1, i, 0, 0, 0) = 1 < \hat{f}(1, i, 0, 1, 1) = \hat{f}(0, i, 0, 0, 1) <$$
  $$\min\{\hat{g}(1, i, 0, 1, 1), \hat{f}(1, i, 0, 0, 1)\} < \hat{g}(1, i, 0, 0, 1)$$

  If player i reveals to be of type 1, $\hat{f}(1, i, 1, \phi_j^{obs}, \phi_j) = \infty^+$ for all j, and

  $$\hat{g}(1, i, 1, 0, 0) = \hat{f}(1, i, 0, 1, 1) < \hat{f}(1, i, 0, 0, 1) < \hat{g}(1, i, 1, 1, 1) < \hat{g}(1, i, 1, 0, 1)$$

  If player i reveals to be of type 0, $\hat{f}(1, i, -1, \phi_j^{obs}, \phi_j) < 0$ for all j, and

  $$\hat{g}(1, i, 1, 0, 0) < \hat{f}(1, i, 0, 1, 1) < \min\{\hat{g}(1, i, 1, 1, 1), \hat{f}(1, i, 0, 0, 1)\} < \hat{g}(1, i, 1, 0, 1)$$

  If $\mathcal{H}(\theta) \in (1, \hat{f}(0, i, 0, 0, 1))$, then if i pools, all judges believe preferences match; if i reveals $M(h_i)$, all judges believe preferences match; if i reveals $1 - M(h_i)$, $(0, 1 - M(h_i))$ and possibly revealer judges believe preferences match. If revealer judges believes preferences match when i reveals $1 - M(h_i)$, then separating is the unique PBE strategy with a D1 refinement (point 4 of Lemma 5). If only $(0, 1 - M(h_i))$ judges believe preferences match when i reveals $1 - M(h_i)$, then by point 6 of Lemma 5, there is either a separating or a pooling equilibrium strategy, and by points 3 and 5 we know there is no pooling or semi-separating on $1 - M(h_i)$, and that if separating is the equilibrium strategy, it is the unique strategy of a PBE equilibrium with the D1 refinement. As I decreases or i increases, the proportion of revealer judges increases, while $\hat{g}(1, i, \phi_i^{obs}, \phi_j^{obs}, \phi_j)$ increase if j is a revealer or $(0, 1)$ judge. But then the judges who believe preferences match when i reveals $1 - M(h_i)$ decreases, so incentives to separate decrease.

If $\mathcal{H}(\theta) \in (\hat{f}(0,i,0,0,1), \hat{g}(0,i,1,0,1))$, then there are two possible cases. If $\hat{g}(1,i,0,1,1) < \mathcal{H}(\theta)$, then since $\hat{g}(0,i-1,1,0,1) < \hat{g}(1,i,1,1,1)$ by points 3 and 5 of Lemma 4, so separating is an equilibrium strategy by point 4 of Lemma 5. If $\mathcal{H}(\theta) < \hat{g}(1,i,0,1,1)$, then if i pools $(0,1)$ and $(0,0)$ judges believe preferences match. If i reveals $M(h_i) = 1$, then revealer and $(0,1)$ judges believe preferences match. If i reveals $1-M(h_i) = 0$, $(0,0)$ and possibly revealer judges believe preferences match since $\hat{f}(0,i,0,0,1) < \hat{f}(1,i,0,0,1)$. If revealer judges believe preferences match, then separating is an equilibrium strategy. Otherwise, pooling on $M(h_i)$, separating or semi-separating on $M(h_i)$ may be equilibrium strategies (points 3 and 5 of Lemma 4). To show that any of these may be equilibruim strategies, suppose social expectation to type $1 - M(h_i)$ from separating is greater than the social expectation from pooling, and that i of type $M(h_i)$ would not deviate from pooling on $M(h_i)$ for any out-of-equilibrium beliefs. We can then use D1 to rule out a deviation from $M(h_i)$ if i is pooling on $M(h_i)$. Then the equilibrium strategy is either to pool on $M(h_i)$, semi-separate on $M(h_i)$ or separate depending on whether the payoff to type $1 - M(h_i)$ from choosing $1 - M(h_i)$ and revealing type $1 - M(h_i)$ is respectively greater than the payoff from pooling on $M(h_i)$, smaller than the payoff from pooling on $M(h_i)$ but greater than the payoff from choosing $M(h_i)$ and revealing $M(h_i)$, or smaller than the payoff from choosing $M(h_i)$ and revealing $M(h_i)$. As I decreaes or i increases, $\hat{g}(1,i,0,1,1)$ increases. Therefore, for I small enough or i large enough, $\mathcal{H}(\theta) < \hat{g}(1,i,0,1,1)$. Furthermore, as I continues to decrease or i increase, the proportion of revealers increases and $\hat{g}(1,i,\phi_i^{obs},\phi_j^{obs},\phi_j)$ increase if j is a revealer or $(0,1)$ judge. But then the judges who believe preferences match when i reveals $1 - M(h_i)$ decreases, so incentives to separate decrease.

- $\Delta(h_i) = -1$. If i pools,

$$\hat{g}(-1,i,0,0,0) < \min\{\hat{g}(-1,i,0,1,1), \hat{f}(-1,i,0,0,0)\} < \hat{f}(-1,i,0,1,1) =$$

$$\hat{f}(0,i,0,0,0) < \hat{f}(-1,i,0,0,1) = \hat{f}(0,i,0,1,1) = \hat{g}(-1,i,0,0,1) = 1$$

and $\hat{g}(-1,i,0,0,1) < \hat{f}(-1,i,0,0,1)$. If player i reveals to be of type 1, $\hat{f}(-1,i,1,\phi_j^{obs},\phi_j) = \infty^+$ for all j, and

$$\hat{g}(-1,i,1,0,0) < \hat{f}(-1,i,0,1,1) < \hat{f}(-1,i,0,0,1) < \hat{g}(-1,i,1,0,1) = \hat{f}(0,i,0,0,1)$$

and $\hat{g}(-1,i,1,0,0) < \hat{g}(-1,i,1,0,0) < \hat{f}(-1,i,0,0,1)$. If player i reveals to be of type 0, $\hat{f}(-1,i,-1,\phi_j^{obs},\phi_j) < 0$ for all j, and

$$\hat{g}(-1,i,-1,\phi_j^{obs},\phi_j) < \hat{f}(-1,i,0,0,0) < \hat{f}(-1,i,0,1,1) = \hat{g}(-1,i,0,0,1) = \hat{f}(0,i,0,0,0)$$

for revealer or $(0,0)$ judges j.

If $\mathcal{H}(\theta) \in (1, \hat{f}(0,i,0,0,1))$, then if i pools, all judges believe preferences match; if i reveals $M(h_i)$, all judges believe preferences match; if i reveals $1 - M(h_i) = 1$, $(0, 1 - M(h_i))$ judges believe preferences match. We know by point 6 of Lemma 5 that there is either a separating or a pooling equilibrium strategy, and by points 3 and 5 we know there is no pooling or semi-separating on $1 - M(h_i)$, and that if separating is the equilibrium strategy, it is the unique strategy of a PBE equilibrium with the D1 refinement. As I decreases or i increases, the proportion of revealer judges increases, while $\hat{g}(1, i, \phi_i^{obs}, \phi_j^{obs}, \phi_j)$ decreases if j is a revealer or $(0,0)$ judge. But then the judges who believe preferences match when i reveals $1 - M(h_i)$ decreases, so incentives to separate decrease.

If $\mathcal{H}(\theta) \in (\hat{f}(0,i,0,0,1), \hat{g}(0,i,1,0,1))$, then all judges believe preferences match if i pools, and none do if i reveals $1 - M(h_i)$. By point 2 of Lemma 5, i will pool on $M(h_i)$.

- $\Delta(h_i) = -2$. If i pools, $\hat{e}(-2, i, -1, \phi_j^{obs}, \phi_j) \leq 1$ for all j, $e \in \{f, g\}$. Therefore, if $\mathcal{H}(\theta) > 1$, all judges believe preferences match if i pools, and none do if i reveals $1 - M(h_i)$. But then i pools on $M(h_i)$.

- $\Delta(h_i) = 2$. If i pools,

$$1 < \hat{f}(0,i,0,0,1) = \hat{f}(2,i,0,0,0) < \hat{g}(2,i,0,0,0), \hat{f}(2,i,0,\phi_j^{obs},\phi_j) < \hat{g}(2,i,0,\phi_j^{obs},\phi_j)$$

for revealar and $(0,1)$ judges j. If player i reveals to be of type 1, $\hat{f}(2,i,1,\phi_j^{obs},\phi_j) = \infty^+$ for all j, and $\hat{g}(0,i,1,0,1) < \hat{g}(2,i,1,0,1)$. If player i reveals to be of type 0, $\hat{f}(2,i,-1,\phi_j^{obs},\phi_j) < 0$ for all j and

$$1 < \hat{g}(2,i,-1,0,0), \hat{f}(2,i,0,0,0) < \hat{g}(2,i,-1,1,1), \hat{f}(2,i,0,1,1) < \hat{g}(2,i,-1,0,1)$$

If $\mathcal{H}(\theta) \in (\hat{f}(0,i,0,0,1), \hat{g}(0,i,1,0,1))$, then all judges believe preferences match if i pools and if i reveals $M(h_i) = 0$. If i reveals $1 - M(h_i) = 1$, at most $(0, 1 - M(h_i))$ judges believe preferences match. If no judges believe preferences match when i reveals $1 - M(h_i)$, then i pools on $M(h_i)$ by point 2 of Lemma 5. If $(0, 1 - M(h_i))$ judges believe preferences match when i reveals $1 - M(h_i)$, then by point 6 of Lemma 4 we know there is either a separating or a pooling equilibrium strategy. By points 3 and 5 we know there is no pooling or semi-separating on $1 - M(h_i)$, and that if separating is the equilibrium strategy, it is the unique strategy of a PBE equilibrium with the D1 refinement.

If $\mathcal{H}(\theta) \in (\hat{f}(0,i,0,0,1), \hat{g}(0,i,1,0,1))$, then revealer, $(0,1)$ and possibly $(0,0)$ judges believe preferences match if i pools, with out-of-equilibrium beliefs putting full weight on $1 - M(h_i)$. If i reveals $M(h_i) = 1$ all judges believe preferences match. Therefore, if all judges believe preferences match when i pools, we can use past arguments to conclude that there is either a pooling or a separating strategy in equilibrium, there is no pooling or semi-separating on $1-M(h_i)$, and if separating is an equilibrium strategy it is a unique PBE strategy with the D1 criterion. So suppose that when i pools, only revealer and $(0,1)$ judges believe preferences match. If i reveals $1 - M(h_i)$, the set of judges that believe preferences match is either: none; $(0,0)$, $(0,0)$ and revealers, all. If revealers think preferences match when i reveals $1 - M(h_i)$, then we know separating is an equilibrium strategy (point 1 of Lemma 5). Otherwise, since the social expectation from revealing $M(h_i)$ are higher than the social expectation from from pooling, then the equilibrium strategy is either to pool on $M(h_i)$, semi-separate on $M(h_i)$ or separate, depending on whether the payoff to type $1 - M(h_i)$ from revealing $1-M(h_i)$ is respectively greater than the payoff of revealing $M(h_i)$, between the payoff of revealing $M(h_i)$ and pooling, or lower than the payoff of pooling.

For either value of $\mathcal{H}(\theta)$ that we have considered, as I decreases or i increases, the proportion of revealer judges increases, while $\hat{g}(1, i, \phi_i^{obs}, \phi_j^{obs}, \phi_j)$ increase for any judge j. But then the judges who believe preferences match when i reveals $1-M(h_i)$ decreases, and the judges who believe preferences match when i reveals $M(h_i)$ weakly increases, so incentives to separate decrease.

- $\Delta(h_i) = x \geq 3$. If i pools, $\hat{f}(0,i,0,0,1) < \min\{\hat{f}(x,i,0,\phi_j^{obs},\phi_j), \hat{g}(0,i,1,0,1)\} < \hat{g}(x,i,0,\phi_j^{obs},\phi_j)$ for all j, with $\hat{f}(3,i,0,\phi_j^{obs},\phi_j) < \hat{g}(0,i,1,0,1)$. If i reveals $M(h_i) = 1$, then $\hat{g}(0,i,1,0,1) < \hat{g}(x,i,1,\phi_j^{obs},\phi_j)$ for all j. If i reveals $1 - M(h_i)$, then $\hat{f}(x-1,i,0,0,0) < \hat{g}(x,i,-1,0,0)$, $\hat{f}(x,i,0,0,0) < \hat{g}(x,i,-1,1,1)$, and $\hat{g}(0,i,1,0,1) < \hat{f}(x,i,0,1,1) < \hat{g}(x,i,-1,0,1)$.

  If $\mathcal{H}(\theta) \in (1, \hat{f}(0,i,0,0,1))$, then if $\Delta(h_i) = 3$ all judges believe preferences match if i pools or reveals $M(h_i) = 1$, and no judges believe preferences match if i reveals $1 - M(h_i)$. Therefore, by point 2 of Lemma 5, i pools on $M(h_i)$.

  If $\mathcal{H}(\theta) \in (\hat{f}(0,i,0,0,1), \hat{g}(0,i,1,0,1))$, then if $x = 3$ and i pools, revealer, $(0,1)$ and possibly $(0,0)$ judges believe preferences match. If i reveals $M(h_i) = 1$, all judges believe preferences match. We can again use past arguments to conclude that if all judges believe preferences match when i pools, there is either a pooling or a separating strategy in equilibrium, there is no pooling or semi-separating on $1 - M(h_i)$, and if separating is an equilibrium strategy it is a unique PBE strategy with the D1 criterion.

So suppose that when i pools, only revealer and $(0,1)$ judges believe preferences match. If i reveals $1-M(h_i)=0$, $(0,0)$ and possibly revealer judges believe preferences match. Then using the same argument we used when $\Delta(h_i)=2$, we can conclude that the equilibrium strategy is either to pool on $M(h_i)$, semi-separate on $M(h_i)$ or separate. As x increases, $\hat{e}(x,i,\phi_i^{obs},\phi_j^{obs},\phi_j)$ for all ϕ_i^{obs}, all j and $e \in \{f,g\}$ increases. Therefore, social expectations from pooling and from revealing $M(h_i)$ weakly increase, as do social expectations from revealing $1-M(h_i)$. For $\hat{x}$ large enough, $\hat{g}(0,i,1,0,1) < \hat{e}(x,i,\phi_i^{obs},\phi_j^{obs},\phi_j)$, for all ϕ_i^{obs} and j, so i pools on $M(h_i)$. Using arguments similar to before, we can conclude that as I decreases or i increases, incentives to separate decrease.

Finally, I consider the dynamics when $\mathcal{H}(\theta) > \hat{f}(0,i,0,0,1)$ and private information sways. Again, we assume $M(h_1)=0$, or $\mathcal{H}(\theta)>1$, without loss of generality. We begin by considering values of $\Delta(h_i)$ in a history h_i such that all players $k<i$ have separated or pooled.

Suppose $\Delta(h_i)=0$. Then $\hat{g}(0,i,\phi_i^{obs},0,0)<0$, $\hat{g}(0,i,\phi_i^{obs},1,1)<(\hat{f}(0,i,0,0,1)$, and $\hat{g}(0,i,\phi_i^{obs},0,1)=\infty^+$ for all ϕ_i^{obs}. Therefore, if i pools or reveals $M(h_i)=0$, $(0,M(h_i))$ and revealer judges believe preferences match. If i reveals $1-M(h_i)$, $(0,1-M(h_i))$ judges believe preferences match. There is no ϕ_i^{obs} that gives a higher social expectation than from pooling, so we can use D1 to conclude that if i pools on $M(h_i)$, judges believe a deviation would be from a $1-M(h_i)$ type. Then the incentives from either action are the same whether i pools on $M(h_i)$ or separates, so player i either pools on $M(h_i)$ or separates in equilibrium. If player i reveals 0, player $i+1$ will pool on $M(h_i)=0$ (the analysis is identical to the case $\Delta(h_i)=-1$ for $\mathcal{H}(\theta)\in(\hat{f}(0,i,0,0,1),\hat{g}(0,i,1,0,1))$ when private information yields).

Suppose $\Delta(h_i)=1$. Then $\hat{g}(1,i,0,0,0)=\hat{f}(1,i,0,0,0)=1<\hat{g}(1,i,1,0,0)=\hat{f}(0,i,0,0,1)$, $\hat{f}(1,i,0,\phi_j^{obs},\phi_j)<\infty<\hat{g}(1,i,0,\phi_j^{obs},\phi_j)$.Then if player i pools or reveals $1-M(h_i)$, only $(0,0)$ types believe preferences match. We can then conclude that player i does not pool on $M(h_i)$, since $1-M(h_i)$ types believe most players are of type $M(h_i)$ (point 6 of Lemma 4). If i reveals $M(h_i)$, revealer and $(0,1)$ judges believe preferences match. Since i of type $1-M(h_i)$ believes most players are of type $M(h_i)$, the social expectation of revealing is greater than the social expectation of pooling. Then there is either a semi-separating on $M(h_i)$ or separating equilibrium.

Finally, suppose $\Delta(h_i)=x>1$. Then $\hat{f}(x,i,0,0,1)<\hat{g}(x,i,0,0,0)=\infty^+$. The social expectations from pooling depend on whether $\mathcal{H}(\theta)$ is greater or equal to $\hat{f}(x,i,0,\phi^{obs_j},\phi_j)$ for each judge j. The analysis of equilibrium strategies when the context is not strongly ϕ is the same as before. When private information sways, the context is strongly ϕ if $\hat{f}(I-1,i,0,0,1)<\mathcal{H}(\theta)$. That implies that there is no x such that pooling on $1-M(h_1)=1$

leads any judge to believe that i is most likely to be of type 1 if she pools. But then if $M(h_i) = 1$, revealer and type 1 judges believe preferences don't match while type $1 - M(h_i)$ believes most judges are of type $M(h_i)$, so there is no pooling on 1. Furthermore, in any history h_i such that some players semi-separate, $f(h_i, i, 0, \phi_j^{obs}, \phi_j, 1/2; I) < \hat{f}(i-1, i, 0, \phi_j^{obs}, \phi_j)$, so if the context is strongly ϕ there is no history of play such that players will pool on $1 - M(h_1)$.

G Proof of Result 2

If the context is uninformative, i does not semi-separate and will pool on ϕ only if there is at least $n_\phi \in [1, 3]$ (Result 1). Then if $I = 2$, there will be no pluralistic ignorance, since at least half of the players will act according to their type. If $I \to \infty$, the probability of pluralistic ignorance is positive. To see this, note that by the law of large numbers, the probability that the majority preference in a group is $1 - M(h_1)$ goes to zero. The probability that players pool on $1 - M(h_1)$, however, is positive: it is guaranteed if the first 3 players are of type $1 - M(h_1)$.

If the context is strongly ϕ and private information yields, then the probability of pluralistic ignorance is given by the probability that a strict majority of group members are of type $1 - M(h_1)$, equal to $P(\phi = bm \mid h_1)^2$ if $I = 2$ and 0 if $I \to \infty$ (by the law of large numbers). If the context is strongly ϕ and private information sways, then the probability of pluralistic ignorance is given by the probability that players pool on $M(h_1)$ and the majority of the group is of type $1 - M(h_1)$. The probability of this is zero or $P(\phi = bm \mid h_1)^2$ when $I = 2$, depending on whether player 1 separates or pools on $M(h_1)$. The probability becomes arbitrarily close to zero as $P(\theta = M(h_i))$ increases when $I \to \infty$ by the law of large numbers. □

H Proof of Result 3

If players have cooperative utility functions, the incentives for players of type 0 are the same as if they had the original utility function. As γ increases, type 1 players have a larger incentive to choose the action that increases the probability that other players will choose 1. Suppose pooling on 1 would be an equilibrium strategy after some history h_i if both types of player i had the original utility function ($\gamma = 0$). Then, with a cooperative utility function type 1 players would not deviate from pooling on 1 and would deviate from pooling on 0 – by D1, judges would believe the deviation would be done by a type 1 player, which would weakly increase future players of type 0's incentives to pool on 1.

Now suppose separating would be an equilibrium strategy at history h_i if both types of player i had the original utility function. Then with a cooperative utility function type 1 players would not deviate, who again increases future player of type 0's incentives to pool by revealing her type (by Result 8). But then, since we know by Result 8 that players choose pure strategies when the context is uninformative or strongly ϕ, we can conclude that with an initial run of at most two players of type 1, all players pool on 1.

For all $I \geq 5$, there is a sequence of types such that most players are of type 0, but all choose action 1. Therefore, pluralistic ignorance has a positive probability. Since more than half of the players are not acting according to their ideal point, $\sum_{i \in I} -(\phi_i - a_i^*)^2 < 0$, so we can find γ low enough that cooperative pluralistic ignorance is inefficient.□

I Proof of Result 4

By a similar logic to that in Lemma 2, we can write $P(\phi_i = 1 \mid a_{\mathcal{P}}, \phi_j) > 1/2$ as

$$\frac{P(a_{\mathcal{P}} \mid \theta = 1)}{P(a_{\mathcal{P}} \mid \theta = 0)} \frac{P(\phi_j \mid \theta = 1)}{P(\phi_j \mid \theta = 0)} > \frac{P(\theta = 0)}{P(\theta = 1)}$$

Since $P(a_{\mathcal{P}} \mid \theta) = 1 - P(1 - a_{\mathcal{P}} \mid \theta)$, then if $P(a_{\mathcal{P}} \mid \theta) > P(a_{\mathcal{P}} = 1 \mid 1 - \theta)$, then $P(1 - a_{\mathcal{P}} \mid 1 - \theta) > P(1 - a_{\mathcal{P}} \mid \theta)$. Therefore, if action $a_{\mathcal{P}}$ makes players believe it is more likely the state of the world is θ, $1 - a_{\mathcal{P}}$ makes players believe it is more likely the state of the world is $1 - \theta$. If the context is strongly ϕ, and $\mathcal{P}$'s actions are informative about the state of the world, $P(\theta = \phi \mid 1 - a_{\mathcal{P}}) > 1/2$ implies a context strongly $1 - a_{\mathcal{P}}$. But then action $1 - a_{\mathcal{P}}$ will not change equilibrium behavior. By Result 2, if $I \leq \tau_1(G)$, the principal would minimize pluralistic ignorance with action $a_{\mathcal{P}}$ if she set her strategy so that $P(\theta = \phi \mid a_{\mathcal{P}})$ implies the context is uninformative. This would be her optimal strategy if she can commit. If $\mathcal{P}$ cannot commit, she would deviate from this strategy to always choose $a_{\mathcal{P}}$.

If $I \geq \tau_2(G)$, we know by Result 2 that pluralistic ignorance is minimized if all players pool on θ. Therefore, the optimal strategy is to reveal θ.□

J Proof of Result 5

This follows from an analogous reasoning to the proof of Result 8. If $\tilde{\beta}$ and $|P(\phi_{\mathcal{P}} = 1) - .5|$ is large enough, then all survey respondents will pool on $\phi_{\mathcal{P}}$. Therefore, the survey will be uninformative about preferences. If $\tilde{\beta}$ or $|P(\phi_{\mathcal{P}} = 1) - .5|$ are sufficiently low, then survey respondents will reveal their type in the survey. Therefore, the survey will perfectly reveal the average preference in the group.□

K Proof of Result 6

Suppose $\mathcal{P}$ is supported. Then if $l = 1$ and an equilibrium strategy is such that there is an action a such that $\mathbb{E}(\mathbb{E}(\mathcal{K}_{j,i}(a_i = a) \mid \hat{\phi}_{-i})) > \mathbb{E}(\mathbb{E}(\mathcal{K}_{j,i}(a_i = 1 - a) \mid \hat{\phi}_{-i}))$, all players will prefer to choose a for δ large enough. Indeed, since $\beta > 1$, for δ large enough group members' main incentives are to choose whichever action maximizes $\mathbb{E}(\mathbb{E}(\mathcal{K}_{j,i} \mid \hat{\phi}_{-i}))$. If $l = 1$ and $\mathbb{E}(\mathbb{E}(\mathcal{K}_{j,i}(a_i = a) \mid \hat{\phi}_{-i}))\mathbb{E}(\mathbb{E}(\mathcal{K}_{j,i}(a_i = 1 - a) \mid \hat{\phi}_{-i}))$, then it must be the case that player i of type $(0, 1)$ is choosing a different action than player i of type $(1, 1)$. But then, for δ large enough, whoever of these two types is not choosing $\psi_{\mathcal{P}}$ has an incentive to deviate.

We can use a similar logic to show that if $\mathcal{P}$ is not supported and $l = 1$, all players pool on $1 - \mathcal{P}$. □

L Proof of Result 7

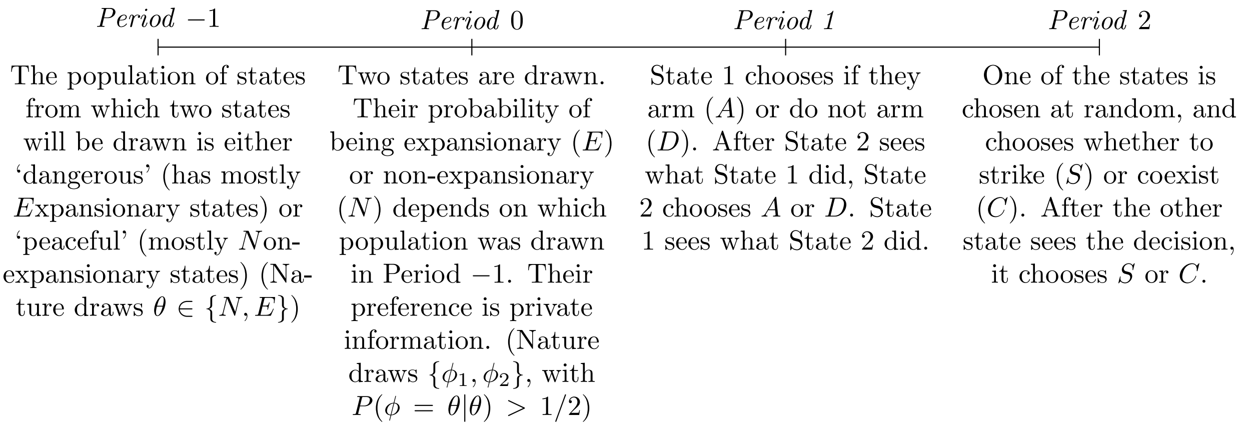
Suppose $\mathcal{P}$ is uninformed. If she is not brash but is benign, by Result 1 she will choose ϕ whenever the context is strongly ϕ. If she is brash but not benign, by Result 1 she will choose her optimal action, which will not guarantee all players pool on $\phi_{\mathcal{P}}$ for an uninformative context. If she is brash all players will learn her type through her action. If she is benign, players who learn about her action learn about the group's majority preference. Therefore, a benign and brash principal's actions reveals the group's majority preference without i for all i, which all players pool on.□

M A Version of the Model to Analyze Inter-State Conflict Bargaining

M.1 Setup

The model represents two states interacting under anarchy. There will be two type of states: expansionary states that gain from an inefficient war, and non-expansionary states that do not. Non-expansionary states will want to find out what type of state they are dealing with in order to decide whether they must be prepared for and engage in war. The timeline of the game is represented in Figure 3.

In order to represent this situation, we will imagine that the states are randomly drawn from a population of states. The probability that a randomly drawn state is expansionary will affect non-expansionary states' optimal choice. In populations where it is very likely that the state is expansionary, the non-expansionary state will anticipate that the state



Period -1

The population of states from which two states will be drawn is either ‘dangerous’ (has mostly *E*xpansionary states) or ‘peaceful’ (mostly *N*on-expansionary states) (Nature draws $\theta \in \{N, E\}$)

Period 0

Two states are drawn. Their probability of being expansionary (E) or non-expansionary (N) depends on which population was drawn in Period -1. Their preference is private information. (Nature draws $\{\phi_1, \phi_2\}$, with $P(\phi = \theta|\theta) > 1/2$)

Period 1

State 1 chooses if they arm (A) or do not arm (D). After State 2 sees what State 1 did, State 2 chooses A or D. State 1 sees what State 2 did.

Period 2

One of the states is chosen at random, and chooses whether to strike (S) or coexist (C). After the other state sees the decision, it chooses S or C.

Figure 3: Timeline

it is interacting with is expansionary, so will want to prepare for and engage in war. In populations where it is very likely that the state is non-expansionary (a ‘peaceful’ population of states), a non-expansionary state will want to avoid preparing for and going to war. It will be important to my model that we are able to study what happens as we vary states’ certainty over what type of state they are likely to deal with. We do this by having an initial stage of the game where one of two populations are drawn. In one population (a ‘dangerous’ population of states), there are more expansionary types ($\theta = E$). In the alternative, ‘peaceful’ population of states, there are more non-expansionary types ($\theta = N$). The probability that the dangerous population of states is drawn is given by $P(\theta = E)$.

Once the population is drawn, two states will be randomly drawn from this population. If they are drawn from population E, they will be more likely to be expansionary types. If they are drawn from population N, they are more likely to be non-expansionary types. States know their type, but they don’t know the other player’s type, nor the state of the world. They do know $P(\theta = E)$ and the probability that their type was drawn given the state of the world. In fact, I assume that these probabilities are common knowledge. Further, I assume that the probability of a state of type E being drawn in state of the world E is the same as a state of type N being drawn in state of the world N: $P(\phi = N \mid \theta = N) = P(\phi = E \mid \theta = E) > 1/2$. Notice that preferences are informative: a state of type N assigns a higher probability to the population being N than a state of type E. States will use their preferences and their priors over the state of the world to assess the other state’s preferences.

After the population and the states are drawn, they will take turns choosing whether to arm, and then choosing whether to strike. One of the two states is randomly assigned to be State 1, the other is State 2. In period 1, State 1 decides if it arms (A) or does not arm (D). State 2 observes State 1’s decision, and then chooses A or D. At the end of Period 1, State 1 observes what State 2 chose. In period 2, one of the two states is chosen with equal probability to choose whether to strike (S) or coexist (C). The other state sees the decision, and chooses S or C. We assume that if the other state has struck first, it will always be in the interest of a state to strike second. If the second state to get an opportunity to strike is the first to strike, the other state gets the payoffs as if it struck second.

We assume that because of commitment problems, honoring a bargaining agreement may not be an option. We follow Fearon (1995) in assuming that there is a set of issues represented by the interval $[0, 1]$. State 1 prefers issues closer to 1, while state 2 prefers issues closer to 0. There is an initial endowment or status quo of the issue $x \in (0, 1)$. The interval can be thought of as territory, and the status quo as the border. The states' utility for the outcome $x \in [0, 1]$ is $u_1(x; \phi)$ and $u_2(1 - x; \phi)$, where $u_1(\cdot\phi)$ and $u_2(\cdot; \phi)$ are continuous, increasing and weakly concave. Further, set $u_i(1; \phi) = 1$ and $u_i(0; \phi) = 0$ for $i \in \{1, 2\}$. If the states fight a war, the state prevails with a probability $p : \{f, s\} \times \{N, E\}^2 \times \{A, D\}^2 \to [0, 1]$, which depends on whether the state was a first (f) or second (s) striker, both states' types and the arming decision of both states. The cost of engaging in war is $c > 0$. Whoever wins the war keeps the issue. Therefore, the expected utility of engaging in war is $p(\cdot, \cdot, \cdot) - c$. We make the following assumptions about p:

- The probability of winning for non-expansionary states at war with each other does not depend on whether they are armed: $p^f \equiv p(f, N, N, w_1, w_2) > p^s \equiv p(s, N, N, w_1, w_2)$ for $w_1, w_2 \in \{A, D\}$.Further, war between expansionary states is inefficient: $p^o - c < u_1(x; N)$ and $1 - p^o - c < u_2(1 - x; N)$ for $o \in \{f, s\}$. These are strong assumptions, but they are meant to capture the idea that non-expansionary states prefer the status quo than to fight with each other, an idea that can be captured with a weaker but more drawn out set of assumptions.[16]

- Expansionary types do better off than the status quo if they strike first. Further, they do better off the higher their relative advantage in terms of arms: $u_1(x; E) < p^s(l) - c < p^f(l) - c < p^s(e) - c < p^f(e) - c < p^s(m) - c < p^f(m) - c$, where l stands for 'less' armed ($p^o(l) \equiv p(o, E, \phi_2, A, D)$ with $o \in \{f, s\}$), e stands for 'equally armed' ($p^o(e) \equiv p(o, E, \phi_2, A, A) = p(o, E, \phi_2, D, D)$ with $o \in \{f, s\}$), and m stands for 'more armed' ($p^o(m) \equiv p(o, E, \phi_2, D, A)$ with $o \in \{f, s\}$). The utility for an expansionary state 2 is defined analogously. Notice that as long as the expansionary state is a

[16] Suppose non-expansionary states are satisfied with the status quo. Formally, they don't get higher utility from getting more of the issue: $u_1(x; N) = u_2(1 - x; N) = 1$. This makes them truly non-expansionary states, in the sense that they have no taste for doing so. Then any war would be inefficient. This can be complemented by assuming that any war between expansionary states would make them do worse than the status quo in expectation: $p(\cdot, N, N, \cdot, \cdot) - c < u_1(x; N)$ and $1 - p(\cdot, N, N, \cdot, \cdot) - c < u_2(x; N)$.

first mover, it does not matter whether it is at war with an expansionary or non-expansionary state. Further notice that as long as both states are equally armed, the expected utility is the same. These assumptions are meant to avoid notational burden, and can be weakened.

- If non-expansionary types strike first against an expansionary type, they manage to defend themselves. We capture this by having $p = p(f, N, E, w_1, w_2)$ for all $w_1, w_2 \in \{A, D\}$. Once again for simplicity, I assume that that whether either state is armed does not affect this probability.

These assumptions capture the intuition that expansionary states want to go to war with an arms or first mover advantage, and non-expansionary states only want to go to war to defend themselves. An example of payoffs that satisfy the above conditions is captured in Figure 4, which shows the expected payoff of war for state 1.

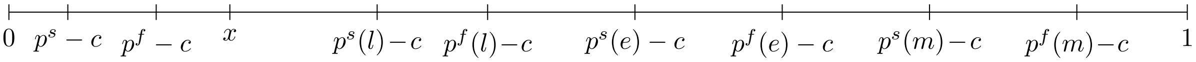


Figure 4: Expected Payoffs From Going To War For State 1

Notice that the only situation in which the status quo is part of the bargaining range are the conditions that lead to probability p: non-expansionary states fighting against each other, or a non-expansionary state striking first against an expansionary state. This implies that non-expansionary states prefer to bargain than to go to war with each other, but there is no bargaining range for an expansionary state fighting another state.

M.2 Analysis

In this section I will give an intuition of the results.

Expansionary states will always want to strike. Because of this, non-expansionary states who make the second choice about whether to strike will know that if the other state did not strike, it is a non-expansionary state. Therefore, non-expansionary states who make the second choice about whether to strike will do so if and only if the other state chose to strike. Knowing this, a non-expansionary state who makes the first choice to strike will do so if they believe it is sufficiently likely that the other state is an expansionary state. If i is the state who chooses to strike first and j is the other state, i strikes first if and only if:

$$P(\phi_j = E \mid \phi_i = N, h_i) > \frac{u_1(x; N) - p + c}{u_1(x; N) - p^s(r) + c > 0} \tag{20}$$

where h_i is the history of play and $r \in \{l, e, m\}$ is determined and known in equilibrium. We will summarize condition (20) with the 'judgment function' $J(h_i)$, equal to one if (20) holds, and zero otherwise.[17]

In the arming stage, states must compare the benefits of four possible strategies: arming and striking, arming and not striking, not arming and striking and not arming and not striking. We know that non-expansionary states will always want to strike, so when analyzing their decision we only need to consider the benefits of arming and striking versus the benefits of not arming and striking. Let state $i \in \{1, 2\}$ be a non-expansionary state, and j is the other state. The benefit of arming and striking for i is:

$$P(\phi_j = E \mid h)[.5(p^f(r_E) - c) + .5(p^s(r_E) - c)] + \\ P(\phi_j = N \mid h)\Big[.5(p^f(r_N) - c) + .5[\mathbb{E}(J(h_i))(p - c) + (1 - \mathbb{E}(J(h_i)))(p^f(r_N) - c)\Big]$$

where $r_E, r_N \in \{l, e, m\}$ are the relative power of expansionary state i with respect to the other state of type E or N, respectively. Conditional on j's type, in equilibrium i will know the value of r_{ϕ_j} when making its decision whether it chooses first or second. The top line represents the expected outcome from encountering an expansionary state. Each state can choose to strike first with fifty percent probability, and they will. The bottom line represents the expected outcome from encountering a non-expansionary state. With fifty percent probability, the expansionary state strikes first, which gives it utility $.5(p^f(r_N) - c)$. If the non-expansionary state j makes the choice of whether to strike first, however, the outcome will depend on j's beliefs over what i wants. However, state j has to think what i would think about whether j is an expansionary state. This is captured by the term $\mathbb{E}(J(h_i)) = P(J(h_i) = 1)$, or the expectation that i believes i and j have different preferences. This is the heart of a second order conformity model.

When deciding whether to arm, expansionary states trade off having an arms advantage in war versus losing the first strike advantage with non-expansionary states who interpret the decision to arm as a signal of their type.[18] Non-expansionary states trade off being able to defend themselves from an expansionary state by arming with having non-expansionary states misinterpret them for expansionary states and enter into war. Importantly, for non-expansionary states to not go to war, they need to believe that it is likely that the other state is non-expansionary, *and* that the other non-expansionary state believes it is likely that it is non-expansionary. Without these two conditions on beliefs, non-expansionary states will be better off by striking first to defend themselves.

[17] I'm ignoring the knife-edge case of equality for this draft. It does not change any of the results.

[18] Behind the scenes, I am using the assumption that expansionary states get a higher relative payoff from arming than from not arming relative to non-expansionary states. This allows whoever observes the decision to arm to think it is weakly more likely for an expansionary state to have armed.

Without further justification, I state the result. Since the logic of the equilibrium is very similar to that of Result 8, I will state it as a claim.

Claim 1. *When there is a high probability that the population is mostly of expansionary types, and there are many expansionary types in that population ($P(\theta = E)$ and $P(\phi = E \mid \theta = E)$ higher than some threshold), states will arm and strike in the unique equilibrium.*

When there is a low probability that the population is mostly of expansionary types, and there are many non-expansionary types in a population of mostly non-expansionary types ($P(\theta = N)$ and $P(\phi = N \mid \theta = N)$ higher than some threshold),states will not arm. Expansionary states will strike. Non-expansionary states will not strike first.

When the probability that the population is mostly of expansionary states is in between these two extremes, an expansionary state 1 will arm, and a non-expansionary state 1 will not arm. State 2's choice of arming will be the same as that of state 1. Expansionary states will strike. Non-expansionary states will not strike first.

When states are convinced that the other state is expansionary, it arms and strikes. This happens if the prior probability that a state is expansionary is high, or if the state has revealed itself to be expansionary. But if the prior probability that a state is expansionary is high, then states of all types want to arm, so behavior does not reveal who the expansionary state is. Since states rely on their prior probabilities at the striking stage, all prefer to strike. This is the type of misunderstandings between non-expansionary states that leads to war.

The converse happens when the prior probability that the state is non-expansionary is low: no one arms, which means that at the striking stage states use their priors to decide whether to strike. Since non-expansionary states believe it is likely that the other state is non-expansionary, it does not strike. However, an expansionary state will be able to strike an unarmed non-expansionary state. This equilibrium tracks the analysis in Fearon (1995) regarding Germans' efforts to sneak attack other European powers in the July crisis.

Finally, when there is an intermediate probability, the first expansionary state arms and the first non-expansionary state does not arm. The reason is that their priors over the population from which states are drawn is weak, so they rely heavily on their private information – their own preferences – to determine whether the population is composed mostly of expansionary or non-expansionary states. Expansionary states believe it is the first, so it arms. Non-expansionarly states believe it is the second, so it does not arm. I will discuss this case further in Section IV.