Job Market Paper

Pluralistic Ignorance and Social Change*

A Model of Conformity to the Perceived Majority

Mauricio Fernández Duque†

October 27, 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. 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'. Strong beliefs over others' views increases pluralistic ignorance in small groups, but decreases it in large groups. Behavior may be affected by leaders, laws or surveys that influence what is thought to be the majority preference, possibly creating pluralistic ignorance. Abrupt social change may come about through an everyman who reveals what everyone wishes they were doing. The model integrates insights from scholarship on collective action, public opinion, and social meaning regulation, and then applies these insights to political phenomena such as misperceived support for discriminatory policies, the rise of the Arab Spring, beliefs about climate change beliefs, and the impact of get out the vote campaigns.

*The latest version of the paper can be found at https://scholar.harvard.edu/duque/research. I would like to thank Samuel Asher, Nava Ashraf, Robert Bates, Iris Bohnet, Jeffry Frieden, Benjamin Golub, Michael Hiscox, Horacio Larreguy, Scott Kominers, Pia Raffler, 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 situation in which '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 also been found in small groups. In Southern universities in the sixties, white students would avoid openly supporting integration in conversations, fostering a mistaken impression of the extent of support for segregation (Turner, 2010). While strong beliefs over others' views increased the probability of pluralistic ignorance among students, they decrease pluralistic ignorance in public opinion polls (Shamir and Shamir, 2000).

In this paper I provide a rational model of pluralistic ignorance in large and small groups via *second-order conformity*, in which group members are motivated to act according to the perceived majority preference. The 'second-order' modifier refers to the fact that individuals form beliefs over others' beliefs about preferences.[1] Second-order conformity provides insights into questions about public opinion and collective action. When will misperceptions of opinion arise in micro and macro interactions? How is social change effected in large group interactions? How is it effected in small group interactions? What explains abrupt social change? I address these questions by considering how pluralistic ignorance arises in public opinion, communication and collective action problems, as well as the types of leaders and policies that make pluralistic ignorance less likely.

Pluralistic ignorance can be broken down into three features, which in the model result from equilibrium misinformation over what to conform to. First, group members act *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. A perverse failure of collective action may result from pluralistic ignorance, with individuals reluctantly pursuing what they mistakenly think is socially optimal.

[1] Past models of conformity have assumed that individuals know which preference others consider valuable (e.g. Bernheim, 1994, Ellingsen, Johannesson et al., 2008, Benabou and Tirole, 2012).

To introduce the model, consider a stylized conversation between a group of acquaintances in which each announces whether they support or oppose segregationist housing. A concern for conforming motivates them to express a view that matches their acquaintances' views. Each individual must use *context* to form beliefs about others' views – the probability that most acquaintances in a given situation are pro-segregationists. In the Southern universities in the sixties, white students believed acquaintances' views were very likely to be pro-segregationist. In Northern universities, where views on segregation had been changing more rapidly, it was less clear what acquaintances' views were.

Dynamics of how opinions are expressed depend on the context. While an informative context leads no one to reveal their private view, first speakers in an uninformative context reveal their private view and have a strong impact on the opinion others express. For example, when the context strongly indicates that acquaintances are pro-segregationist, all speakers conform by expressing support for segregationist housing. When context is uninformative, the first speaker relies on her own views to infer what the majority view is. The first speaker then expresses her private view, which impacts the opinions expressed by others.[2] Pluralistic ignorance – a conversation in which most express an opinion that differs from their private view – can therefore arise from a strong context or from the minority views of the first speakers.

We can reinterpret the model to analyze public opinion formation. Consider a 'conversation' among *all* whites in the U.S. in 1968, who take turns publicly expressing their opinion on segregationist housing. This captures, in a simple way, how individuals form their opinions by observing what others in society have said. If we further assume individuals are motivated to conform to what they believe is the majority view, we are back to the second-order conformity model, in this case with a large group. By considering small and large group interactions in the same framework, we can compare the conditions under which pluralistic ignorance arises. I show the probability of pluralistic ignorance is lowest in small groups when the context is uninformative, and lowest in large groups when the context is informative. That is, the dynamics that lead to pluralistic ignorance in small groups are the same that impede pluralistic ignorance in large groups, and vice versa.

With these tools for studying pluralistic ignorance in hand, I turn to the question of social

[2]The uninformative context dynamic is analogous to *herding* (Banerjee, 1992) or *information cascades* (Bikhchandani, Hirshleifer and Welch, 1992). In the model, individuals may herd on what they think the majority group preference is, leading them to conform to the minority view. 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. Whereas group size does not affect the probability of herding, it affects the probability of pluralistic ignorance. Furthermore, a motivation to conform leads to different behavior when individuals think they will be judged and when they do not. I discuss experimental evidence for this behavior in section 4.2.2.

change. Consider a broad interpretation of the model that allows us to analyze decisions in which conformity matters. Instead of expressing an opinion, individuals may be deciding whether to protest, vote, or contribute to a charity, to name a few examples. In an important sense, all of these examples are different from expressing an opinion because they affect the decision maker and others' outcomes more directly. However, in all of these decisions there are pressures towards conformity, and individuals look to others to decipher what to conform to. Individuals are motivated to vote if they think others think voting is important, or to protest if they think others dislike the regime, or to contribute to a charity they think others value. My objective is to look for novel predictions that arise when we focus on conformity.

To motivate how the framework helps us think of social change in large groups, consider the puzzling variation in the impact of social information campaigns. These campaigns measure a group's behavior or opinion and report the average response back to the group Kenny et al. (2011). The popularity of these campagins has been growing, and they have been used to affect behavior as varied as voting (Gerber and Rogers, 2009), charitable contributions (Frey and Meier, 2004), and tax evasion (Wenzel, 2005). The underlying assumption behind these campaigns is that there is a misunderstanding of what the average behavior is, and that individuals are motivated to conform to the average behavior (e.g. Wenzel, 2005). Despite sometimes large impact on behavior, their success has been mixed.

How information is collected may explain the mixed success of social information campaigns. Some behaviors and opinions can only be measured through surveys, such as unethical practices (Buckley, Harvey and Beu, 2000), intention to vote, climate change beliefs (Mildenberger and Tingley, 2016), or policy positions. However, respondents tend to conform to what they think the surveyor wants to hear – the so-called 'social desirability bias'. Because respondents do not know what the surveyor wants to hear, we can naturally think of this in the context of second-order conformity. The success of a social information campaign will then depend on the perceived extend of social desirability bias. To the best of my knowledge, this straightforward hypothesis has not received attention in the literature. The framework further predicts that surveyors who want truthful responses will try to obscure their own views and to avoid seeming judgmental, which are best practices in the profession.

Abrupt social change can also be profitably analyzed with second-order conformity. When there is pluralistic ignorance, a social movement 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.

While protesting, voting and charitable giving can all be thought of as large group interactions, second-order conformity also arises in small group interactions, perhaps in surprising

ways. Consider dueling, which will provide a good illustration of how social change can come about in a small group interaction. If a member of the elite felt that they were being disrespected by another, they would demand 'satisfaction' via a duel (Williams, 2000). What was considered grounds for a duel, moreover, was often determined by social custom. An elite member might be expected to call for a duel after a slight, even if he did not consider it to be worthy of a duel. Furthermore, it was considered dishonorable to not provide satisfaction to someone who demanded a duel.

The decision to duel can be thought of with the help of the model. An individual who is slighted by another must decide whether to propose a duel, and the offender must decide whether to accept. We can think of this as a two person interaction where each announces whether they would like to be part of a duel. Individuals have a private view on whether they would like to duel, and must trade off acting according to their view and acting according to what they think the other wants. Their private view captures whether they feel that 'satisfaction' of a slight is necessary. If neither feels that this is the case, they could go their separate ways without losing honor. However, it is dishonorable to not want to duel someone who wants to duel.

In the eighteenth and nineteenth century in the U.S., U.K. and other European countries, dueling was a common social practice among the elites. However, by the twentieth century, dueling was no longer practiced. To understand the policies and campaigns that brought about this social change, it is helpful to think of the concerns with dueling as concerns over pluralistic ignorance. In a context that strongly indicates that individuals value dueling, two individuals that would rather not duel may end up doing so to avoid losing honor (Schwartz, Baxter and Ryan, 1984). If pluralistic ignorance captures the concern over dueling, then policies should diminish dueling by facilitating the communication of a desire to not duel, or by changing beliefs over how much dueling is valued. Indeed, these objectives were pursued through codes of honor, which allowed intermediaries to call off an unwarranted duel (O'Neill, 2003), laws to ban duelers from holding prestigious public office (Lessig, 1995), and a campaign against the 'false' honor of dueling (Shoemaker, 2002, Andrew, 1980). Policies that did not address the problem of pluralistic ignorance in dueling, such as fines or jail time for duelers, were unsuccessful.

Beyond dueling, second-order conformity arises in small-group interactions such as sensitive conversations, open voting in committees, state decisions to go to war, and more social examples such as dating, drinking among friends, and gift-giving exchanges. As suggested by the dueling and social information campaign examples, the second-order conformity framework can shed light on current policies to effect social change. Furthermore, the approach provides tools to avoid miscommunication and improve collective action with testable impli-

cations.

Pluralistic ignorance has been found in a variety of issues, including religious observance (Schanck, 1932), climate change beliefs (Mildenberger and Tingley, 2016), tax compliance (Wenzel, 2005), political correctness (Van Boven, 2000), unethical behavior (Halbesleben, Buckley and Sauer, 2004), and alcohol consumption (Prentice and Miller, 1993). 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). Past models do not capture all three features of pluralistic ignorance I described. In particular, they typically do not feature equilibrium misunderstandings about what the majority prefers, a point I develop in more detail in section 4.2. These explanations of pluralistic ignorance have been used to analyze social change (e.g Kuran, 1997, Lohmann, 1993), and are related to a large rational choice literature on collective action (Olson, 1965, Hardin, 1982, Esteban, 2001, Siegel, 2009). The approach to pluralistic ignorance presented here allows us to capture forces that are not typically found in game-theoretic models. In particular, the model complements scholarship emphasizing the importance of social meaning and context in understanding and affecting behavior (Bates, de Figueiredo Jr and Weingast, 1998, Finnemore, 1996, Tarrow, 2011, McAdam, McCarthy and Zald, 1996). Second-order conformity also links scholarship on collective action with scholarship on public opinion (Noelle-Neumann, 1974, Katz and Allport, 1931, Shamir and Shamir, 2000, J O'Gorman, 1986), two strands of literature that have largely developed independent of each other. My work also relates to a literature on equilibrium misaggregation of information (Acemoglu et al., 2011, Bikhchandani, Hirshleifer and Welch, 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, Cain and Dawes, 2006, Dana, Weber and Kuang, 2007, DellaVigna, List and Malmendier, 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 leaders, laws and surveys 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, provide a couple of interpretations, 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 indexed by $\phi_i \in \{0, 1\}$. These ideal points are drawn from an unobserved population distribution $\theta \in \{0, 1\}$. Nature selects the population distribution θ with probability $\chi \equiv P(\theta = 1) \in [0, 1]$. Probability χ will capture commonly held priors over population from which the group is drawn, 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 $\pi \equiv P(\phi_i = \phi \mid \theta = \phi) > 1/2$ – population $\theta = \phi$ is more likely to produce groups with majority type ϕ. Since π indicates how likely it is to find a type in a given population, I will refer to it as the 'precision' of the population. Although θ is not observed and ϕ_i is private information, both the context χ and the precision π are a commonly known part of the environment.

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.

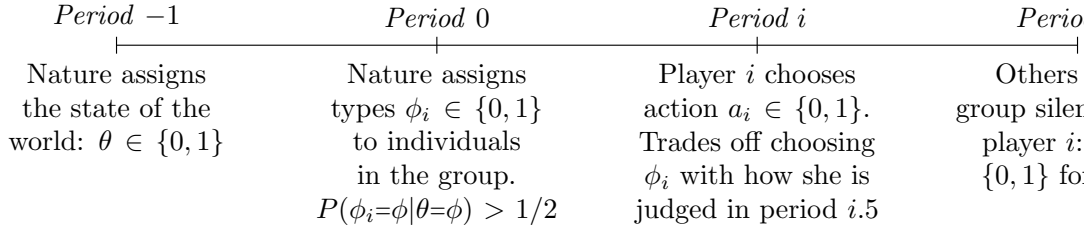

Figure 1: Timeline

After i takes an action, and before $i + 1$ does, all players other than i judge i based on her action. Player i trades off choosing her ideal point with choosing whichever action she

expects will be best judged by others. Let $\mathcal{J}_{j,i}(a_i) \in [0,1]$ be j's judgment of i, and let $-i$ be the set of players in the group other than i, or the 'group without i'. As I elaborate below, j judges whether i's preference match her own preference. Player i's payoff is affected by the average of judgments $\mathbb{E}_{j\neq i}(\mathcal{J}_{j,i} \mid \hat{\phi}_{-i})$, where $\hat{\phi}_{-i}$ is the distribution of preferences in the group without i. If i knew the true distribution of others' preferences, $\mathbb{E}_{j\neq i}(\mathcal{J}_{j,i} \mid \hat{\phi}_{-i})$ would be a deterministic function of a_i. Since i does not know the distribution, she maximizes her expected utility: $a_i^* = \arg\max_{a_i\in\{0,1\}} \mathbb{E}u(a_i; \phi_i, h_i, |I|) =$

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

where the outer expectation operator in the second summand is over all the distributions of preferences $\hat{\phi}_{-i}$. I refer to the expectation over average judgments as the '*expected judgment*' or the '*social expectation*'. Note that I will generally use i for the current period's decision maker, and j for a judge of i.

When player j judges player i, j must decide whether i's preference matches her own preference ($\mathcal{J}_{j,i} = 1$) or not ($\mathcal{J}_{j,i} = 0$). If there is uncertainty over i's preference, then j sets $\mathcal{J}_{j,i} = 1$ if it is most likely that i's preference matches her own, or $P(\phi_i = x \mid h_i, a_i, \phi_j = x) > 1/2$. If j believes it is more likely that preferences do not match, or $P(\phi_i = x \mid h_i, a_i, \phi_j = x) < 1/2$, then $\mathcal{J}_{j,i}(a_i) = 0$. When there are no mixed strategies in equilibrium, the intermediate case of $P(\phi_i = x \mid h_i, a_i, \phi_j = x) = 1/2$ is a knife edge scenario, and $\mathcal{J}_{j,i}(a_i)$ takes an arbitrary value in $[0,1]$. These judgments are not observed by others.

The uncertainty over judgments is the heart of second-order conformity. If there were no uncertainty over the distribution of group preferences, 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).

2.2 Interpretation

Before moving on to defining strategies and equilibrium, I will first relate the setup of the model to the example of segregationist preferences given in the introduction. I then provide a more general interpretation of the model.

2.2.1 Segregationist Preferences

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 ($\theta = 1$), or from a population which consists mostly of anti-segregationists ($\theta = 0$), but group members do not observe which population they were drawn from. Further, there is common knowledge over the context χ – the probability they were drawn from population $\theta = 1$. Individuals are motivated to conform by choosing an action that reveals their prefererence matches that of the majority in the group.

We can use this setup to think of small scale and large scale interactions. An example of a small scale interaction is a conversations among students. Southern universities in the sixties were a strongly pro-segregationist context ($P(\theta = 1)$ is high), whereas the context in Northern universities was more uninformative about others' preferences ($P(\theta = 1)$ is close to $1/2$).

An example of a large scale interaction is society-wide public opinion. The model provides a stylized way of capturing how public opinion is formed, in which past public announcments of opinion affect which opinions others are willing to state. The U.S. in the 50's are a context where the majority of white Americans were believed to favor segregationist preferences, whereas the distribution of preference was much less clear in the 70's.

Notice the difference in the large scale and small scale examples. With the small scale examples, context could be thought of in terms of a larger set of individuals from which a group was drawn – such as participants in a conversation drawn from a specific university. It is not natural to think of context in this 'frequentist' way for large scale interactions. However, we can still think of context in a more 'subjectivist' sense, as capturing the beliefs over the underlying probability with which Americans take on a certain preference. I will elaborate on this distinction in section 4.3.

2.2.2 A More General Interpretation

We can think of the model as capturing situations where individuals take an action where they care about revealing that they want what others want. Generally, these are situations individuals have a strong opinion on, or that affect others: controversial conversations, protesting, turning out to vote, voluntarily providing a public good, engaging in a fight, courtship, and so on. Individuals are assumed to trade-off acting according to their private views and to what they think others think is what most in a group prefer. To be clear, individuals surely care about more than this trade-off. I abstract from these other considerations simply to highlight the novel predictions that come out of a second-order conformity approach. However, in section 3.2.3 I consider some modifications to the utility function to

allow for externalities.

I now provide more detail on some of the examples I've discussed. In section 4 I discuss further examples.

Protesting. Consider a society comprised of $|I|$ citizens who, one by one, decide whether they will protest ($a_i = 1$) or not ($a_i = 0$). The sequential nature of the decision captures in a simple way how waves of protests may spread or dissipate. Individuals either disfavor the regime ($\phi_i = 1$) or favor it ($\phi_i = 0$), and are assumed to want others to think that she shares their view on the regime. The context captures individuals' common knowledge over whether the society is likely made up of mostly supporters or mostly opponents of the regime.

Dueling. Consider two individuals ($|I| = 2$) who must decide whether to duel, perhaps because one made a minor offense that can be grounds for dueling. The first player decides whether to propose a duel, and the second decides whether to accept. Put another way, they both decide whether to make engage in a fight ($a_i = 1$) or not ($a_i = 0$). Individuals either want to duel ($\phi_i = 1$) or do not ($\phi_i = 0$). Suppose that if a player wants to duel, the other would lose honor from not wanting to duel. However, if a player does not want to duel, the other would not lose honor from not dueling. Therefore, both players would like to match the preferences of the other – they want the other to think that both want to duel, or that neither wants to duel.[3] The context captures individuals' common knowledge over whether they likely want to or do not want to duel.

2.3 Strategies and equilibrium

Denote by $G \equiv \{\chi, \pi, \beta, I\}$ the game defined above with context χ, precision π, 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 will be interested in Perfect Bayesian Equilibrium, and use the intuitive criterion to refine out-of-equilibrium beliefs (Fudenberg and Tirole, 1991). The intuitive criterion establishes that, if an action is observed that no type would choose on the equilibrium path of play, individuals will infer that the deviation came from a type who would deviate for *some* out-of-equilibrium beliefs.

[3] In the current setup, an individual who wants to fight would want to signal that he does not want to fight if that is what he thinks the other player wants. An alternative setup would allow for an individual to prefer the other to not want to duel when he wants to duel. In the words of Mark Twain: "Well, out there, if you abused a man, and that man did not like it, you had to call him out and kill him; otherwise you would be disgraced. So I challenged Mr. Lord, and I did hope he would not accept; but I knew perfectly well that he did not want to fight, and so I challenged him in the most violent and implacable manner.' (Twain, 1872) The main point of the analysis can be obtained with this alternative specification as long as the following holds: An individual who does not want to fight does not want to miscommunicate if the other player also does not want to fight.

I say that k *pools on* $x \in \{0,1\}$ *at history* h_k if she chooses x no matter her type: $\alpha_k(h_k, \phi_k) = x$ for all ϕ_k. I call k a *withholder at history* h_i if one of two conditions hold: (a) k has already made a decision ($k < i$) and she pooled, or (b) k has not made a decision ($k > i$). I say i *separates* at history h_i if her types make different decisions: $\alpha(h_i, \phi_i) \neq \alpha(h_i, 1-\phi_i)$. I call k a *revealer at history* h_i if k has already made a decision ($k < i$) and she separated. If it does not lead to ambiguity, I will not mention the history when referring to revealers, withholders, separating and pooling.

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. 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 and informative Perfect Bayesian Equilibrium satisfying the intuitive criterion for refining out-of-equilibrium beliefs.*

As a consequence of requiring the equilibrium to be informative, I select a separating equilibrium whenever it exists. This allows me to obtain results that are biased against information stagnation (i.e., subjects not revealing their type).

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. 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 pursue what they mistakenly believe is socially optimal.

3.1 Equilibrium Dynamics

I first define informative and uninformative contexts, which are subsets of χ I will focus on. I then introduce some terms with which to characterize equilibrium dynamics.

3.1.1 Informative and Uninformative Context

I begin with a more specific way of thinking about context. I'll focus on contexts where individuals are very certain or very uncertain of the majority preference.

Definition 2. *The context is* ***informative*** *if and only if, at the beginning of the game, player i of type 0 strongly believes most judges are of type 1:*

$$\mathbb{E}_{j\neq i}\big(P(\phi_j = 0 \mid h_1, \phi_i = 0)\big) < \frac{\beta - 1}{2\beta} \in (0, 1/2)$$

The context is ***uninformative*** *if and only if, at the beginning of the game, player i of type $x \in \{0, 1\}$ believes most judges are of type x:*

$$\mathbb{E}_{j\neq i}\big(P(\phi_j = 1 \mid h_1, \phi_i = 0)\big) < \frac{1}{2} < \mathbb{E}_{j\neq i}\big(P(\phi_j = 1 \mid h_1, \phi_i = 1)\big)$$

By the symmetry of the game, the definition of an informative context is without loss of generality – a context where player i of type 1 would believe most judges are of type 0 is defined analogously, and yields similar results.

Uninformative and informative contexts are of particular interest, since they respectively capture high uncertainty and certainty about the distribution of preferences from which the group was drawn. Furthermore, equilibrium strategies are pure in informative and uninformative contexts. If the context is neither informative nor uninformative, then mixed strategies may arise which substantially complicate the model but add little insight.

3.1.2 Characterizing Equilibrium Dynamics

In order to keep track of the information that has been revealed by players' actions, I follow Ali and Kartik (2012) and define the *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 lead for action 0.

Result 1. *Suppose $\beta \in [0, 2]$ and $\pi > 3/4$. If the context is informative, pooling on 1 is the unique equilibrium strategy.*

If the context is uninformative, player 1 separates. Suppose that along the equilibrium path of play, all players $k < i$ have separated. Then there is some threshold $n > 0$ such that if the lead for action $a \in \{0, 1\}$ is equal to n, pooling on a is the unique equilibrium strategy for player i and all players $l > i$. The threshold n is finite, decreases in β, χ and in π.

The result shows that there are two types of dynamics, depending on the context.

When the context is informative, the first speaker will strongly believe most judges are of type 1 independent of her private view. The belief is strong enough that she will pool on 1 in order to be judged positively. Since player 2 does not learn anything from player 1's action, she will have the same incentives as player 1 and will therefore also pool on 1. So will all other players.

When context is uninformative, players at the beginning of the game rely heavily on her private view to form expectations about the group. Judges and decision makers believe most group members are of the same type as they are. Therefore, player 1 believes that if she reveals her type, most judges will believe preferences match. Player 1 therefore separates, which gives player 2 higher incentives to pool on player 1's preference. Once enough players create a sufficiently large lead for some action, all players pool on that action. For example, suppose the weight on social expectations β is 2, the context χ is 0.5, and the precision π is 0.8. Then a lead for action x of 3 ($\Delta(h_i)$ equal to $+3$ or -3) results in all players pooling on x. If instead the precision is π is 0.9, all players pool on x with a lead of 2.

3.2 Pluralistic Ignorance in Large and Small Groups

Result 1 described equilibrium dynamics for informative and uninformative contexts. 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 ($\beta = 0$). I can illustrate these definitions with the example of white segregationist preferences I presented in section 2.2.1. 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. *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 probability that types are realized in a way that results in pluralistic ignorance.

Result 2. *Suppose $\beta \in (0,2)$ and $\pi > 3/4$. There exist thresholds $\tau_1(\chi,\pi,\beta) \geq 2$ and $\tau_2(\chi,\pi,\beta) < \infty$ with the following characteristics:*

- *Suppose the context is informative, $|I| \geq \tau_2$, and $\chi > p$ for some p which decreases with $|I|$. Then the probability of pluralistic ignorance is lower than for any other context, and is lower than if $|I| \leq \tau_1$.*
- *Suppose the context is uninformative and $|I| \leq \tau_1$. Then the probability of pluralistic ignorance is lower than for any other context, and is lower than for $|I| \geq \tau_2$.*

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 χ, or $P(\theta)$, 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 approximates the majority preference of the population θ – the probability of pluralistic ignorance becomes vanishingly small as group size and strength of beliefs over θ increases. 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 would rely on few opinions 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 (as presented in section 2.2.1). The condition $\gamma > 0$ 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 coopera-*

tive 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. *Suppose* $\beta \in (0, 2)$ *and* $\pi > 3/4$*. Individuals will be in cooperative pluralistic ignorance with a positive probability if (a)* $\chi > 1/2$*, the context is uninformative, and I sufficiently large, or (b) the context is informative. Cooperative pluralistic ignorance 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 if the context χ is 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.[4] In the next section I will argue that this situation captures an important aspect of pluralistic ignorance.

4 From Public Opinion to Collective Action

Sections 2 and 3 focused on the motivating example of white segregationist preferences for ease of exposition. I focused on the question of when misperceptions of public opinion arise

[4]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.

in micro and macro interactions, and I showed how pluralistic ignorance can give rise to a perverse failure of collective action. This result provides a first approach to using the framework as a link between public opinion and collective action.

I will now turn to analyzing social change more generally. I will address the questions: How is social change effected in small group interactions? How is change effected in large group interactions? What explains abrupt social change? To transition to these questions, 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 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. *Player k of type ϕ_k considers behavior $a \in \{0,1\}$ after history h_i to be **appropriate** if and only if she believes it makes most judges believe preferences match in equilibrium: $P(\mathcal{J}^*_{j,i}(\alpha^*_i = a) = 1 \mid h_i, \phi_i) > 1/2$.*

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. 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.[5]

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, only if individuals pool on an action is that action a norm. Moreover, a norm would emerge the first period an individual pools. 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 emphasized the importance of thinking about endogenous preference formation in social analysis (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.

4.2 Alternative Approaches to Pluralistic Ignorance

I now present a more detailed review of the literature on pluralistic ignorance, and consider an alternative approach to the model. I will argue that past formalizations do not capture the features of pluralistic ignorance that the second-order conformity model does.

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

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

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 underrestimation of reluctance leads individuals to act reluctantly. 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 ohters' threshold for action (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' threshold for action 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 which action they would all prefer to coordinate on if the threshold for action were low enough. Further, these models often focus on a continuum of players (e.g. Angeletos, Hellwig and Pavan, 2006, Morris and Shin, 1998), making the question of group size moot.

Some authors, such as Kuran (1997), Lohmann (1994), Chwe (2000) have studied social change in situations of pluralistic ignorance. The paper follows in this tradition, and I will emphasize policy implications and testable hypothesis that are unique to the second-order conformity approach to pluralistic ignorance.

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, Johannesson et al., 2008).[6] Unlike with second-order conformity, the preference that provides

[6]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.

reputational benefits is exogenously given. Since there is common knowledge about what this preference is, individuals will not incorrectly believe others are acting reluctantly.[7] 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, this approach seems unsatisfying from both a conceptual and an empirical perspective. Conceptually, it is hard to motivate why altruistic individuals would not simply reveal their types so that they can choose the optimal action. Empirically, 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, Weber and Kuang, 2007, Dana, Cain and Dawes, 2006).

For example, consider the experiment by Dana, Cain and Dawes (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. Although I remain agnostic as to why individuals in the model care about being judged, the experiment suggests that individuals care absent reputational concerns.

[7] 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 is the distribution of preferences across individuals.

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 in the sixties will be more pro-segregation at a Southern university than at a Northern University.

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 a discussion on the evidence of pluralistic ignorance in large and small groups. In section 4.4.2 I argue that the model can be used to study norm variation across groups. This will link the model to a question of social change in small group interactions. In section 4.4.3 I provide a further example of how second order conformity can analyze small group interactions. I argue inadvertent wars are the result of states acting according to what they think other states want, with an explicit model provided in appendix K. 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) regarding large groups, discuss some qualitative evidence regarding small groups, 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. Consider evidence 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.

Qualitative Evidence for Opinion Formation in Small Group Interactions. White students in the beginning of the 60's had very different experiences in conversations over segregation in the North and South of the U.S. In the North, where individuals perceived a growing acceptance of anti-segregationist views (O'Gorman, 1975), many white anti-segregationist student groups formed. In contrast, these types of groups were much slower to form in the South (Turner, 2010). This was not only because anti-segregationist views were more widespread in the North. Anti-segregationist whites in the South were often afraid to speak up. Consider the following quote:

> In the early days of the Atlanta student movement, Constance Curry, Southern Project director for the U.S. National Student Association (NSA), tried to recruit students from Atlanta's white colleges for sit-ins. "Somehow or other I scraped up a white representative from every college, even Georgia Tech," she later recalled. "They only came to one meeting because they were terrified. . . . [T]hey took one look at [Morehouse student] Lonnie King and all those students and never

said a word and went home that night and we never heard of 'em again." Turner (2010)

The Southern context – the belief that most whites supported segregation – silenced anti-segregationist whites. As demonstrations of support for integration increased, however, white anti-segregation groups started to form. Consistent with the model, diversity of opinions increased in small group interactions when individuals were less certain about others' anti-segregationist preferences.

Experimental Design. I have given some qualitative examples of how in small groups, an informative context may lead individuals to pluralistic ignorance. However, we do not have much experimental evidence for the implication of Result 2 on small groups. Here I sketch a 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, Cialdini and Kallgren, 1993), voting (Gerber and Rogers, 2009), protesting (McClendon, 2014), tax compliance (Wenzel, 2005), or as I've argued, dueling. Therefore, the model can be thought of as a stylized approach to norm emergence in large and small groups. 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 2 is that the distribution of norms across groups will vary much more when the context is uninformative than when the context is informative. We will use this interpretation to think of social change. For example, when considering dueling, we will be interested in the question of how it was that interactions of pairs of slighted elites would often result in duels in the eighteenth and nineteenth century, when dueling was the norm in these interactions, but disappeared by the twentieth century.

4.4.3 Second Order Conformity in Interstate Conflict Bargaining

I have argued that second order conformity arises in dueling, and provides a useful framework for understanding the success of policies intended to diminish its frequency. Here I extend the logic to another situation of conflict: that of states deciding whether to go to war.

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 appendix K 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 when the context is informative. 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.[8] In contrast,

[8]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."

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 (Wendt, 1992) – analogous to Result 1 in which the first movers reveal their type when the context is uninformative.

4.4.4 Examples Outside of Politics

In this section I informally present some examples of social phenomena which can be thought through the framework 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. Since they have uncertainty about what the other person wants, we think of this as a problem of second order conformity. Assume they are uncertain as to whether the other is most likely interested (in the model, 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 context 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).

5 The Impact of Principals and Policies on Social Change

I have presented the basic model in sections 2 and 3, and argued in section 4 that it can be applied to several situations. 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? Through considering different policies and principals, I will address how policies led to a decline in dueling, and why social information campaigns have had such mixed success.

In subsection 5.1, I consider an informed and trustworthy principal who makes an announcement with the objective of minimizing pluralistic ignorance. I will consider two mo-

tivating examples. First, I consider a teacher who would like the class to engage in a lively discussion on a controversial topic. Second, I consider a campaign to diminish dueling by arguing over the preferences of 'true' gentlemen. I show that principals cannot always avoid pluralistic ignorance 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 how a law may 'regulate' social meaning by affecting what a group considers appropriate behavior ('appropriateness' and 'social meaning' are defined in section 4.1). 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. Although the first four subsections can be read somewhat independently, the discussion sections discuss the relationships between the predictions of having different principals and policies. Subsection 5.5 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 'Pronouncement' period 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.

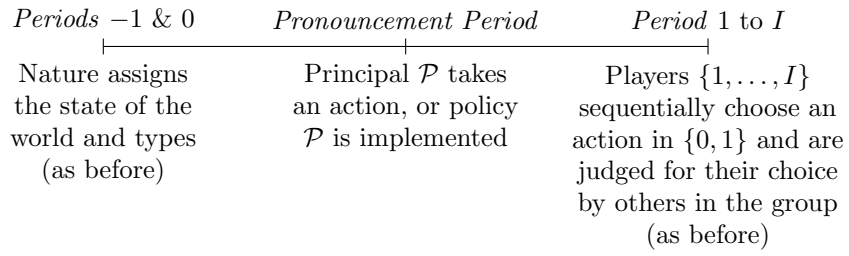

Figure 2: Timeline with principal or policy $\mathcal{P}$

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 examples I have 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 is 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. This fits nicely into the model if we imagine a conversation in which students take turns announcing their preference.

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.

As a second motivation, consider a campaign against dueling that took place in London (Shoemaker, 2002). The tactics used by abolitionists was to argue that dueling promoted a 'false' honor. A true gentleman, they argued, was compassionate towards others. This campaign may be thought of as providing information about what was the distribution of preferences in the population of gentlemen, with the objective of convincing those who duel that they would not be seen well if they displayed preferences for dueling.

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.[9] 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

[9] 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.

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. To think of this commitment problem, suppose that before the principal sees the population, she can announce which action she will follow after seeing the population. After she sees the population, she then follows a strategy. I'll say $\mathcal{P}$ *can commit* if she has to follow the strategy she announces.

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 τ_1 and τ_2 from Result 2.*

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

- *Suppose the group size is large, $|I| \geq \tau_2$. 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 informative. If the principal wishes to increase uncertainty about the context, her message must be informative about the population. But then 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 informative. 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

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.

The London campaign to abolish dueling was not trying to just impede pluralistic ignorance, as the setup assumes. The model can be easily extended to show that abolitionists may have a commitment problem analogous to the one we discussed. Their objectives will weaken their message if it makes them more willing to say whatever decreases dueling Lupia and McCubbins (1998). However, abolitionists' campaign seems to have succeeded in diminishing support for dueling. The context shifted from one in which the elite strongly believed most valued dueling to one in which there was more uncertainty about what the elite believed (Shoemaker, 2002). As predicted by the model, dueling remained mostly among strong supporters of the practice, who would duel in private in part to avoid social judgment.

The two motivating examples I provided, of conversations among students and of dueling, show that the logic applies to both opinion formation and collective action problems. However, they are both small group examples. A large group example may be found in 'norm entrepreneurs' Finnemore (1996), Finnemore and Sikkink (1998), who start social movements by informing a group what a 'true' group member values. For example, Gandhi argued that a 'true' Indian would seek independence from British rule through non-violent protest (Dalton, 2012). Henry Dunant argued that a 'true' Christian nation did not leave those hurt by war unattended (Finnemore, 1996). Both of these examples can be thought of through the model. Individuals will take turns choosing an action (whether to non-violently protesting, whether to attend the wounded in war), and they have a preference over this action. Once again, the model assumes that individuals trade off choosing their ideal point with how they will be judged by group members (Indians, Christian nations). The norm entrepreneur then has private information about the group's preference.

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. One could object that neither Gandhi nor Henry Dunant were informing the group about the distribution of preferences in the group, but rather about the distribution of preferences the group 'should' have. Indeed, both non-violent protest and attending the wounded in war did not seem to be what the majority preferred before the norm entrepreneurs. One way to address this concern is to reinterpret the population from which the group is drawn as the distribution of preferences the group would have in the absence of some constraint. For example, it is the preference Indians would have if they were better informed, or less oppressed. This line of thought is a significant break from the model, so I will not pursue it. A second way to address this problem is to argue that these norm entrepreneurs are creating a link between a behavior and how others value them. I will return to this interpretation below.

Imperfect information about the population distribution of preferences will lead to pluralistic ignorance. A principal may therefore consider 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, Reno and Kallgren (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 having subjects interact with a surveyor, such as with energy consumption (Allcott, 2011), contributions to public parks Alpizar, Carlsson and Johansson-Stenman (2008), or past votes (Gerber and Rogers, 2009), 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 the Pronouncement period, a principal surveys each player privately about their ideal point, and then reports the mean response $\mu \in [0, 1]$ to the group. I further modify the model so that the principal has a preference 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, the Pronouncement period is subdivided into three. First, $\mathcal{P}$'s type is drawn by Nature. Second, $\mathcal{P}$ surveys the group. Third, $\mathcal{P}$ reports the mean response μ 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.

To give a few examples of the actions that are taken at the sequential decision stage, individuals may be deciding whether to contribute to a public park, vote, or express their opinion on a topic.

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 mean response μ back to the group at the Pronouncement period. 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: $P(\phi_k \mid h_i, \phi_j, \mu) = P(\phi_k \mid h_i, \phi_j)$ 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.

To the best of my knowledge, we do not have experimental evidence on whether perceptions of the presence of surveyor demand effects impacts the effectiveness of a social information campaign. However, this can be manipulated by varying what the surveyor knows about the conditions under which surveyors were asked questions. A simple example of this would be to run an anonymous survey, and vary whether the social information campaign reports that surveys were done anonymously.

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 if respondents are motivated to conform to the surveyor's judgment.

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.

Although Result 5 does not depend on group size, social information campaigns are most naturally thought of as policies to address behavior change in large groups. If pluralistic ignorance depended on the distribution of preferences of a small group interaction, then a social information campaign would have to survey and report back to each small group. A more suitable policy for small group interactions may be to propose mechanisms through which group members could determine the distribution of preferences in the group without being judged negatively by others. Below, I will argue that this was an important function of honor codes in dueling.

5.3 Regulating Social Meaning

I now turn to a type of policy that changes a behavior's social meaning (as defined in Section 4.1).

5.3.1 Motivation

Consider the norm for dueling among the elite in the U.S., whose social inefficiency has been argued by Lessig (1995), Schwartz, Baxter and Ryan (1984). Dueling was considered by many to be inefficient: a duel could be started for a petty reason, the result could be death, and the person who died need not be the aggressor (Schwartz, Baxter and Ryan, 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.

[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 pool.

5.3.2 Setup

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, such as announcing an intention to duel. The second preference is over some seemingly unrelated issue. For example, $\psi_i \in \{0, 1\}$ is i's views serving in public office, with $\psi_i = 1$ indicating a favorable view. Preference ψ_i is uninformative about the population θ. Individuals' preference ψ_i is not observable, but there is a commonly known probability $\hat{\pi}$ that a randomly chosen individual is of type $\psi_i = 1$.

I will call $a = 1$ the action of 'dueling', and $a = 0$ 'not dueling'. Before group members make their sequential decisions, 'lawmaker' $\mathcal{P}$ may link dueling and holding public office by requiring action $a = 0$ to be taken to be able to hold office. If the lawmaker establishes a link, she sets $l = 1$, and $l = 0$ otherwise. The lawmakers chooses l to minimizes the amount of dueling, or minimize the expected sum of a_i for all i.

Group members' utility function is a sum of two components. The first component is the original utility function. The second component matters only if $\mathcal{P}$ establishes a link, and captures supporters' concerns for holding public office. Expected utility is then:

$$\underbrace{\mathbb{E}u(a_i; \phi_i)}_{\text{Original utility function}} + l\delta\big[-\psi_i a_i + \beta \mathbb{E}_{\hat{\psi}_{-i}}(\mathbb{E}_{j \neq i}(\mathcal{K}_{j,i} \mid \hat{\psi}_{-i}))\big] \tag{4}$$

The term in square brackets is analogous to the original utility function, and is given weight $\delta > 0$. Those who favor holding public office want to avoid dueling (first summand in the square brackets). All players want others to think their preference over holding public office matches theirs. The expectation $\mathbb{E}_{\hat{\psi}_{-i}}$ sums over the possible distribution of views regarding holding public office. Judge j decides whether i's preference over holding public office matches her own, analogous to before. The judgment function $\mathcal{K}_{j,i} \in [0, 1]$ is equal to one if j believes i's preference over holding public office most likely matches her own, or $\mathcal{K}_{j,i}(a_i) = 1$ if $P(\phi_i = x \mid h_i, \alpha_i, \phi_j = x) > 1/2$; it is equal to zero if j believes i's preference most likely does not match her own, or $\mathcal{K}_{j,i}(a_i) = 0$ if $P(\psi_i = x \mid h_i, \alpha_i, \psi_j = x) < 1/2$; it takes an arbitrary value between 0 and 1 otherwise, or $\mathcal{K}_{j,i}(a_i) \in [0, 1]$ if $P(\psi_i = x \mid h_i, \alpha_i, \psi_j = x) = 1/2$.

In the result below, I will consider the probability of pluralistic ignorance as defined originally. That is, I will only consider pluralistic ignorance in preferences over dueling (over ϕ_i, or the first dimension of preferences).

5.3.3 Result

Result 6. *Suppose a principal $\mathcal{P}$ decides whether to link behavior $a \in \{0,1\}$ with a second behavior, by restricting the second behavior for those who choose $a = 1$. The principal $\mathcal{P}$ chooses whether to link behavior in order to minimize the expected sum of action a across all players. Individual i has preferences over action a, given by $\phi_i \in \{0,1\}$, and over the second behavior, given by $\psi_i \in \{0,1\}$. Individual i's utility function is given by* (4)*.*

If there is a strong belief others in the group favor the second behavior ($\hat{\pi}$ large enough), and group members place a lot of weight on the link established by $\mathcal{P}$ (δ large enough), then it is an equilibrium for $\mathcal{P}$ to establishes a link and all players pool on $\psi_{\mathcal{P}}$. In this equilibrium:

- *The probability of pluralistic ignorance diminishes if most group members prefer to not duel ($\sum_{i \in I} \phi_i < 1/2$) and the context is uninformative.*
- *The probability of pluralistic ignorance increases if most group members prefer to duel ($\sum_{i \in I} \phi_i > 1/2$).*

If most in the group favor holding office and the lawmaker establishes a link, then group members will believe there are added social reasons to not duel – either because they value holding public office or because they want most group members to judge them favorably.

Suppose in the absence of a law linking behavior, group members would all choose to duel (such as when the context is informative). If most in the group favor holding public office and the lawmaker establishes a link between holding office and dueling, then it is no longer the case that all group members will duel. Individuals may value holding public office enough to avoid dueling even if every other type preferred to duel (this requires δ to be large enough). But then the social meaning of dueling is 'ambiguated' (Lessig, 1995), since now some types whose preferences over holding public office match the group would not duel, which makes it less socially costly for those who prefer to not duel to avoid doing so. Intuitively, duelers may get out of dueling by saying 'I would love to fulfill my gentlemanly duty to duel you, but that would impede me from fulfilling my more important gentlemanly duty to hold public office.'

5.3.4 Discussion

By making an action a signal of whether a player values holding public office, the 'social meaning' of dueling changes (see subsection 4.1 for a discussion of social meaning). Once group members believe that some will avoid dueling in order to hold public office, not dueling signals something different. The lawmaker will therefore be able to impact behavior not by informing the group about the group's preferences, but by linking a behavior that the

lawmaker wants to avoid to a second behavior the group values. However, enough group members must value the second behavior sufficiently for it to have an impact.

Many laws seeking to ban dueling failed to make the elite ignore their social responsibility to duel. If an elite was not convinced that a 'true' elite would avoid dueling to avoid a fine or jail time, not dueling continued to signal not being an elite. Banning those who duel from holding public office was successful because it was something the elite highly valued.

The challenge of this type of policy is that it may also lead to pluralistic ignorance: two individuals who want to duel may now avoid doing so because they think the other does not. Indeed, some of these concerns were reflected in the uneven application of the law. Although lawmakers were interested in diminishing the extent of dueling, they were reticent to apply the law to those who continued to cherish the practice (Williams, 2000).

There is a growing body of work on the expressive function of the law, which do not emphasize how laws may regulate social meaning through linking behaviors. Benabou and Tirole (2012) shows how the law can provide information about how *intensely* a population values an exogenously given trait. For example, in a society where all agree that litterers are not valued, the law provides information about just how much dislike there is for littering. 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.

We can also use this setup to think about norm entrepreneurship. In section 5.1 I discussed how Gandhi mobilized individuals to non-violent resistance (Dalton, 2012) and Henry Dunant mobilized states to tend to the wounded in war (Finnemore, 1996). They did so by defining what a 'true' Indian or a 'true' Christian nation did. I argued that one way to interpret what they did was that they were providing information about the distribution of preferences. However, I pointed out that interpretation is not entirely satisfactory. Another way to interpret what the were doing is that they were regulating social meaning in an analogous way to what lawmakers did to decrease dueling. For example, Gandhi may have been establishing a link between himself as a highly valued Indian to the act of non-violent protest.[11] Thus, those who did avoided non-violent protest were signaling that they did not value Gandhi, thus revealing that their preferences did not conform to those of other Indians. Consider one implication of the difference between an informed leader and a social meaning regulator. An informed principal will be more effective the more group members learn about

[11]Now individual i has a preference over protesting given by ϕ_i, and a preference over Gandhi given by ψ_i. Gandhi is the principal $\mathcal{P}$, who can establish a link between the decision to non-violently protest a_i and himself. In effect, he is saying that those who do not peacefully protest do not support me.

others' preferences from her messages, and less about the principal's message. In contrast, a social meaning regulator will be more effective the more group members learn about her preference.

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.[12]

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, Willer and Macy, 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

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

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 coordination 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

Individuals 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

	Good regime	Bad regime
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	Has no clothes
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)

person is someone whose preference match those of the majority in the group.[13] Thus, if the regime is 'good' – that is, at least 50% plus one citizens support it – then a benign principal favors the regime ($\phi_{\mathcal{P}} = 1$). Otherwise, the benign 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 and there will be zero probability of pluralistic ignorance*

[13] 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.

for any χ 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 abrupt 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, Herron and Shotts, 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).[14] 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.

I now consider other candidates for uninformed leadership.

[14]The reader may be wondering whether an uninformed catalyst may be modeled in a standard herding model such as Banerjee (1992) or Bikhchandani, Hirshleifer and Welch (1992). In those models, individuals all have the same preference, but don't know what the best action is that satisfies that preference. Individuals receive a private, informative signal about the state of the world, and the state of the world determines their optimal choice. Therefore, in order for an individual to be 'benevolent', we would have to say that she observes the true state of the world. This already strains the story we would like to capture, both because this suggests the individual is informed and because it suggests individuals face no social pressure to hide their private view. We can ignore the second objection, and further say that the individual is uninformed if only others observe her benevolence. However, the analogy to brashness is also strained. If an individual does not know she is benevolent, she would not rationally follow her private information in a context that is strongly ϕ. We could perhaps say that a brash individual mistakenly believes that the context is uninformative, but examples such as Bouazizi seem to be only too aware of the social pressure to acquiesce to the regime.

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 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. As suggestive evidence of 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.' As suggestive evidence 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 Discussion

I close this section by by discussing codes of honor in dueling, and considering the role of common knowledge in the second-order conformity model.

Honor codes. Honor codes carefully regulated duels. The Wilson honor code was an

important code in the U.S., which required the aggrieved to privately send a note to the aggressor through an intermediary (Williams, 2000). The aggressor had an opportunity to apologize, also privately. If he didn't, his intermediary would work with the other intermediary to figure out if the duel could be called off on the grounds that it was based on a misunderstanding. That is, the honor code was designed to provide an opportunity to avoid a duel if there was pluralistic ignorance. In fact, the person who wrote the honor code explains in the introduction that his motive to write it was not because he favored duels, but because he wanted to make sure there was a way to get out of it. The popularity of the honor code is attributed to this feature (Williams, 2000, O'Neill, 2003).

Common knowledge. 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, Ho and Chong, 2004). More importantly, the difference in predictions over the importance of higher order believes allows for some leverage to test between theories.

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. 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

Abouzeid, Rania. 2011. "Bouazizi: The Man Who Set Himself and Tunisia on Fire." *TIME* .

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

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

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

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

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

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

Andrew, Donna T. 1980. "The code of honour and its critics: The opposition to duelling in England, 1700–1850." *Social history* 5(3):409–434.

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

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

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

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

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

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

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

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

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

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

Buckley, M Ronald, Michael G Harvey and Danielle S Beu. 2000. "The role of pluralistic ignorance in the perception of unethical behavior." *Journal of Business Ethics* 23(4):353–364.

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

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

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

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

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

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

Cialdini, Robert B, Raymond R Reno and Carl A Kallgren. 1990. "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.

Cukierman, Alex and Mariano Tommasi. 1998. "When does it take a Nixon to go to China?" *American Economic Review* pp. 180–197.

Dalton, Dennis. 2012. *Mahatma Gandhi: Nonviolent power in action*. Columbia University Press.

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

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

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

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

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

Ellingsen, Tore, Magnus Johannesson et al. 2008. "Pride and Prejudice: The Human Side of Incentive Theory." *American Economic Review* 98(3):990–1008.

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

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

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

Fernández-Duque, Mauricio and Michael Hiscox. 2017. "Leadership and Social Expectations." *unpublished manuscript* .

Finnemore, Martha. 1996. "National interests in international society.".

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

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

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

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

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

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

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

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

Halbesleben, Jonathon RB, M Ronald Buckley and Nicole D Sauer. 2004. "The role of pluralistic ignorance in perceptions of unethical behavior: An investigation of attorneys' and students' perceptions of ethical behavior." *Ethics & Behavior* 14(1):17–30.

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

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

J O'Gorman, Hubert. 1986. "The discovery of pluralistic ignorance: An ironic lesson." *Journal of the History of the Behavioral Sciences* 22(4):333–347.

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

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

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

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

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

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

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

Light, Richard J. 2015. "How to Live Wisely." *The New York Times* .

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

Lohmann, Susanne. 1994. "The dynamics of informational cascades: The Monday demonstrations in Leipzig, East Germany, 1989–91." *World politics* 47(1):42–101.

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

Majumdar, Sumon and Sharun Mukand. 2008. "The Leader as Catalyst-On Leadership and the Mechanics of Institutional Change.".

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

Maskin, Eric and Jean Tirole. 2004. "The Politician and the Judge: Accountability in Government." *The American Economic Review* 94(4):1034.

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

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

McClendon, Gwyneth H. 2014. "Social esteem and participation in contentious politics: A field experiment at an LGBT pride rally." *American Journal of Political Science* 58(2):279–290.

Mildenberger, Matto and Dustin Tingley. 2016. "Beliefs about Climate Beliefs: The Problem of Second-Order Climate Opinions in Climate Policymaking." *unpublished manuscript* .

Morris, Stephen and Hyun Song Shin. 1998. "Unique equilibrium in a model of self-fulfilling currency attacks." *American Economic Review* pp. 587–597.

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

Noelle-Neumann, Elisabeth. 1974. "The spiral of silence a theory of public opinion." *Journal of communication* 24(2):43–51.

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

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

O'Neill, Barry. 2003. "Mediating national honour: lessons from the era of dueling." *Journal of Institutional and Theoretical Economics JITE* 159(1):229–247.

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

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

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

Schanck, Richard Louis. 1932. "A study of a community and its groups and institutions conceived of as behaviors of individuals." *Psychological Monographs* 43(2):i.

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

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

Shoemaker, Robert B. 2002. "The taming of the duel: masculinity, honour and ritual violence in London, 1660–1800." *The Historical Journal* 45(3):525–545.

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

Sliwka, Dirk. 2007. "Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes." *American Economic Review* 97(3):999–1012.

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

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

Taylor, Justin. 2015. "5 Myths about Rosa Parks, the woman who had almost a 'biblical quality'." *The Washington Post* .

Theoharis, Jeanne. 2015. "How history got the Rosa Parks story wrong." *The Washington Post* .

Turner, Jeffrey A. 2010. *Sitting in and speaking out: student movements in the American South, 1960-1970*. University of Georgia Press.

Twain, Mark. 1872. "How I Escaped Being Killed in a Duel.." *Bangor Daily Whig & Courier* 23.

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

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

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

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

Williams, Jack K. 2000. *Dueling in the old South: Vignettes of social history*. Texas A&M University Press.

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

A Proof of Result 1

In order to establish the proof, I will work my way towards establishing the following:

Claim 1. *Player i of type x deviates from a separating strategy if and only if:*

$$\mathbb{E}_{j\neq i}\big(P(\phi_j = x \mid h_i, \phi_i = x) \mid h_i\big) < \frac{\beta - 1}{2\beta} \tag{5}$$

Suppose $\beta \in (0, 2)$. If this condition holds in an uninformative or informative context, pooling on $1 - x$ is a unique strategy of a Perfect Bayesian equilibrium with an intuitive criterion refinement.

Inequality (5) follows from considering the utility of type $\phi_i = x$ under a separating strategy. Only judges of type x believe preferences match if i chooses x.[15] Therefore, type x compares the social expectation from choosing x and having only players of type x believe preferences match, or choosing $1-x$ and having only players of type $1-x$ believe preferences match. This leads to the threshold rule (5).

If it is an equilibrium strategy for player i to pool, player $i + 1$ will have the same information about the group that player i did. Player $i+1$ will then face the same incentives as player i, so it will also be an equilibrium strategy for $i + 1$ to pool. In fact, if (5) holds at the beginning of the game (when i is equal to 1), we can conclude from Claim 1 that all players will pool – this condition will characterize an informative context. Then we just need to consider the history of play that leads a player to deviate from separating in an uninformative context, knowing that all future players will pool in equilibrium.

To calculate a player's social expectations, we need to know how a decision maker forms her beliefs over how she is judged by others. A player j who has already made a decision will continue to judge others, but may have revealed her preference with her decision. Therefore, decision maker i's beliefs about how j will judge her will depend on whether j has separated. Let $\phi_j^o : H \to \{-1, 0, +1\}$ be j's *observed type*, a map from history h_i to the value -1 if j has revealed type 0, to the value $+1$ if j has revealed type 1, and to the value 0 if j is a withholder. Note that the observed type of a withholder j is indicated by $\phi_j^o = 0$, which does not specify her type $\phi_j \in \{0, 1\}$. This notation emphasizes that a withholder's type is not observed to player i.

When it is useful, I will explicitly condition j's belief on her observed type ϕ_j^o, for example I will write $P(\phi_i = 1 \mid h_i, \phi_j, \phi_j^o)$. Although the conditioning on ϕ_j^o is implicit in the expression

[15] Recall the terminology for talking about social expectations, or $\mathbb{E}_{\hat{\phi}}\big(\mathbb{E}_{j\neq i}\big(\mathcal{J}_{j,i}|\hat{\phi}\big) \mid h_i, \phi_i\big)$: judge j 'believes preferences match' if she thinks it is more likely that i's type is the same as hers, or $P(\phi_i = x \mid h_i, \phi_j = x) > 1/2$.

$P(\phi_i = 1 \mid h_i, \phi_j)$, it will sometimes make it easier to keep track of the information j has when forming beliefs.

The judges' and decision makers' decisions will depend on their beliefs. When decision maker i pools, judge j's judgment of i will depend on her beliefs over i's most likely type. In turn, decision maker i's decision will depend on her belief over the distribution of judges' types. To capture what a player k knows about the group, let the *observed type lead* be $\Delta^o(h_i, \phi_k) \equiv \sum_{l \neq k} \phi_l^o(h_i) + (2\phi_k - 1)$, or the difference between the number of type one and type zero signals observed by player k of type ϕ_k at history h_i.

Lemma 1. *Suppose at history h_i, players either separated or are withholders.*

Judge j will believe a withholder i is of type 1 with probability at least z, or $P(\phi_i = 1 \mid h_i, \phi_j) > z$, if and only if:

$$f_j\big(z, \phi_j^o, \textstyle\sum_{l \neq j} \phi_l^o\big) \equiv \left(\frac{\pi}{1-\pi}\right)^{\Delta^o} [\pi - z] - \frac{1-\chi}{\chi} [z - (1-\pi)] > 0$$

Decision maker i will believe the average judge is of type 1 with probability at least z, or $\mathbb{E}_{j \neq i}\big(P(\phi_j = 1 \mid h_i, \phi_i) \mid h_i\big) > z$, if and only if:

$$g_i\big(z, \phi_i^o, \textstyle\sum_{l \neq i} \phi_l^o\big) \equiv \left(\frac{\pi}{1-\pi}\right)^{\Delta^o} \big[\mathbb{E}_{j \neq i}\big(P(\phi_j = 1 \mid h_i, \theta = 1) \mid h_i\big) - z\big] -$$

$$\frac{1-\chi}{\chi} \big[z - \mathbb{E}_{j \neq i}\big(P(\phi_j = 1 \mid h_i, \theta = 0) \mid h_i\big)\big] > 0$$

Expressions $f_j > 0$ and $g_i > 0$ in Lemma 1 characterize the beliefs that shape decision makers' and judges' choices, in terms of primitives and the observed type lead. The first multiplicand of f_j and of g_i is the likelihood ratio of the precision π, to the power of the observed type lead Δ^o. This term captures the information an individual has about the population based on the signals she has observed. Expressions f_j and g_i also depend positively on the context χ, or the prior probability that the population is $\theta = 1$. The expressions only differ in the terms in square brackets.

Lemma 2. *The terms f_k and g_k increase in ϕ_k^o, $\sum_{l \neq k} \phi_l^o$, and χ.*

Suppose players j and i have the same type ($\phi_j = \phi_i$) and j is a withholder. Then $f_j(1/2, \phi^o, y)$ is greater, equal or less than $g_i(1/2, \phi^o, y)$ if y is respectively greater, equal or less than zero.

The context is uninformative if and only if $g_1(1/2, 0, 0) < 0 < g_1(1/2, 1, 0)$. The context is informative if and only if $g_1((\beta - 1)/2\beta, 0, 0) < 0$, or equivalently (5) holds for $i = 1$.

I now show that when the weight on social expectations $\beta \in (0, 2)$, either a separating strategy exists, or there is at most one pooling strategy that is part of a Perfect Bayesian equilibrium with the intuitive criterion refinement (I will refer to such an equilibrium as a 'PBE', and to such a strategy as a 'PBE strategy'). When $\beta \in (0, 1]$, players always prefer to choose their ideal point. When $\beta \in (1, 2)$, a type chooses her ideal point x if and only if the difference in social expectation from choosing x versus choosing $1 - x$ is less than $1/2$:

$$\mathbb{E}_{\hat{\phi}_{-i}}\big(\mathbb{E}_{j\neq i}(\mathcal{J}_{j,i}(x) \mid \hat{\phi}_{-i}) \mid h_i, \phi_i\big) - \mathbb{E}_{\hat{\phi}_{-i}}\big(\mathbb{E}_{j\neq i}(\mathcal{J}_{j,i}(1-x) \mid \hat{\phi}_{-i}) \mid h_i, \phi_i\big) < \frac{1}{\beta} \in \left(\frac{1}{2}, 1\right)$$

By Lemma 2, a type 1 player believes more of her judges are of type 1 than does a type 0 player: $\mathbb{E}_{j\neq i}\big(P(\phi_j = 1 \mid h_i, \phi_i = 1) \mid h_i\big)$ is larger than $\mathbb{E}_{j\neq i}\big(P(\phi_j = 1 \mid h_i, \phi_i = 0) \mid h_i\big)$. Each type may believe most judges are of their same type, in which case separating is the unique PBE strategy. Alternatively, both types may believe most judges are of type x. I now show that in this case, pooling or semi-pooling on $1 - x$ is not a PBE strategy.

Suppose players pool on x in a PBE. Then type $1 - x$ must believe at least half of judges believe preferences match if she chooses x. Type x will also believe half of judges believe preference match (by assumption), so would not deviate for any out-of-equilibrium belief. We can then use the intuitive criterion to determine that judges believe only $1 - x$ would deviate from pooling on x. But then if type $1 - x$ believes most judges are of type $1 - x$, pooling on x would not be a PBE strategy. Furthermore, a 'semi-pooling on x' strategy where type $1 - x$ randomizes and type x chooses x would not be a PBE strategy: type $1 - x$ would deviate to always choosing $1 - x$.

We rule out PBE strategies in which both types mix with the following:

Lemma 3. *If separating is not a PBE strategy for player i, it is not a PBE strategy for both types of player i to choose a mixed strategy.*

I now close the argument by describing the PBE dynamics when $\beta \in (1, 2)$. Suppose first that there is an equal number of publicly observed signals of type 0 and of type 1, or $\sum_{k\neq i} \phi_k^o = 0$. Then by Lemma 2, it follows that $f_j(1/2, -1, 0) = g_i(1/2, -1, 0) < f_j(1/2, 1, 0) = g_i(1/2, 1, 0)$ for any history of play. Then separating is a PBE strategy if $g_i(1/2, -1, 0) < 0 < g_i(1/2, 1, 0)$. Otherwise, for some $x \in \{0, 1\}$ only type x judges believe preferences match if player i pools. Further, decision makers believe most judges are of type x. Decision maker i pools if she has strong enough beliefs that players are of type x, and pooling is the unique PBE strategy if expression (5) holds.

Because $f_j(1/2, \phi_j^o, 0)$ and $g_i(1/2, \phi_i^o, 0)$ are not affected by the history of play, then either the first player pools when $\sum_{k\neq i} \phi_k^o = 0$, or no players pool. Therefore, whether all players

pool is determined completely by the environment: the context χ and the precision π. The conditions that lead all players to pool are the conditions that define an informative context.

Now suppose that there are more public signals of type 1 than of type 0, or $\sum_{k \neq i} \phi_k^o = y > 0$ (this is without loss of generality; an analogous argument holds if there are more public signals of type 0). Then by Lemma 2, it follows that in an uninformative context, $f_j(1/2, -1, y) < g_i(1/2, -1, y) < g_i(1/2, 1, y)$, $f_j(1/2, -1, y) < f_j(1/2, 1, y)$, and $0 < f_j(1/2, 1, y) < g_i(1/2, 1, y)$ for any history of play. Consider players' incentives in different cases:

- If $0 < f_j(1/2, -1, y)$, then only type 1 judges believe preferences match if i pools, and the decision maker believe most judges are of type 1. As y grows, the decision maker's expected proportion of judges of type 1, or $\mathbb{E}_{j \neq i}(P(\phi_j = 1 \mid h_i, \phi_i = 1) \mid h_i)$, becomes arbitrarily close to 1. Mechanically, as y grows, the decision maker observes at least proportion $y/|I|$ of the group prefers 1. Moreover, an increase in y also increases the probability that the population is $\theta = 1$, which makes it more likely that withholder judges are of type 1. For any $\beta > 1$ and finite I, there is then a y large enough such that pooling on 1 is a PBE strategy, and separating is not.

- If $f_j(1/2, -1, y) < 0 < g_i(1/2, -1, y)$, all judges except for revealers of type 0 believe preferences match, since $0 < f_j(1/2, 1, 0) \leq f_j(1/2, 0, y)$ by Lemma 1 and because the context is uninformative. But then the social expectation from pooling is higher than the social expectation from revealing 1. Notice that this case can only arise when $y = 1$. When there are at least 2 more observed 1 types than observed 0 types, all judges believe withholders are most likely of type 1: $f_j(1/2, -1, 2) > 0$.

- If $g_i(1/2, -1, y) < 0 < g_i(1/2, 1, y)$, then separating is a PBE strategy. Notice that this case can only arise when $y = 1$. When there are at least 2 more observed 1 types than observed 0 types, all decision makers believe most judges are most likely of type 1: $g_i(1/2, -1, 2) > 0$.

The cases above show that when $y > 1$ and the context is uninformative, the social expectations from pooling are the same than from revealing 1: only judges of type 1 believe preferences match. These same type 1 judges are the only ones who believe preferences match if the decision maker chooses 1 when following a strategy of semi-pooling on 1. Further, the social expectation from choosing 0 are the same for these same strategies of pooling on 1, semi-pooling on 1 or separating – only judges of type 0 believe preferences match (this uses the intuitive criterion as argued above). Therefore, the optimal choice for each type is the same for any of these strategies. From our earlier results we can further conclude that

pooling on 0, semi-pooling on 0, or a strategy where both types are mixing are not PBE strategies when $y > 1$ and the context is uninformative – the decision maker believed most judges were of type 1. There is then a unique PBE strategy when $y > 1$ and the context is uninformative. Through a similar reasoning, I can show that there is a unique PBE strategy when $y < -1$ and the context is uninformative.

When $y = 1$, more judges may believe preferences match if the decision maker pools than if she reveals 1. If type 0 deviates from a separating strategy, pooling on 0 is a PBE strategy. If she deviates from pooling on 0, separating is an equilibrium strategy. However, I cannot rule out multiple equilibria. An analogous statement can be obtained for $y = -1$.

This establishes Claim 1, and out result. □

B Proof of Lemma 1

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

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

$P(\phi_i = 1 \mid \theta, h_i, a_i, \phi_j) = P(\phi_i = 1 \mid \theta, h_i, a_i)$, since once we condition on the state of the world θ, the only reason to condition on history (h_i, a_i) is to determine i's type-dependent strategy, which is publicly known in equilibrium. We can apply Bayes' rule to obtain

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

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

$$P(\phi_i = 1 \mid h_i, a_i, \phi_j) = \frac{\sum_{\theta \in \{0,1\}} P(\theta) \prod_{k \in \{1,...,i-1\}} [P(a_k \mid h_k, \theta)] P(\phi_j \mid \theta) P(\phi_i = 1 \mid \theta, h_i, a_i)}{\sum_{\theta \in \{0,1\}} P(\theta) \prod_{k \in \{1,...,i-1\}} [P(a_k \mid h_k, \theta)] P(\phi_j \mid \theta)}$$

By setting $P(\phi_i = 1 \mid h_i, a_i, \phi_j) > z$, noting that $P(\phi_i = 1 \mid \theta, h_i, a_i) = P(\phi_i = 1 \mid \theta)$

since i is a withholder, and rearranging the terms, we get the first part of the Lemma.

We can follow an analogous set of steps to obtain that $\mathbb{E}_{j\neq i}\big(P(\phi_j = 1 \mid h_i, \phi_i)\big)$ is equal to

$$\frac{\sum_{\theta\in\{0,1\}} P(\theta)\prod_{k\in\{1,\ldots,i-1\}}[P(a_k \mid h_k,\theta)]P(\phi_i \mid \theta)\mathbb{E}_{j\neq i}\big(P(\phi_j = 1 \mid h_i,\phi_i,\theta)\big)}{\sum_{\theta\in\{0,1\}} P(\theta)\prod_{k\in\{1,\ldots,i-1\}}[P(a_k \mid h_k,\theta)]P(\phi_i \mid \theta)}$$

By setting $\mathbb{E}_{j\neq i}\big(P(\phi_j = 1 \mid h_i, \phi_i)\big) > z$, noting that $\mathbb{E}_{j\neq i}\big(P(\phi_j = 1 \mid h_i, \phi_i, \theta)\big)$ is equal to $\mathbb{E}_{j\neq i}\big(P(\phi_j = 1 \mid h_i, \theta)\big)$ since players have either separated or are withholders, and rearranging the terms, we get the second part of the Lemma. □

C Proof of Lemma 2

Proof. An increase in χ increases f_j and g_i straightforwardly. An increase in ϕ_j or $\sum_{l\neq j}\phi_l^o(h_i)$ increases the first multiplicand of f_j and of g_i. The sum of observed types $\sum_{l\neq j}\phi_l^o(h_i)$ also affects g_i through $\mathbb{E}_{j\neq i}\big(P(\phi_j = 1 \mid h_i,\theta) \mid h_i\big)$, which we can rewrite as follows:

$$P(\phi_{j\neq i}^o = 0 \mid h_i)P(\phi_{j\neq i} = 1 \mid \phi_j^o = 0,\theta) + P(\phi_{j\neq i}^o = 1 \mid h_i)$$

The term $P(\phi_{j\neq i}^o = x \mid h_i)$ is the probability that a player other than i is of observed type x, and does not depend on θ because equilibrium strategies cannot condition on an unobserved parameter. The term $P(\phi_{j\neq i} = 1 \mid \phi_j^o = 0,\theta)$ is the probability that a withholder is of type one when the population is known to be θ, which is equal to the precision π. Once the population is known, history does not provide information about a withholder's type.

The expectation $\mathbb{E}_{j\neq i}\big(P(\phi_j = 1 \mid h_i,\theta)\big)$ is a weighted sum of revealers of type 1, withholders, and revealers of type 0 – revealers of type 1 receive a weight of one, withholders receive a weight between zero and one, and revealers of type 0 receive a weight of zero. Therefore, the expectation would increase if and only if $\sum_{l\neq k}\phi_l^o(h_i)$ increases – say by replacing a revealer of type 0 with a withholder, or a withholder with a revealer of type 1. This also increases $\Delta^o(h_i,\phi_l)$. Since $g_i(z)$ increases in $\mathbb{E}_{j\neq i}\big(P(\phi_j = 1 \mid h_i,\theta)\big)$ for $\theta\in\{0,1\}$, the result follows.

The only difference between f_j and g_i are the bracketed terms. Notice that, since $\pi - 1/2 = 1/2 - (1-\pi)$, the bracketed terms in $f_j(1/2)$ can be canceled out. I will show that the bracketed terms in $g_i(1/2)$ can also be cancelled out when $\sum_{k\neq j}\phi_k^o(h_i) = 0$. The rest of the result follows from the fact that a change in $\sum_{k\neq j}\phi_k^o(h_i)$ impacts the bracketed terms in $g_i(z)$, but not in $f_j(z)$.

Suppose that there are an equal number of publicly observed signals of type 1 and of type 0, or $\sum_{k\neq j}\phi_k^o(h_i) = 0$. Players i and j at history h_i observe the same number of signals

of type 1 and of type 0: $P(\phi^o_{j\neq i} = 1 \mid h_i)$ is equal to $P(\phi^o_{j\neq i} = -1 \mid h_i)$. Further, notice that by the symmetry of the distribution of preferences conditional on the population, $P(\phi_j = 1 \mid \phi^o_j = 0, \theta = 1)$ is equal to $P(\phi_j = 0 \mid \phi^o_j = 0, \theta = 0)$. Using these facts, plus rewriting some probabilities in terms of their complements, we can express $\mathbb{E}_{j\neq i}\big(P(\phi_j = 1 \mid h_i, \theta = 1)\big) - 1/2$ as follows:

$$(1 - P(\phi^o_{j\neq i} = 1 \mid h_i) - P(\phi^o_{j\neq i} = -1 \mid h_i))(1 - P(\phi_j = 1 \mid \phi^o_j = 0, \theta = 0))$$
$$+\frac{P(\phi^o_{j\neq i} = 1 \mid h_i) + P(\phi^o_{j\neq i} = -1 \mid h_i)}{2} - \frac{1}{2}$$

Canceling out terms, this expression is equal to:

$$1/2 - (1 - P(\phi^o_{j\neq i} = 1 \mid h_i) - P(\phi^o_{j\neq i} = -1 \mid h_i))P(\phi_j = 1 \mid \phi^o_j = 0, \theta = 0)$$
$$-\frac{P(\phi^o_{j\neq i} = 1 \mid h_i) + P(\phi^o_{j\neq i} = -1 \mid h_i)}{2}$$

By using the complement of probabilities and the fact that $\sum_{k\neq j} \phi^o_k(h_i) = 0$, we get the following expression:

$$1/2 - P(\phi^o_{j\neq i} = 0 \mid h_i)P(\phi_j = 1 \mid \phi^o_j = 0, \theta = 0) - P(\phi^o_{j\neq i} = 1 \mid h_i)$$

which is equal to $1/2 - \mathbb{E}_{j\neq i}\big(P(\phi_j = 1 \mid h_i, \theta = 0)\big)$.

The last claim in the Lemma follows directly from applying Lemma 1 to the definition of informative and uninformative contexts.□

D Proof of Lemma 3

If $\beta \in (0, 1)$, player i chooses ϕ_i. If $\beta > 1$, for ϕ_i to be indifferent between actions, it must be that

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

But this only holds for both types if each believes there are more judges of their own type. But then separating is a PBE strategy.□

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

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

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

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

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

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

E Proof of Result 2

If the context is uninformative, player i will pool on x only if the lead for action x is at least $n > 0$ (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 the minority preference in the population goes to zero. The probability that players pool on the minority preference of the population, however, is positive: it is guaranteed if the first n players's preference is the minorty preference in the population.

If the context is informative, then the probability of pluralistic ignorance is given by the probability that a strict majority of group members prefer 0, equal to $[(1-\pi)\chi + \pi(1-\chi)]^2$ if $I = 2$. As the group size increases, the probability that the majority preference of the group is equal to the majority preference of the population comes arbitrarily close to one. The probability that the group is drawn from population $\theta = 1$ comes arbitrarily close to 1 as χ increases. So the probability of pluralistic ignorance goes to zero for high enough $|I|$ and θ. $\square$

F 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 the intuitive criterion, 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 also not deviate from a separating strategy, who again increase future player of type 0's incentives to pool by revealing her type (by Result 1). But then, since we know by Result 1 that players choose pure strategies when the context is uninformative or informative, we can conclude that with an initial run of at most $n > 0$ players of type 1, all players pool on 1.

If $I \geq 2n + 1$, where n is as defined in Result 1, 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.□

G Proof of Result 4

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

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

Since $P(a_{\mathcal{P}} \mid \theta) = 1 - P(1 - a_{\mathcal{P}} \mid \theta)$, then if $P(a_{\mathcal{P}} \mid \theta) > P(a_{\mathcal{P}} = 1 \mid 1 - \theta)$, it follows that $P(1 - a_{\mathcal{P}} \mid 1 - \theta) > P(1 - a_{\mathcal{P}} \mid \theta)$. Therefore, if action $a_{\mathcal{P}}$ makes players believe it is more likely the population is θ, action $1 - a_{\mathcal{P}}$ makes players believe it is more likely the population is $1 - \theta$. If the context is informative, and $\mathcal{P}$'s actions are informative about the state of the world, there is some action $1 - a_{\mathcal{P}}$ that leads to an informative context, and therefore does 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.□

H Proof of Result 5

This follows from an analogous reasoning to the proof of Result 1. 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.□

I Proof of Result 6

As δ becomes arbitrarily large, individuals give negligible weight to the original utility function if $l = 1$. Since $\beta > 1$, for δ large enough group members' main incentives are to choose whichever action maximizes $\mathbb{E}_{\hat{\psi}_{-i}}(\mathbb{E}_{j \neq i}(\mathcal{K}_{j,i} \mid \hat{\psi}_{-i}))$. If $\hat{\pi}$ is large, then all individuals will think group members are most likely of type $\psi_i = 1$. Player 1 of type $\psi_1 = 0$ would deviate from a separating equilibrium in order to signal that she is of type $\psi_i = 1$. If the first player pooled, group members would believe she is most likely of type $\psi_i = 1$, and player 1 thinks her judges are most likely of type $\psi_i = 1$. Therefore, player 1 of type $\psi_i = 1$ does not deviate from pooling on $a_i = 0$, so by the intuitive criterion a deviation must come from type $\psi_i = 0$. Since type $\psi_i = 1$ would not deviate from pooling on 0 for β and $\hat{\pi}$ large enough, this is an equilibrium.

J 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.□

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

K.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.

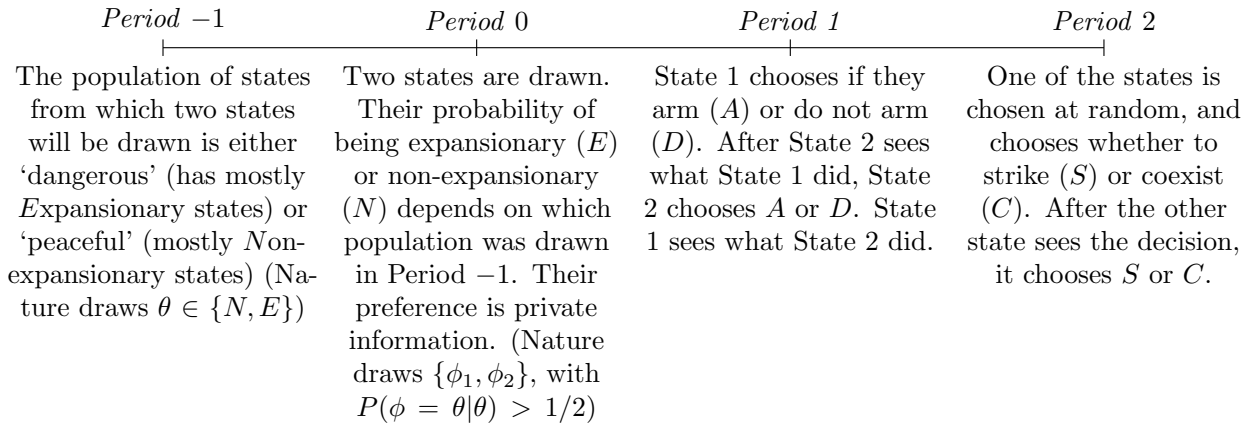

Figure 3: Timeline

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 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,

[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 $1 - p(\cdot, N, N, \cdot, \cdot) - c < u_2(x; N)$.

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 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.

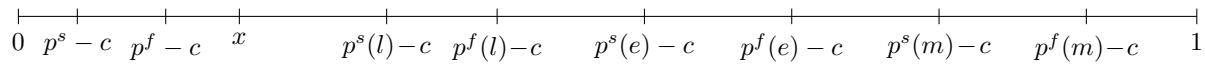

Figure 4: Expected Payoffs From Going To War For State 1

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.

K.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:

$$P(\phi_j = E \mid \phi_i = N, h_i) > \frac{u_1(x; N) - p + c}{u_1(x; N) - p^s(r) + c > 0} \tag{6}$$

where h_i is the history of play and $r \in \{l, e, m\}$ is determined and known in equilibrium. We will summarize condition (6) with the 'judgment function' $J(h_i)$, equal to one if (6) 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

[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.

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.

Without further justification, I state the result. Since the logic of the equilibrium is very similar to that of Result 1, I will state it as a claim.

Claim 2. *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.