Disliking to disagree*

Florian Hoffmann† Kiryl Khalmetski‡ Mark T. Le Quement§

January 24, 2019

Abstract

Abundant experimental and field evidence suggests that people tend to dislike open disagreement. We propose a formalization of perceived disagreement and study the implications of perceived disagreement aversion in disclosure games involving agents with different priors. Across a variety of settings, the ideal conditions for disclosure involve identical prior variances and differing prior means. When equilibrium disclosure is partial, it is biased towards evidence that is congruent with the most confident agent's prior bias. Perceived disagreement aversion leads to assortative matching in prior beliefs that provides a theoretical basis for echo chambers. Equilibria may feature higher average perceived or actual disagreement than a hypothetical full disclosure scenario. Perceived disagreement aversion arises endogenously within simple games of delegation and competitive authority assignment.

Keywords: strategic disclosure, psychological games, disagreement aversion

JEL classification: D81, D83, D91

*We thank Pierpaolo Battigalli, Oana Borcan, Gary Charness, Martin Dufwenberg, Ángel Hernando-Veciana, Martin Kocher, Matias Nunez, Axel Ockenfels, Amrish Patel, Karl Schlag, Rajiv Sethi, Joel Sobel, Peter Norman Sørensen, Robert Sugden, Joel van der Weele and Jörgen Weibull for helpful comments and suggestions. We thank participants at the following workshops and seminars: Southampton research seminar 2019, ESEM 2018, Toulouse IAS 2018, Paris-Cergy IAS 2018, Warwick Dr@w seminar 2018, 2nd workshop on Psychological Game Theory 2017, UEA Theory workshop 2017, Political Economy workshop in Konstanz 2017 and seminar of the DFG Research Unit "Design and Behavior" 2017. Khalmetski gratefully acknowledges financial support of the German Research Foundation (DFG) through the Research Unit "Design and Behavior" (FOR 1371).

†Erasmus University Rotterdam. E-mail: hoffmann@ese.eur.nl.

‡University of Cologne. E-mail: kiryl.khalmetski@uni-koeln.de.

§University of East Anglia. E-mail: m.le-quement@uea.ac.uk.

1 Introduction

Decentralized information exchange within social networks is an important channel shaping public opinion, which is ever more important in the digital era (Internet, social media).[1] While avoiding some of the distortions that are particularly relevant for centralized information flows, this source of information itself exhibits many forms of bias. In particular, people do not talk equally easily about all topics, are not equally willing to disclose all facts or opinions, and are not equally likely to talk to everyone. A 2016 poll by the online employment website *CareerBuilder* finds that 42 percent of respondents avoid talking politics at the office while 44 percent may talk about it but interrupt the conversation if it becomes heated.[2,3] Social-psychologists have developed a wide repertoire of concepts to describe informational biases arising in social networks, e.g. *Taboos, Overton windows, opinion corridors, political correctness, conversational minefields, echo chambers, confirmation bias, pluralistic ignorance, information avoidance.*

An important role in generating these biases can be attributed to the tendency to avoid conflict in opinions (i.e. to the desire to be perceived as having similar beliefs as the counterparty), or *perceived disagreement aversion.* A large body of experimental and field evidence documents that individuals tend to state opinions that conform to what they believe others think. In the seminal experiments conducted by Asch (1955), subjects wrongly evaluated the length of a line in public after being exposed to other participants' (artificially induced) wrong assessment. Deutsch and Gerard (1955) showed that this effect is weaker if subjects report their judgment privately, so

[1] See Sunstein (2007), p. 52.: "*In contrast to television, many of the emerging technologies are extraordinarily social, increasing people's capacity to form bonds with individuals and groups that would otherwise have been entirely inaccessible. Email, instant messaging, texting and Internet discussion groups provide increasingly remarkable opportunities, not for isolation, but for the creation of new groups and connections.*"

[2] *Political Talk Heats Up the Workplace, According to New CareerBuilder Survey*, CareerBuilder.com, Press Releases, July 2016.

[3] See also for example the following recommendation from the gentleman's manual "*Hills Manual of Social and Business Forms*" from (1879): "*Do not discuss politics or religion in general company. (...) To discuss those topics is to arouse feeling without any good result.*"

that others' *perceived* disagreement is unaffected. Mutz (2006) reviews a number of studies showing that Americans avoid discussing politics with non like-minded people for fear of creating tensions.[4] Bursztyn et al. (2017) found that subjects were more likely to publicly reveal immigration-critical views two weeks after Donald Trump's victory than two weeks before it (i.e. before it became apparent that such views might be shared by a large fraction of the population). Prentice and Miller (1993) established that a large fraction of students refrained from expressing dissent with campus alcohol practices for fear of stigma, vastly underestimating the share of people sharing their opinion.[5]

While there is ample evidence of the relevance of perceived disagreement aversion, to the best of our knowledge it has not been formally modeled. This paper is a first step towards filling this gap. We suggest a formalization of perceived disagreement aversion and analyze its consequences for the incentives to generate and share hard (verifiable) information with other agents which may have different prior beliefs.[6] Given mutually known priors, our baseline specification measures i's perceived disagreement between herself and another agent j simply as the (absolute) distance

[4] See Mutz (2006), p. 107: "*There is already ample evidence in support of the idea that people avoid politics as a means of maintaining interpersonal harmony. For example, in the mid 1950s, Rosenberg noted in his in-depth interviews that the threat to interpersonal harmony was a significant deterrent to political activity. More recent case studies have provided further support for this thesis. Still others have described in great detail the lengths to which people will go in order to maintain an uncontroversial atmosphere. Likewise, in focus group discussions of political topics, people report being aware of, and wary of, the risks of political discussion for interpersonal relationships. As one focus group participant put it, "It s not worth it to try and have an open discussion if it gets them [other citizens] upset."*

[5] Disagreement aversion has many potential causes (see Golman et al., 2016, for a general review of what the authors term a preference for *belief consonance*). Individuals might experience an intrinsic psychological discomfort from being explicitly confronted with disagreement in views (Festinger, 1957; Domínguez et al., 2016). The aversion may instead be driven by the anticipation of adverse consequences stemming from disagreement. For instance, political practice in north-western Europe (e.g. Netherlands and the so-called Polder model, Scandinavia) puts a strong emphasis on reaching consensus, in particular in negotiations between different labor market organizations.

[6] Heterogeneous priors are an integral part of many social situations. Instances range from views on general questions (climate change, immigration, free trade, religion, minority rights) to how to manage a firm or optimize an investment portfolio. A key underlying source is that people have different personal histories (experiences, socialization, education). See Morris (1995) for an early general discussion, and Acemoglu et al. (2016), Banerjee and Somanathan (2001), Gentzkow and Shapiro (2006), and Dixit and Weibull (2007) for modeling applications.

between i's expected value of the state (i's first order belief) and i's expectation of j's expected value of the state (i's second order belief), the utility of a perceived disagreement averse agent being strictly decreasing in this distance.

A main source of tension for information sharing originates in the mechanics of Bayesian updating: Though agents update their prior expectation in the same direction whatever the observed signal, the magnitude of belief adjustment depends on the prior belief distribution. It follows that disagreement may well increase following the disclosure of particular signal realizations.[7] As a consequence, a perceived disagreement averse agent has incentives to selectively reveal or hide his private information. The same robust intuition also implies that the benefits of generating costly information, as regards the reduction in (perceived) disagreement, also depends on agents' prior distributions. One of our contributions is to characterize how different specifications of prior heterogeneity may facilitate or hinder information generation and sharing within a given group of disagreement averse agents, as well as affect the choice of conversation partners.

Our baseline model is a simple game of strategic disclosure by a potentially informed sender (S) who is averse to disagreement as perceived by an uninformed receiver (R). The sender privately observes, with some commonly known probability, an informative signal drawn from a known distribution, and can decide whether to disclose it to the receiver or not.[8]

Information transmission in equilibrium can be characterized based on the differences in means and variances of the heterogeneous prior distributions. This is interesting because these quantities have a natural interpretation. The prior mean represents an agent's prior stance. The prior variance represents his confidence in his prior stance and his willingness to revise his stance as new information becomes available. While we consider various specifications of the state space, the prior dis-

[7] This is most easily seen by comparing the belief adjustment with a degenerate prior (which is zero) to the one with, e.g., a uniform prior (which is strictly positive).

[8] As is standard (see, e.g., Jung and Kwon, 1988), scope for selective disclosure emerges when the probability of being informed is interior, preventing full unraveling.

tributions and the signal structure, the basic intuition is most apparent within the simple setup in which the state of the world is either 0 or 1 and S's signal is binary and of known precision (call this the binary-binary setting). In this setup, we denote the commonly known prior beliefs that the state is 1 by β_i, $i \in \{S, R\}$.

With heterogeneous priors, this game has (almost always) a unique pure-strategy equilibrium that always features some information transmission, as at least one signal realization is disclosed. Whether full disclosure is feasible however crucially depends on the prior profile. For any signal precision, full disclosure is feasible if β_S is close enough to $1 - \beta_R$ while for low enough precision, full disclosure is not feasible if β_S is close enough (but not identical) to β_R. The profile of priors that makes it easiest to achieve full disclosure thus features similar prior variances and a potentially large difference in prior means. In such a profile, agents' willingness to revise their stance, and hence the magnitude of their belief adjustments, in the face of confirming and contradictory evidence is similar.[9] In contrast, given a small difference in prior variances, a potentially significant difference in prior means ($\beta_S \approx 1 - \beta_R$) is better than almost none ($\beta_S \approx \beta_R$).[10] The reason is that sufficiently different means ensure that a player with a higher (lower) mean will be relatively less affected by a higher (lower) signal, which in turn implies convergence in posterior beliefs whatever signal is disclosed.

We find that if disclosure is partial, the information selectively revealed by S is biased towards evidence that is congruent with the more confident player's prior belief. As an illustration, in the binary-binary setting, consider the case in which the most confident player assigns higher probability to state 1. Then, if equilibrium involves partial disclosure, only 1-signals are shown.[11]

[9] This result extends beyond the binary world: In the canonical normal priors - normal signals setting, full disclosure is possible if and only if prior variances are identical, and full disclosure is the only equilibrium outcome if and only if prior means furthermore differ.

[10] In the normal-normal setting, when prior variances differ, the set of disclosed signals has zero measure under identical prior means and instead positive measure when prior means differ.

[11] Within the normal-normal setting, consider a situation in which both prior means and variances differ across players. Then, only signals within a bounded interval are disclosed, and this interval is

We demonstrate in the binary-binary setting that perceived disagreement aversion generates echo chamber-like dynamics in simple matching scenarios. If receivers are randomly matched with senders and priors are publicly observed, a more confident receiver is less likely to encounter contradicting information, this probability tending to zero as his prior variance tends to 0. Allowing for repeated random pairwise encounters, this leads to inertia in learning dynamics. We then show that confirmatory information bias is further strengthened if disagreement averse senders can choose whom to be matched with, while society exhibits a sufficiently high degree of polarization in priors. Senders, rationally anticipating the nature of equilibrium disclosure, only interact with receivers whose prior mean is similar to their own (assortative matching). Our equilibrium characterization then implies that in the exclusively like-minded matches that are formed, only information congruent with the shared bias will be disclosed.

Our theory of perceived disagreement aversion hence offers a putative explanation of the following two stylized facts. First, many citizens are exposed disproportionately to information that confirms their worldview (echo chambers). Second, there is very significant positive assortative matching in communicative behavior on the basis of worldviews (worldview homophily), partially as a result of the Internet. These stylized facts are often presented and discussed together.[12] Our tentative explanation of these facts rests on rationality, heterogeneous priors and aversion to (perceived)

closer to the prior mean of the more confident player in terms of Hausdorff distance.

[12] See Mutz (2006), p. 9: *"Social network studies have long suggested that likes talks to likes; in other words, people tend to selectively expose themselves to people who do not challenge their view of the world. Network survey after network survey has shown that people talk more to those who are like them than to those who are not, and political agreement is no exception to this general pattern."* See also Sunstein (2007), p. 145: *"because of self-sorting, people are often reading like-minded points of view, in a way that can breed greater confidence, more uniformity within groups, and more extremism. Note in this regard that shared identities are often salient on the blogosphere, in a way that makes polarization both more likely and more likely to be large."* See also Sunstein (2007), p. 63: "*The phenomenon of group polarization has conspicuous importance for the communications market, where groups with distinctive identities increasingly engage in within-group discussion. (...) New technologies, emphatically including the Internet, make it easier for people to surround themselves (virtually of course) with the opinions of like-minded but otherwise isolated others, and to insulate themselves from competing views. For this reason alone, they are breeding ground for polarization, and potentially dangerous for both democracy and social peace.*"

disagreement.[13]

We extend our analysis in various directions.[14] First, we adapt our baseline measure of perceived disagreement, which presumes commonly known priors, to allow for uncertainty about priors. Within our baseline bilateral disclosure game, (expected) prior heterogeneity in means continues to be conducive to information transmission, echoing the results obtained under known priors. Furthermore, we show that uncertainty about priors may actually be beneficial for information sharing in equilibrium.

In our baseline model of equilibrium disclosure the sender aims at being "politically correct," in the sense of selectively disclosing only signals that reduce *perceived* disagreement. There is an ongoing debate about the value of such self-censorship. In particular, critics of this view of political correctness often point to the benefits of encouraging people to freely speak their minds. Linking to this debate, we evaluate the value of commitment to a full disclosure strategy - or, respectively, the hidden cost of political correctness - and find that equilibrium disclosure by a perceived disagreement averse sender induces higher perceived disagreement in expectation than commitment to a full disclosure strategy if and only if the sender is more confident. Interestingly, equilibrium disclosure might also dominate full disclosure with respect to expected *actual* disagreement. To see this, we take the perspective of a third party who cares about minimizing the expected ex post *actual* disagreement between S and R. Then, while it is immediate that full disclosure is the optimal commitment strategy whenever the third party shares the prior of either S or R, this need no longer be the case when the third party has a different prior, highlighting once again the role

[13] An alternative theory explaining these stylized facts is that people talk in order to make the right decisions (say, match the state) and induce others to do the same. This theory predicts that more similar worldviews lead to better information transmission but thereby fails to explain why homogeneity in groups seems to correlate with (confirmation) biased learning. One might assume that individuals underestimate the extent to which peer group members' information correlates with their own, and as a consequence overweight others' information. This theory of so-called correlation neglect has been explored in Levy and Razin (2015) and Glaeser and Sunstein (2009). The theory assumes an element of bounded rationality, which is not the case of ours.

[14] For expository reasons we chose to present these extensions within the binary state - binary signal model.

of prior heterogeneity.

Our main analysis exogenously assumes perceived disagreement aversion and derives implications for information transmission. The main results on equilibrium information transmission hold independently of whether perceived disagreement causes purely psychological costs or indirectly affects material payoffs via its effect on decisions in equilibrium of a larger game. As an illustration, we consider simple games of delegated decision making and competition for authority, in which a disclosure stage is followed by one or several stages of decision making that generate material payoffs for both the sender and the receiver. We show that in equilibrium, in his quest to influence subsequent decision making, the privately informed party acts *as if* disagreement averse at the disclosure stage, giving rise to the same dependence of equilibrium information transmission on prior heterogeneity as in our main (reduced form) setup.

While our main focus is on the implications of perceived disagreement aversion for the incentives to share information, it also generates novel insights regarding the generation of information. To see this we consider a game of costly collective acquisition of public signals by parties who are (perceived) disagreement averse. Though the game is strategically different from our disclosure game, it addresses a similar problem of learning in (two person) groups with heterogeneous prior beliefs. We find that moderate heterogeneity in prior means optimally incentivizes information acquisition, which echoes previous findings for strategic disclosure.

Literature review In its foundations, our paper relates to a literature studying how (public) information relates to disagreement in beliefs. The literature so far focused on actual (instead of perceived) disagreement, trying to explain phenomena such as polarization, which refers to situations where individuals update in opposite directions on the basis of the same information. This may result from different prior beliefs (Dixit and Weibull, 2007; Acemoglu et al., 2007; Sethi and Yildiz, 2012), different privately observed prior signals (Andreoni and Mylovanov, 2012) as well as

ambiguity (Baliga et al., 2013).[15] Under certain conditions, disagreement in beliefs may persist in the long run, i.e. asymptotically (Acemoglu et al., 2016; Andreoni and Mylovanov, 2012).[16]

An extensive body of research dating back to Grossman (1981), Milgrom (1981), and Milgrom and Roberts (1986) studies strategic disclosure of verifiable signals by a privately informed sender.[17] These models typically involve a difference in players' preferences over the receiver's action conditional on the state. Newer papers study the case of different prior beliefs, often featuring identical material preferences given the state, such as Banerjee and Somanathan (2001) and Kartik et al. (2015). While these papers, thus, share important elements of our analysis, perceived disagreement does not play a role in shaping equilibrium incentives for disclosure. Relatedly, Che and Kartik (2009) examine the effect of prior belief misalignment on the sender's incentives to privately acquire costly information. Prior misalignment hurts disclosure but increases information acquisition, so that the receiver may ultimately benefit from more misalignment. While potential benefits of prior misalignment also feature prominently in our analysis, this results from a different mechanism.[18] In particular, when information transmission is driven by perceived disagreement aversion, (some) prior misalignment *encourages* disclosure independently of whether information is given exogenously or acquired at some cost.

A strand of the literature on strategic information transmission features an endogenous preference for belief conformity arising from reputational concerns. Morris (2001) (see also Sobel, 1985; Benabou and Laroque, 1992; Ely and Välimäki,

[15] Several papers in network economics consider the effect of individual conformity to the beliefs or opinions of others on belief polarization (Dandekar et al., 2013; Buechel et al., 2015; Golub and Jackson, 2012).

[16] Sethi and Yildiz (2016) focus on the fact that observing others' opinion over time, an observer learns both about their subjective prior and about their private information concerning some objective state, thereby triggering non-trivial dynamics in belief updating.

[17] See in particular also Jung and Kwon (1988) for the baseline model of random disclosure as well as Shin (1994a,b). See Sobel (2013) for a general review of the literature on strategic information transmission.

[18] This is apparent by noting that their result rests on strictly positive costs of information acquisition.

2003) studies a sender-receiver game with an endogenous reputational concern of the sender for being perceived as unbiased, which leads to distorted communication.[19] In Gentzkow and Shapiro (2006), the sender wishes to signal a high quality of her information to the receiver who may remain uninformed about the actual state. This leads her to bias her message towards the receiver's prior belief.[20] Similarly, in our setup if the sender is less confident, she omits signals contradicting the receiver's prior. The motivation is however very different: In our model the sender wants to mitigate perceived ex post disagreement with the quality of her information being known. This same objective will as a matter of fact lead the sender to omit signals that confirm the receiver's prior if the latter is less confident.

Our study also contributes to the growing body of literature on psychological game theory, which posits preferences that directly incorporate beliefs (of arbitrary order) about others' strategies or beliefs (Geanakoplos et al., 1989; Battigalli and Dufwenberg, 2009). Here, our analysis is related to Ely et al. (2015) who consider the behavior of a principal wishing the beliefs of an agent to follow a specific time path exhibiting suspense or surprises. While this paper as well as our baseline specification focus on pure belief-based preferences, several more applied models allow preferences to depend on the interplay between beliefs and material payoffs, see, for instance, the models of reciprocity (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004) or guilt aversion (Battigalli and Dufwenberg, 2007), as well as the settings in our section 3.3.

Our paper also relates to a rich theoretical and empirical literature in social psy-

[19] Loury (1994) offers a stimulating discussion of self-censorship and political correctness in public discourse stemming from such concerns.

[20] The models in Ottaviani and Sørensen (2006a) and Ottaviani and Sørensen (2006b) embed a similar setting with ex post verifiable reports, resulting in S's reporting conforming to his own prior. Visser and Swank (2007) study deliberative committees whose members want to signal high expertise. This gives them an incentive to pretend to have similar signals (i.e. to agree) and to decide against the prior. Within a similar setup Levy (2007) focuses on the impact of transparency rules on decision making. In a principal-agent setting, Prendergast (1993) examines the agent's incentive to match the (noisy) information of the principal in his report. Bursztyn et al. (2017) consider a setting where a sender has to communicate his type to a receiver and has an incentive to appear of the same type as the receiver. Bénabou (2012) shows that agents with anticipatory utility may converge to each other's wrong beliefs due to the dependence of one's payoffs on the actions of the others.

chology on biases in network formation, communication and norm adoption, dating back to the 1950s, 1960s and 1970s (see Newcomb, 1961; Homans, 1961; Asch, 1955; Lazarsfeld and Merton, 1954; Festinger, 1950; Rosenberg, 1954; Huston and Levinger, 1978; Goffman, 1959). Finally, our paper also links to a research agenda in political economy and political theory on deliberative and so-called epistemic democracy (see Estlund, 2009; Landemore and Elster, 2012; Sunstein, 2007; Mutz, 2006; Huckfeldt et al., 2004; Feddersen and Pesendorfer, 1998; Coughlan, 2000; Austen-Smith and Feddersen, 2006) originating in Condorcet's seminal work on majority voting. The agenda evaluates democratic institutions and practices in terms of their ability to aggregate information (their so-called truth-tracking properties), which ultimately rests on citizens' incentive or ability to use their private information as well as share it with each other.

The remainder of the paper is organized as follows. Section 2 presents the benchmark model and the main theoretical results. Section 3 considers extensions of the model and possible microfoundations. Section 4 concludes. All proofs, unless explicitly stated otherwise, are relegated to the online Technical Appendix.

2 Main analysis

2.1 The disclosure game

There are two agents - the sender (S, he) and the receiver (R, she) and a state of Nature $\omega \in \{0,1\}$. Player $i \in S,R$ assigns prior probability $\beta_i \in (0,1)$ to $\omega = 1$. Priors are common knowledge.[21] S holds with probability $\varphi \in (0,1)$ a privately observed informative signal which has a value of either 0 or 1. Thus, S holds information $\sigma \in \{0,1,\varnothing\}$, where $\varnothing$ stands for no signal. If S obtains a signal, it is identical to the state with probability $p \in \left(\frac{1}{2},1\right]$, i.e. $P(\sigma = \omega) = p$ for $\sigma \neq \varnothing$. If S obtains a signal, he can disclose it to R or not. Denote S's disclosed information by d, where

[21] We consider the case of privately known priors in section 2.5.

$d \in \{0, 1, \varnothing\}$ and where $\varnothing$ stands for no disclosure. R simply observes S's signal if disclosed and subsequently updates beliefs. Let $\widetilde{\beta}_S(\sigma)$ and $\widetilde{\beta}_R(d)$ denote the posterior probability assigned by S and R, respectively, to $\omega = 1$ given obtained information (σ or d) and respective priors β_S and β_R.

After the disclosure stage, R evaluates how much S's expected posterior belief is different from her own. In particular, R's perceived disagreement is

$$\begin{aligned} \Delta(d, \beta_S, \beta_R) &= |E_R[E_S[\omega|\sigma, \beta_S]\,|d] - E_R[\omega|d, \beta_R]| \\ &= \left|E_R[\widetilde{\beta}_S(\sigma)\,|d] - \widetilde{\beta}_R(d)\right|. \end{aligned} \tag{1}$$

The expression $\Delta(d, \beta_S, \beta_R)$ captures the extent to which R thinks that S beliefs are ex ante biased in a specific direction relative to her own, conditional on disclosure d.[22]

S is averse to perceived disagreement on the part of R, i.e. wants to minimize R's ex post perception of disagreement. Hence, S's utility function for given priors β_S and β_R is given as

$$U_S(\beta_S, \beta_R, d) = -\Delta(d, \beta_S, \beta_R). \tag{2}$$

In other words, S's utility is maximized if R *thinks* that S holds the same posterior belief as she. Note also that S's *actual* posterior belief does not enter S's utility function. R's preferences are left unspecified, this player being entirely passive.

Our equilibrium concept throughout is Perfect Bayesian equilibrium: Players' strategies are sequentially rational given their beliefs and others' equilibrium strategies. Second, beliefs are derived via Bayes' rule whenever possible.

A disclosure strategy of S specifies a probability of disclosing at each information set of S, and a disclosure strategy is informative if S discloses with positive ex ante probability. The three informative and pure disclosure strategies are respectively full

[22] Note that the values of $E_R[\widetilde{\beta}_S(\sigma)\,|\varnothing]$ and $\widetilde{\beta}_R(d)$ depend on the assumed disclosure strategy of S, whereas $\Delta(1, \beta_S, \beta_R)$ and $\Delta(0, \beta_S, \beta_R)$ do not. To avoid any ambiguities, we often explicitly write $\Delta^X(d, \beta_S, \beta_R)$, where X is the putative equilibrium disclosure strategy of S. In the Appendix, we similarly introduce $E_R^X[\widetilde{\beta}_S(\sigma)\,|\varnothing]$ and $\widetilde{\beta}_R^X(\varnothing)$.

disclosure (called FD), disclosure of only 1-signals or only 0-signals (called D1 and D0, respectively). We denote by ND the strategy of never disclosing. An equilibrium featuring disclosure strategy $X \in \{FD, D1, D0, ND\}$ is called an X-equilibrium. An equilibrium featuring an informative disclosure strategy is called informative. If $\beta_i > (<)\frac{1}{2}$, we say that i's prior is *biased towards* state 1 (0). If $\beta_i > \frac{1}{2}$, a 1-signal is *congruent with* i's prior bias and a 0-signal *contradicts* it (vice versa if $\beta_i < \frac{1}{2}$). If β_i is strictly closer to the boundary than β_j, then i is said to be more *confident* than j (or i holds a more *confident* prior than j).

2.2 Equilibrium characterization

As our next proposition shows, S's optimal disclosure strategy depends on the relation between players' prior beliefs, i.e. on the position of β_S relative to the following thresholds:

$$\begin{aligned} \beta_S^*(\beta_R, p) &= \frac{(1-p)(1-\beta_R)}{1-p+\beta_R(2p-1)}, \\ \beta_S^{**}(\beta_R, p) &= \frac{p(1-\beta_R)}{\beta_R + p(1-2\beta_R)}. \end{aligned}$$

The above two functions have the following properties. For $\beta_R \in (0,1)$ and $p \in \left(\frac{1}{2}, 1\right]$, it always holds that $0 \leq \beta_S^*(\beta_R, p) < \beta_S^{**}(\beta_R, p) \leq 1$. Also, $\beta_S^*(\beta_R, p)$ is decreasing in p while $\beta_S^{**}(\beta_R, p)$ is increasing in p. Finally, $\beta_S^*(\beta_R, \frac{1}{2}) = \beta_S^*(\beta_R, \frac{1}{2}) = 1-\beta_R$ while $\beta_S^*(\beta_R, 1) = 0$ and $\beta_S^{**}(\beta_R, 1) = 1$. As we shall see, for any given β_R these two functions divide the parameter space into three regions, each of which features a unique equilibrium prediction.

Proposition 1 *1. If $\beta_S = \beta_R$, then any disclosure strategy of S is an equilibrium disclosure strategy.*

2. Given $\beta_S \neq \beta_R$:

a) There exists no ND-equilibrium.

b) The D0-equilibrium exists if and only if $\beta_S \in (0, \beta_S^*(\beta_R, p)]$.

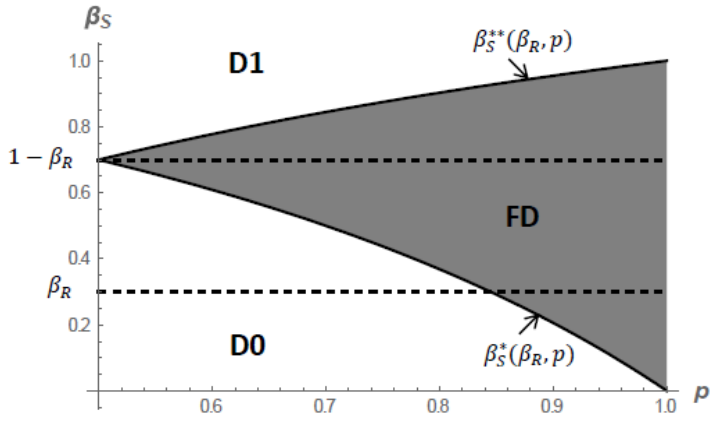

Figure 1: Equilibrium characterization in the baseline model.

c) The FD-equilibrium exists if and only if $\beta_S \in [\beta_S^*(\beta_R, p), \beta_S^{**}(\beta_R, p)]$.

d) The D1-equilibrium exists if and only if $\beta_S \in [\beta_S^{**}(\beta_R, p), 1)$.

e) Equilibria in mixed disclosure strategies exist if and only if $\beta_S \in \{\beta_S^*(\beta_R, p), \beta_S^{**}(\beta_R, p)\}$.

Figure 1 provides an illustration of our characterization for $\beta_R = 0.3$. The thick curves correspond to $\beta_S^*(0.3, p)$ and $\beta_S^{**}(0.3, p)$. Strictly between the two thick curves (in the gray area), only the FD-equilibrium exists. Instead, strictly above (below) of the upward (downward) sloping thick curve, only the D1 (D0) equilibrium exists. Finally, for $\beta_S = \beta_R$, the FD-, D0-, D1- and ND-equilibria exist for any $p \geq \frac{1}{2}$. Note that φ does not affect the parameter values for which the different types of equilibrium exist, and it is thus left unspecified for this figure.

Proposition 1 leads to the following corollary.

Corollary 1 *a) If* β_S *is sufficiently close to* $1-\beta_R$*, then FD is the unique equilibrium.*

b) Given $p < \max\left\{\frac{(1-\beta_R)^2}{(1-\beta_R)^2+(\beta_R)^2}, \frac{(\beta_R)^2}{(1-\beta_R)^2+(\beta_R)^2}\right\}$ *and* $\beta_R \neq 1/2$*, if* β_S *is sufficiently close (but not equal to)* β_R*, then FD is not an equilibrium.*

c) For given β_i*, the set of* β_j *for which FD exists is increasing in* p*. It is* $(0, 1)$ *if* $p = 1$.

d) If equilibrium features partial disclosure of the D0 or D1 type, the signal that is disclosed is the one that is congruent with the bias of the more confident player.

Summarizing, our characterization exhibits the following key properties:

1. Except under knife-edge conditions, there is a unique equilibrium.

2. Unless $\beta_S = \beta_R$, there exists no ND-equilibrium. The reason is that for any p and $\beta_S \neq \beta_R$, the disclosure of at least one type of signal (either 0 or 1) leads to a strict decrease in disagreement with respect to prior disagreement. This follows from the statistical property that (in this binary setup) from S's *ex ante* perspective an informative signal always reduces disagreement by moving R's belief towards his own prior in expectation.

3. Full disclosure is not always feasible. The intuition comes from contemplating the fact that belief updating has two dimensions: Direction and intensity. In our setup, players both update in the same direction after any signal (no polarization), but they update with different intensities. In Figure 2 below, the thin continuous curve shows $\widetilde{\beta}_i(1)$ as a function of β_i for $p = 0.85$ and the thick curve plots $\widetilde{\beta}_i(1) - \beta_i$, which is single peaked and concave in β_i. We see that very confident types update very little no matter the signal, while maximum updating arises for a prior moderately biased against the observed signal. Disagreement will increase after a signal if the player assigning the largest prior probability to the state indicated by the signal is also the player who updates the most. A 1-signal, for example, will increase disagreement if $\beta_i < \beta_j$ and β_j shifts upward more than β_i.

4. Point a) of Corollary 1 states that for any p, FD is possible if β_S is close enough to $1 - \beta_R$. Such S-prior can have a very different mean than R's prior but it has the same variance (i.e. exhibits the same confidence). A technical intuition is as follows. As noted above, for any β_S, β_R, p at least one signal (0 or 1)

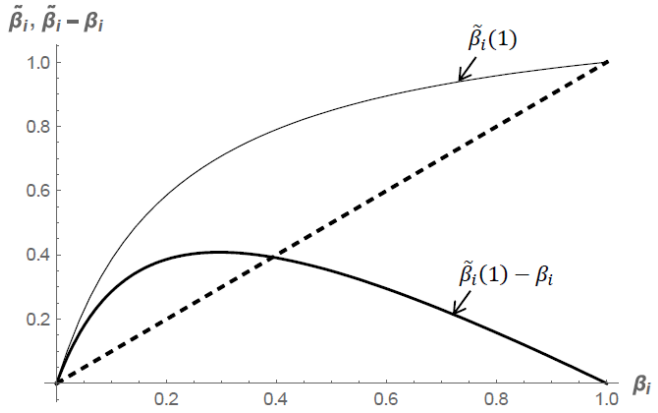

Figure 2: Intensity of belief updating given a 1-signal as a function of β_i.

strictly decreases disagreement with respect to the status quo. Next, note that if $\beta_S = 1 - \beta_R$, both signals yields the same posterior disagreement, i.e.

$$\widetilde{\beta}_S(1) - \widetilde{\beta}_R(1) = \widetilde{\beta}_S(0) - \widetilde{\beta}_R(0).$$

Since at least one signal strictly reduces disagreement, the other must achieve the same. Hence, FD is achievable for any p for $\beta_S = 1 - \beta_R$.[23] A more concrete intuition (see Figure 2) is that when priors are symmetric around $\frac{1}{2}$, the prior with the highest (lowest) prior mean moves strictly less than the other prior after a 1-signal (0-signal). Thus the difference in posterior means is always smaller than the difference in prior means.

5. Point b) of Corollary 1 states that for p low enough, FD is impossible if β_S is close enough (but not identical) to β_R. Note that prior variances are very similar if either prior means are approximately symmetric around $\frac{1}{2}$, or if these are approximately identical. Yet, full disclosure is significantly less robust in the latter case. The result shows that prior variances are not the only factor affecting disclosure and that prior means also play an important role. For an

[23] Note furthermore that updating prior β_S^* with a 1-signal or instead β_S^{**} with a 0-signal yields $1 - \beta_R$.

intuition, let both priors be strongly biased towards 0 and very close to each other. Given that $\widetilde{\beta}_i(1) - \beta_i$ is single peaked in β_i (see Figure 2), we see that after a 1-signal the player with the highest prior updates more intensely. In consequence, the spread between beliefs will increase after this signal.

6. Point c) of Corollary 1 means that a sufficiently precise signal allows for full disclosure. For an intuition, note that in the limit case of $p = 1$ any signal trivially reduces disagreement to 0. Low signal quality thus triggers two types of costs for R; exogenous and endogenous (i.e. strategic). The first is the lower informativeness of S's signals and the second is the lower informativeness of S's disclosure strategy.

7. For the intuition behind Point d) of Corollary 1, consider the case where the two players have opposite prior biases and let the most confident player be very confident and the other player's prior be close to $\frac{1}{2}$. The first player updates very little no matter the signal, so that her posterior is virtually identical to her prior no matter the signal observed. The moderate player instead updates significantly. Now, note that a signal congruent with (in contradiction with) the confident player's bias moves the belief of the moderate player closer to (away from) the confident player's prior.

2.3 Matching

2.3.1 Non-selective matching

Within a simple random matching setup, Point d) of Corollary 1 naturally implies that the more R's prior is biased towards a given state, the less likely she is to be exposed to information contradicting her prior. Assume that a given receiver $\widetilde{R}$, whose prior $\beta_{\widetilde{R}}$ is publicly observed, is randomly matched with a perceived disagreement averse sender whose publicly observed prior is drawn from the uniform distribution on $[0, 1]$.

Given $\beta_S, \beta_{\widetilde{R}}$ and p, the standard disclosure game ensues. We call this game *non-selective matching*.

The following result characterizes the confirmatory bias arising under non-selective matching.

Remark 1 *For any ω, p, under non-selective matching, the ex ante probability that $\widetilde{R}$ observes a 0-signal (1-signal) is decreasing (increasing) in $\beta_{\widetilde{R}}$.*

For instance, consider the case of a 0-signal. By Proposition 1 the ex ante probability that $\widetilde{R}$ is exposed to a 0-signal is the probability that S's signal is 0 and that β_S is such that the equilibrium is D0 or FD. This equals

$$\Pr[\sigma = 0\,|\omega\,]\Pr[\beta_S < \beta_S^{**}(\beta_{\widetilde{R}}, p)] = \Pr[\sigma = 0\,|\omega\,]\beta_S^{**}(\beta_{\widetilde{R}}, p),$$

which is strictly decreasing in $\beta_{\widetilde{R}}$.

Within a dynamic version of the above non-selective matching scenario where $\widetilde{R}$ repeatedly plays the same one-shot disclosure game against short-sighted senders (i.e. who do not update from observing $\widetilde{R}$'s period-t prior), perceived disagreement aversion on the part of senders thus slows down $\widetilde{R}$'s learning of the true state (i.e. causes inertia in beliefs) if the state is not congruent with $\widetilde{R}$'s extreme prior bias. Note that $\widetilde{R}$'s learning is only slowed down as opposed to entirely impeded, as $\widetilde{R}$ acknowledges that no disclosure by S does not necessarily imply that he holds no information.

2.3.2 Selective matching

In reality, individuals often choose their conversation partners and we now explore this possibility within the context of our model. We find that selective matching further increases the prospect of echo chambers: Individuals select matching partners with similar priors, which in turn induces partial and confirmatory disclosure.

We define the game of *selective matching* as follows. Suppose a large population of senders and receivers, all senders being (perceived) disagreement averse. As before,

senders and receivers are randomly matched and observe each others' priors. Yet, in contrast to non-selective matching, a match becomes *active* if and only if the sender accepts it. We focus on the sender's payoffs. If the match does not become active, the sender obtains a payoff of 0, which represents his outside option. If the match becomes active, the standard disclosure game introduced in section 2.1 ensues and the sender's final payoff equals $W > 0$ minus the ex post perceived disagreement after the disclosure stage. The interpretation of the payoffs is that psychological payoffs only arise once a sender explicitly decides to become involved in conversation. For simplicity, assume that the sender's match acceptance decision is made before his information is realized.

Let $E_S[\Delta|\beta_S, \beta_R]$ denote the sender's ex ante expectation of the receiver's ex post perceived disagreement given β_S, β_R conditional on the match becoming active.[24] It follows that S will accept a match with R if and only if

$$W \geq E_S[\Delta|\beta_S, \beta_R]. \tag{3}$$

Furthermore, $E_S[\Delta|\beta_S, \beta_R]$ satisfies the following description.

Proposition 2 *$E_S[\Delta|\beta_S, \beta_R]$ is continuous and V-shaped with respect to β_R, and it reaches its minimum of 0 at $\beta_R = \beta_S$.*

Figure 3 illustrates the above proposition. We assume $p = 0.9$, $\varphi = 0.6$, $\beta_S = 0.7$. The thick curve shows $E_S[\Delta|\beta_S, \beta_R]$ as a function of β_R. For $W = 0.1$ (represented by the horizontal dotted line), only values of β_R situated between β_{R1} and β_{R2} satisfy (3). This is formalized in the following corollary.

Corollary 2 *Given W, p, there are thresholds $\underline{\beta}_R < \beta_S < \overline{\beta}_R$ such that S accepts a match with R if and only if $\beta_R \in \left[\underline{\beta}_R, \overline{\beta}_R\right]$, where $\lim_{W \to 0}\{\underline{\beta}_R, \overline{\beta}_R\} = \beta_S$.*

[24] Note that $E_S[\Delta|\beta_S, \beta_R]$ is uniquely defined. In particular, if X and X' are two equilibrium disclosure rules given β_S, β_R, then $E_S[\Delta^X|\beta_S, \beta_R] = E_S[\Delta^{X'}|\beta_S, \beta_R]$ (as is shown in the proof of Proposition 2). Recall furthermore that by Proposition 1 there is a unique equilibrium disclosure rule except under knife-edge conditions.

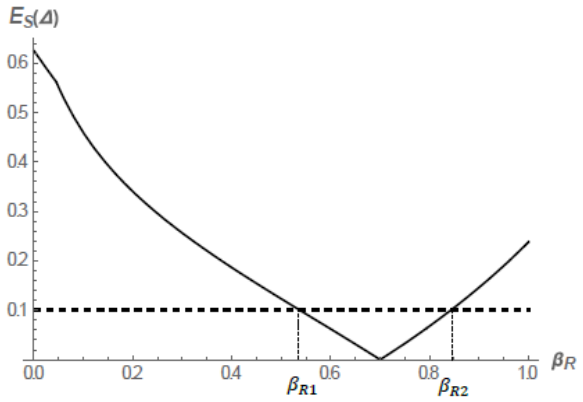

Figure 3: Perceived disagreement expected by S in equilibrium as a function of β_R.

The above corollary states that the only matches that become active are those involving players whose priors are sufficiently similar.

Proposition 1 and this corollary imply that the prospect of confirmatory information bias is strengthened under selective matching in societies that are sufficiently polarized. To see this, consider the following scenario. Priors belong either to $\left(0, \underline{\beta}\right)$ or $\left(\overline{\beta}, 1\right)$, where $\underline{\beta} < \frac{1}{2} < \overline{\beta} = 1 - \underline{\beta}$ and the conditional distribution of priors on each of the two intervals is uniform. Call this society $\left(\underline{\beta}, \overline{\beta}\right)$. We consider throughout a receiver $\widetilde{R}$ with $\beta_{\widetilde{R}} > \overline{\beta}$ (i.e. biased towards 1) and compare outcomes under respectively non-selective and selective matching.

Remark 2 *Consider a society* $\left(\underline{\beta}, \overline{\beta}\right)$. *Let* $\overline{\beta} > \frac{p-\sqrt{p(1-p)}}{2p-1}$ *and* $\beta_{\widetilde{R}} \geq \overline{\beta}$. *Given* ω, *for* W *small enough* $\widetilde{R}$ *observes a 0-signal with probability weakly larger than* $\frac{1}{2}\left(\frac{1-\beta_{\widetilde{R}}}{\underline{\beta}}\right) P[\sigma = 0\,|\omega]$ *under non-selective matching and instead with probability zero under selective matching.*

Under non-selective matching, with a probability bounded below by $\frac{1}{2}\left(\frac{1-\beta_{\widetilde{R}}}{\underline{\beta}}\right)$, $\widetilde{R}$'s match is such that the implied equilibrium is either FD or D0. To see this, note first that $\widetilde{R}$ is matched half of the time with a sender satisfying $\beta_S \in (0, \underline{\beta})$. Second, conditional on $\beta_S \in (0, \underline{\beta})$ the probability that $\beta_S \leq 1 - \beta_{\widetilde{R}}$ (yielding FD or D0 by

Corollary 1) is $\frac{1-\beta_{\widetilde{R}}}{\underline{\beta}}$. As long as $\beta_{\widetilde{R}}$ is not extremely high, there is thus a significant probability, weakly larger than $P[\sigma = 0\,|\omega]\frac{1}{2}\left(\frac{1-\beta_{\widetilde{R}}}{\underline{\beta}}\right)$, that $\widetilde{R}$ encounters a 0-signal.

Selective matching yields a very different picture. For W very small, by Corollary 2 a sender S will accept a match with a receiver R if and only if $\beta_R \approx \beta_S$. As a result, any active match in which $\widetilde{R}$ (recall $\beta_{\widetilde{R}} > \overline{\beta}$) participates will involve an S satisfying $\beta_S > \overline{\beta}$. If furthermore $\overline{\beta} > \frac{1}{2p-1}\left(p - \sqrt{p\,(1-p)}\right)$, it holds true that $\overline{\beta} > \beta_S^{**}(\beta_{\widetilde{R}}, p)$ so that by transitivity, any active match involving $\widetilde{R}$ satisfies $\beta_S > \beta_S^{**}(\beta_{\widetilde{R}}, p)$ and thus yields the D1 equilibrium by Proposition 1. Hence, $\widetilde{R}$ never observes a 0-signal no matter the state. Finally, note that $\frac{1}{2p-1}\left(p - \sqrt{p\,(1-p)}\right)$ is increasing in p and not very large as long as p is not very high (for $p = \frac{3}{4}$ it equals 0.634), which shows that weak societal polarization suffices to create strong echo chamber dynamics under selective matching.

2.4 The hidden cost of political correctness

Can S's attempt to minimize perceived disagreement be counter-productive from an ex ante perspective, thereby revealing a hidden cost of political correctness (relative to a hypothetical case of full transparency)? We address this question in two different ways: first, from S's own perspective in terms of perceived disagreement, and then from the perspective of a third party (e.g., a social planner) who cares about actual disagreement (i.e. would like to reduce social polarization).

First, from S's ex ante perspective, can the expected value of ex post *perceived* disagreement be higher in a (partial disclosure) equilibrium than it would be under (non-equilibrium) full disclosure? In such a case, S would prefer to commit to full disclosure if he could. This question is answered in the next proposition.

Proposition 3 *1. Let parameters be such that D1 is the unique equilibrium. If $\beta_S > \beta_R$, then S ex ante strictly prefers full disclosure over the D1-equilibrium. Vice versa if $\beta_S < \beta_R$.*

2. Let parameters be such that D0 is the unique equilibrium. If $\beta_S < \beta_R$, then S ex ante strictly prefers full disclosure over the D0-equilibrium. Vice versa if $\beta_S > \beta_R$.

In a partial disclosure equilibrium, S would thus ex ante prefer to instead commit to full disclosure if and only if he is the most confident player, which always holds true in D1 (D0) when $\beta_S > \beta_R$ ($\beta_S < \beta_R$). The intuition is as follows. In a partial disclosure equilibrium, e.g. D0, the omission of 1-signals has two countervailing effects. The upside is that S benefits from hiding a 1-signal once he holds it. The downside is that when S holds no signal, R interprets silence as a possible concealment of a 1-signal, which increases perceived disagreement relative to prior disagreement. The negative effect of equilibrium concealment overweighs its positive effect if S is the most confident party. Recall that in this case, S omits signals contradicting his bias in a partial disclosure equilibrium (see Corollary 1.d). But R places a higher weight on the state corresponding to the omitted signal that S does, which leads R to overweight (in S's eyes) the probability that such a signal is held (and omitted) by S, thereby inflating perceived disagreement after no disclosure. Instead, under full disclosure, R's prior does not affect her ex post perception of S's posterior (which is always common knowledge).

A second key question is whether from the ex ante perspective of a third party (TP) endowed with a prior $\widehat{\beta}$, the expected value of ex post *actual* disagreement can be higher in equilibrium than it would be under FD. I.e. would TP prefer a truthful sender or a perceived disagreement averse sender if aiming at minimizing the expected ex post actual disagreement? Note that actual disagreement is different from perceived disagreement. The actual disagreement given that S holds signal σ and discloses d is $\left|\widetilde{\beta}_S(\sigma) - \widetilde{\beta}_R(d)\right|$, where $\widetilde{\beta}_R(d)$ is pinned down by R's beliefs concerning S's disclosure rule. In what follows, if $\beta_i < \widehat{\beta} < \beta_j$, we say that S and R's priors are on different sides of $\widehat{\beta}$.

Proposition 4 *Let parameters be such that there exists no FD-equilibrium. In the eyes of a third party with prior $\widehat{\beta}$ the expected actual disagreement:*

1. is strictly larger in equilibrium than under FD if at least one of the following conditions holds:

a) S's and R's priors are on different sides of $\widehat{\beta}$,

b) R's prior is further away from $\widehat{\beta}$ than is S's prior.

2. is strictly smaller in equilibrium than under FD if the following two conditions hold simultaneously:

a) S's and R's priors are either both strictly smaller or both strictly larger than $\widehat{\beta}$,

b) S's prior is further away from $\widehat{\beta}$ than R's prior and is sufficiently close to the boundary.

Part 1 of the proposition finds that equilibrium information omission can indeed be counterproductive while Part 2 identifies conditions under which it is helpful. A general intuition behind our results is that TP expects new information to lead S's and R's beliefs to converge to her prior. The disclosure strategy of S affects only the speed of convergence of R's beliefs, as S's actual posterior beliefs are independent of his disclosure strategy.

In Point 1.a), S's and R's priors are on different sides of $\widehat{\beta}$. Here, given that S's and R's beliefs move closer to $\widehat{\beta}$ in expectation, they must also be moving closer to each other. Hence TP would prefer that both S and R learn as fast as possible and would thus prefer FD over partial disclosure. The second case in Point 1 is when β_S and β_R are on the same side of $\widehat{\beta}$, but R is further away. An instance of this is the case of $\widehat{\beta} < \beta_S < \beta_R$. Again TP expects S and R to converge to her prior $\widehat{\beta}$, i.e. to both decrease. R will move towards S (since R's prior decreases) but S will simultaneously move away from R (since S's prior also decreases). In consequence, TP would prefer to speed up R's convergence by giving her full information.

Point 2 describes the case when β_S and β_R are on the same side of $\widehat{\beta}$, but S is further away from $\widehat{\beta}$ and is close to the boundary. An instance of this is the case of $\widehat{\beta} < \beta_R < \beta_S \approx 1$. Here, both players' beliefs decrease. But decreasing R's belief

moves it away from S's. So TP would prefer to slow down R's learning and thus would choose partial disclosure.

2.5 Strangers' talk

Conversations often take place between parties who do not exactly know each others' priors but who might hold some relevant information concerning these priors (for example, by observing each other's accent, dressing style, profession, social networks). We now characterize equilibrium outcomes for a set of stylized scenarios featuring privately known priors.

Technically, since the measure of disagreement (1) is defined for given priors, the expected disagreement perceived by R under unknown β_S, given disclosure d, is

$$E_{R,\{\beta_S\}}[\Delta(d,\beta_S,\beta_R)] = E_{R,\{\beta_S\}}\left[\left|E_{R,\{\sigma\}}[\widetilde{\beta}_S(\sigma)\,|d] - \widetilde{\beta}_R(d)\right|\right],$$

where $E_{i,\{z\}}$ stands for an expectation over different possible realized values of the random variable z, as computed by i. Note that R takes the expectations sequentially: first, over all possible signal realizations to compute her second-order belief for given β_S, and only then over all possible realizations of β_S to compute the expected disagreement. In other words, she treats different cases of β_S as separate instances of disagreement. For example, an uninformed receiver with prior equal to 0.5 would not consider a sender with the same prior as disagreeing with her if S is known to hold either a 0- or a 1-signal (with 50% chance each), yet would treat their disagreement as having a positive value in case if S is known to hold the prior belief of either 0 or 1. Besides being more tractable, this approach reflects the intuition that disagreement caused by different priors is essentially more severe than the one caused by different private signals. Indeed, while the latter type of disagreement can generally be resolved by information exchange (Aumann, 1976), the disagreement caused by different prior beliefs cannot, since it is driven by a difference in initial worldviews which are beyond discussion.

In turn, the expected utility of S under privately known priors becomes

$$-E_{S,\{\beta_R\}}[E_{R,\{\beta_S\}}[\Delta(d,\beta_S,\beta_R)]] = -E_{S,\{\beta_R\}}E_{R,\{\beta_S\}}\left[\left|E_{R,\{\sigma\}}[\widetilde{\beta}_S(\sigma)\,|d] - \widetilde{\beta}_R(d)\right|\right].$$

These preferences give rise to the following equilibrium characterization.

Proposition 5 *Let priors be privately observed and drawn from publicly known distributions G_S and G_R, endowed with respective probability density functions g_S and g_R.*

a) If g_S and g_R are both symmetric around $1/2$, then there exists an FD-equilibrium.

b) If g_S and g_R are s.t. $g_S(x) = g_R(1-x)$ for all $x \in [0,1]$ (i.e. they are symmetric w.r.t. each other around $\frac{1}{2}$) and $\frac{g_S(x)}{g_R(x)}$ is monotone in x, then there exists an FD-equilibrium.

c) If g_S and g_R are identical and sufficiently skewed to the right (left), then there exists a $D1$ ($D0$) equilibrium, but no FD and $D0$ ($D1$) equilibrium.

d) If S's prior is commonly known and sufficiently close to $1/2$ while g_R is symmetric around $1/2$, then there exists an FD-equilibrium.

Point a) shows that two-sided uncertainty about priors is beneficial to disclosure if none of the two players is a priori biased in one or the other direction. If this condition is satisfied, this provides an argument for not encouraging revelation of information about respective biases (e.g., disclosing one's own prior political stance in a conversation). For an intuition, note that if g_S and g_R are both symmetric around $1/2$, in a putative FD-equilibrium the payoff from disclosing is the same no matter the signal held by S. Since it is impossible that *both* signals increase disagreement under FD (i.e. that an informative experiment increases expected disagreement from the ex ante perspective), this implies that both should at worst leave disagreement unchanged. Full disclosure is thus incentive compatible for S.

Point b) shows that if players are both a priori biased in different directions but in an equivalent (i.e. mirror-image like) fashion, then an FD-equilibrium exists. This

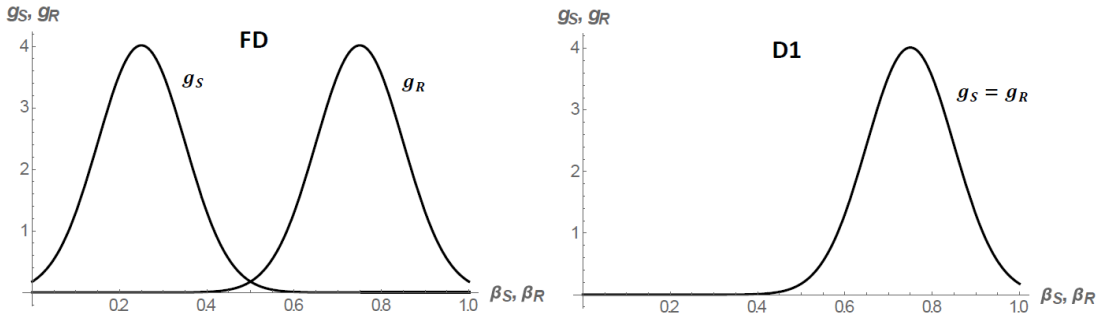

Figure 4: Equilibrium characterization under uncertainty about priors.

contrasts with point c), which states that FD may be infeasible if both priors are drawn from the same biased distribution. The findings of Points b) and c) echo Proposition 1, which highlights the benefit of diversity. For example, for $p = 0.7$, if β_S and β_R are both distributed according to the same truncated normal distribution with mean $3/4$ and standard deviation $\sigma = 0.3$, the only (pure strategy) equilibrium is D1. The FD-equilibrium instead exists if the distribution of β_R stays the same while the distribution of β_S is reflected around $1/2$, i.e. changed to a truncated normal with mean $1/4$ and standard deviation $\sigma = 0.3$. Figure 4 provides examples of profiles of distributions of prior beliefs, complemented by a description (in bold) of the implied equilibrium disclosure.

Finally, Point d) shows that two-sided uncertainty is not strictly necessary to ensure FD. The latter is feasible if S's prior is known and close to $\frac{1}{2}$ while R's prior is symmetrically distributed around $\frac{1}{2}$.

Note that all results in Proposition 5 hold in approximation, i.e. as one density function (uniformly) converges to the other density, or instead to the symmetric reflection around $\frac{1}{2}$ of the other density.[25]

[25] A formal proof is available upon request.

3 Extensions

In what follows, we consider a set of key extensions of our main setup. We first consider a general information structure with continuous signals satisfying the marginal likelihood ratio property (MLRP), while maintaining the assumption of a binary state. In the second subsection, we assume a continuous state space, considering a Normal priors-Normal signals setup. The third subsection finds that perceived disagreement aversion arises endogenously in a variety of simple dynamic games. In our final subsection, we consider disagreement aversion within a game of costly collective acquisition of public signals.

3.1 Binary state and continuous signals

We now show that our equilibrium characterization in the baseline setting carries over qualitatively to the case of an information structure with continuous signals satisfying MLRP. Assume that S's signal s is drawn from an interval $[\underline{s}, \overline{s}]$. Given state $\omega \in \{0, 1\}$, s is distributed according to $F(s|\omega)$ with continuous and differentiable density $f(s|\omega)$. Assume that $\frac{d}{ds}\frac{f(s|1)}{f(s|0)} > 0$ (MLRP), meaning that a higher signal implies a higher conditional probability of state 1. Upon learning s, the updated belief of i is

$$\widetilde{\beta}_i(s) = \frac{\beta_i f(s|1)}{\beta_i f(s|1) + (1-\beta_i) f(s|0)} = \frac{\beta_i}{\beta_i + (1-\beta_i)\frac{f(s|0)}{f(s|1)}},$$

which is increasing in s. Assume furthermore that the extreme signals $\underline{s}$ ($\overline{s}$) are such that $\lim_{s\to\underline{s}} \frac{f(s|1)}{f(s|0)} = 0$ and $\lim_{s\to\overline{s}} \frac{f(s|1)}{f(s|0)} = \infty$. Each of these two extreme signal realizations makes the observer (almost) sure that the state is 0 or 1, respectively. Note that there exists a threshold signal $\tilde{s} \in (\underline{s}, \overline{s})$ such that whatever $\beta_i \in (0, 1)$, it holds true that $\widetilde{\beta}_i(s) \gtreqless \beta_i$ for $s \gtreqless \tilde{s}$. Signal $\tilde{s}$ satisfies $f(s|0) = f(s|1)$ and we call it the uninformative signal. We say that signal $s > (<)\tilde{s}$ indicates state 1 (0). We say that signal $s > (<)\tilde{s}$ is congruent with j's prior bias if $\beta_j > (<)\frac{1}{2}$. We call the above

setup the *binary state-continuous signals environment*. We call *simple disclosure* equilibrium (SD equilibrium) an equilibrium featuring two thresholds $\underline{s} < s_1 < s_2 < \overline{s}$ such that S discloses s if and only if $s \leq s_1$ or $s \geq s_2$. As with the binary signals environment, we call full disclosure (FD) an equilibrium where S discloses all signals. We obtain the following equilibrium characterization.

Proposition 6 *1. If $\beta_S \in \{\beta_R, 1-\beta_R\}$ then there exists an FD-equilibrium. If $\beta_S \notin \{\beta_R, 1-\beta_R\}$, then the unique equilibrium is an SD equilibrium.*

2. In equilibrium, all signals congruent with the bias of the more confident player are disclosed. Signals contradicting the bias of the more confident player are partially disclosed.

The fundamental qualitative features of equilibrium echo those arising under binary signals. Except under knife-edged conditions, the equilibrium is unique. Only signals that are congruent with the prior of the more confident player are fully revealed. Furthermore, if $\beta_S = 1 - \beta_R$, a full disclosure equilibrium exists, implying that increasing prior misalignment can be helpful.

We now reexamine the issue of the hidden cost of political correctness already studied for the case of binary signals. Our original results (Propositions 2 and 3) carry over essentially identically to the continuous signals setup.

Proposition 7 *1. Let parameters be such that in the unique equilibrium, the non-disclosure interval contains signals indicating state 0. If $\beta_S > \beta_R$, then S ex ante strictly prefers full disclosure over the SD equilibrium. Vice versa if $\beta_S < \beta_R$.*

2. Let parameters be such that in the unique equilibrium, the non-disclosure interval contains signals indicating state 1. If $\beta_S < \beta_R$, then S ex ante strictly prefers full disclosure over the D0-equilibrium. Vice versa if $\beta_S > \beta_R$.

Proposition 8 *All the statements in Proposition 4 apply.*

3.2 Continuous state space and continuous signals

Assume that the state space is $\Re$. S and R's commonly known priors are normal and given by respectively $N(\mu_S, \gamma_S^2)$ and $N(\mu_R, \gamma_R^2)$. S is known to hold a signal with probability $\varphi \in (0,1)$. Given realized state ω, S's signal equals $\omega + \varepsilon$, where $\varepsilon \sim N(0, \gamma_\varepsilon)$, this being commonly known.[26] We denote signal realizations by σ. Note the standard result that

$$E_i[\omega|\sigma] = \frac{\mu_i \frac{1}{\gamma_i^2} + \sigma \frac{1}{\gamma_\varepsilon^2}}{\frac{1}{\gamma_i^2} + \frac{1}{\gamma_\varepsilon^2}}.$$

We provide an equilibrium characterization for the same one-shot disclosure game. S, if he holds a signal, is free to either disclose it or omit it. We say that a signal σ increases (decreases) disagreement if and only if $\Delta(\sigma) > (<) |\mu_S - \mu_R|$, i.e. if the distance between posterior means conditional on σ is larger (smaller) than that between prior means. We say that player i is more confident than j if and only if the variance of i's prior belief is smaller, i.e. $\gamma_i < \gamma_j$. This is analogous to our previous definition of confidence for the binary state case, which also implies that a more confident player has a smaller variance of prior beliefs.[27]

We obtain the following equilibrium characterization.

Proposition 9 *1. Let $\gamma_S^2 = \gamma_R^2$. If $\mu_S = \mu_R$, then any signal leaves disagreement equal to the (zero) prior disagreement and any disclosure rule is an equilibrium disclosure rule. If $\mu_S \neq \mu_R$, then any signal strictly decreases disagreement and the only equilibrium is FD.*

2. Let $\gamma_S \neq \gamma_R$ and $\mu_S = \mu_R$. Any signal other than $\sigma = \mu$ strictly increases disagreement and there exists no equilibrium in which any signal other than μ is disclosed. In any equilibrium, S discloses with ex ante probability zero.

[26] A previous version of this paper contains an analysis of the case where priors are beta distributions and signals are drawn according to a state-dependent binomial distribution. Results (available upon request) echo those obtained in the binary and normal environments, in that they highlight the central role of differences in prior variances.

[27] In the binary state case, the variance of prior beliefs of player i is given by $\beta_i(1-\beta_i)$, which is strictly decreasing in the distance of β_i from 1/2.

3. Let $\gamma_S \neq \gamma_R$ and $\mu_S \neq \mu_R$.

a) Any equilibrium features a finite $\eta > 0$ such that S discloses his signal if and only if $\sigma \in I = [\widetilde{\sigma} - \eta, \widetilde{\sigma} + \eta]$, where

$$\widetilde{\sigma} = \frac{\mu_S(\gamma_R^2 + \gamma_\varepsilon^2) - \mu_R(\gamma_S^2 + \gamma_\varepsilon^2)}{\gamma_R^2 - \gamma_S^2}.$$

b) In any equilibrium, the interval of disclosed signals is closer to the prior mean of the more confident player in terms of Hausdorff distance. In particular, $\widetilde{\sigma} \notin (\mu_S, \mu_R)$ and $\widetilde{\sigma}$ is strictly closer to the prior mean of the more confident player.

Point 1 states that if both priors have the same variance, then all signals weakly reduce disagreement, resulting in existence of the FD-equilibrium.[28] This is analogous to the binary state case, where FD exists when players are equally confident as they have identical prior variances (i.e. $\beta_S = \beta_R$ or $\beta_S = 1 - \beta_R$). Note that if and only if $\mu_S \neq \mu_R$, all signals strictly reduce disagreement and FD is the *unique* equilibrium, which indicates a positive role of differences in prior means, as in the binary case.

Points 2 and 3 consider the case of different prior variances. Point 2 assumes $\gamma_S \neq \gamma_R$ and $\mu_S = \mu_R = \mu$. Here, any signal $\sigma \neq \mu$ strictly increases disagreement, as posterior means always differ after disclosure. For any $\sigma \neq \mu$, the posterior mean of the more confident player is closer to μ than that of the other player, as a lower prior variance causes higher inertia in belief updating. In equilibrium, S always conceals $\sigma \neq \mu$ and thereby induces a perceived disagreement of 0. In consequence, S essentially never discloses (i.e. with ex ante probability 0).

Point 3.a states that given $\gamma_S \neq \gamma_R$ and $\mu_S \neq \mu_R$, equilibrium communication features a non-degenerate interval of signals that are disclosed. A difference in means, conditional on different (though potentially arbitrary close) variances, thus improves disclosure in comparison to the case of identical means. This echoes our finding for the binary model (cf. Corollary 1, points a) and b)). Qualitatively, the equilib-

[28] The result that all signals reduce disagreement under $\gamma_i = \gamma_j$ in the normal-learning setup has also been shown in Che and Kartik (2009).

rium exhibits the "opinion corridor" property. Only evidence that belongs to some predetermined interval I is disclosed. Again, the underlying mechanism is that the difference in belief inertias implies that sufficiently high and sufficiently low signals increase disagreement. Finally, Point 3.b is reminiscent of Corollary 1.d, obtained in the binary setting. The set of disclosed signals is biased towards the prior mean of the more confident player. Concluding, our main qualitative insights from the analysis of the discrete state space carry over to this continuous state space setup.

3.3 Instrumental disagreement aversion

Aversion to perceived disagreement on the part of a privately informed party might stem from the fact that it adversely affects subsequent interaction with the uninformed party. We here consider simple dynamic games in which the informed party (S) may disclose her private information in stage 1 to some another party (R), while in subsequent stages players make decisions which are payoff-relevant to both S and R and which depend on players' first- and second-order beliefs. In contrast to the previous analysis, S is not assumed to be intrinsically disagreement averse. We consider two setups matching this description in what follows and find that in both, S is *de facto* averse to perceived disagreement at the initial disclosure stage and acts accordingly. This endogenous preference in turn leads to the informational biases characterized previously. In all setups considered, the underlying environment is as in the main section. The state space is $\{0,1\}$, priors are commonly known and S is known to hold a binary signal of precision p with probability φ.

3.3.1 Delegated decision making

An uninformed principal (R) faces a potentially informed agent (S), both being risk neutral. The principal faces a *technical problem* and there are two potential approaches (0 and 1) for tackling it. One and only one of these actually solves the problem, but its identity is a priori unknown. We call the good approach (either 0

or 1) the state. With probability φ, the agent holds information concerning the state in the form of a binary signal of precision p. If the problem is tackled, this yields a payoff of $1+\tau$ to the principal, where $\tau \in [0,1]$. If not, the principal's payoff is 0. The commonly known prior probability attached by $i \in S, R$ to state 1 is denoted $\beta_i \in (0,1)$.

The game has two stages. Stage 1 is the disclosure game studied in the main section. In stage 2, after observing S's disclosure, R decides whether or not to attempt to tackle the problem by hiring S. If S is not hired, the problem remains untackled and R thus simply obtains a payoff of 0. If S is hired, the contract proposed by R specifies a reward of 1 if and only if the agent tackles the problem successfully (this outcome being observable). By hiring S, R incurs a privately observed and random (transaction) cost c, which is drawn from a uniform distribution on $[0,1]$. Let $I(k)$ be an indicator function, where outcome $k = 1(0)$ indicates success (failure), $I(1) = 1$ and $I(0) = 0$. Conditional on S being hired, outcome $k \in \{0,1\}$ being secured and the transaction cost being c, the payoff of R is thus $I(k)\tau - c$.

If hired, S has in total a unit of work time available and decides freely how much time to dedicate to each approach. S incurs a cost $-\frac{1}{2}e_z^2$ of working e_z units of time on project $z \in \{1,2\}$. The good approach is successful with probability e if e units of time are dedicated to it. The bad approach leads to failure for sure. Thus, conditional on hiring, efforts e_0 and e_1 and outcome $k \in \{0,1\}$, the payoff obtained by S is $I(k) - \frac{1}{2}e_0^2 - \frac{1}{2}e_1^2$. If S is not hired, her payoff is 0.

An equilibrium featuring the full disclosure strategy in stage 1 is called an FD-equilibrium. We refer to the disclosure game studied in the main section of the paper as the *simple disclosure game*. We obtain the following result.

Proposition 10 *For $X \in \{FD, D0, D1\}$, there exists an X-equilibrium if and only if there exists an X-equilibrium in the simple disclosure game.*

We prove the statement in what follows, proceeding by backward induction. We first consider the optimal action choice of the agent if hired. Let $\widetilde{\beta}_i(\sigma)$ denote the

posterior probability assigned by i to state 1 conditional on signal $\sigma \in \{0, 1, \varnothing\}$ in a putative FD-equilibrium, where $\varnothing$ stands for no signal. Given posterior belief $\widetilde{\beta}_S$, the agent solves

$$\max_{e_1,e_2} \left\{ \widetilde{\beta}_S e_1 + (1 - \widetilde{\beta}_S) e_2 - \frac{1}{2}(e_1)^2 - \frac{1}{2}(e_2)^2 \right\} \text{ s.t. } e_1 + e_2 \leq 1.$$

It is straightforward that the agent's optimal total effort will equal 1. Otherwise, increasing one of the two effort levels while keeping the other constant yields an increase in revenue. The maximization problem of the agent thus rewrites as:

$$\max_{x \in [0,1]} \left\{ \widetilde{\beta}_S x + (1 - \widetilde{\beta}_S)(1 - x) - \frac{1}{2}x^2 - \frac{1}{2}(1 - x)^2 \right\},$$

The first-order condition reads $2\widetilde{\beta}_S - 2x^* = 0$, yielding $x^* = \widetilde{\beta}_S$. The agent's optimal effort choice is thus to dedicate to each project a share of her total time equal to the probability that she assigns to the project being the good project.

We now consider the principal's hiring decision after observing the disclosure $d \in \{0, 1, \emptyset\}$. If she decides to hire, and expects the disclosure rule of S to be X, the principal expects to obtain the profit of $\tau \Pi^X(d)$ where, given the optimal effort choice of S shown above

$$\Pi^X(d) = \widetilde{\beta}_R^X(d) E_R^X[\widetilde{\beta}_S(\sigma) \,|d] + (1 - \widetilde{\beta}_R^X(d))(1 - E_R^X[\widetilde{\beta}_S(\sigma) \,|d]).$$

The principal thus hires if and only if c is smaller than $\tau \Pi^X(d)$ (i.e. if and only if hiring yields a net benefit). Note that the above function is maximized if one of the two extreme consensus scenarios is reached: $\widetilde{\beta}_R^X(d) = \widetilde{\beta}_S(d) = 0$ or $\widetilde{\beta}_R^X(d) = \widetilde{\beta}_S(d) = 1$. In other words, S exhibits a form of disagreement aversion at the disclosure stage, in attempting to maximize the probability of being hired (which increases with $\Pi^X(d)$).

We now examine the disclosure choice of the agent if she holds a signal $\sigma \in \{0, 1\}$. In a putative FD-equilibrium, let:

$$\Upsilon(\sigma, \beta_S, \beta_R) = \Pi^{FD}(\sigma) - \Pi^{FD}(\varnothing), \ \sigma \in \{0, 1\}.$$

Note that $\Upsilon(\sigma, \beta_S, \beta_R)\tau$ is thus the change in R's subjective expected payoff from hiring caused by S disclosing signal σ in a putative FD-equilibrium. Clearly, in the FD-equilibrium S has no strict incentive to deviate when holding a σ-signal if and only if $\Upsilon(\sigma, \beta_S, \beta_R) \geq 0$. In words, S discloses her signal only if the disclosure weakly increases the probability that she is hired (thereby obtaining a positive utility instead of 0). Now, it is easily shown that $\Upsilon(0, \beta_S, \beta_R)$ and $\Upsilon(1, \beta_S, \beta_R)$ are both positive if and only if

$$\beta_S \in \left[\frac{(1-p)(1-\beta_R)}{1-p+\beta_R(2p-1)}, \frac{p(1-\beta_R)}{\beta_R + p(1-2\beta_R)} \right].$$

This condition is equivalent to the one for FD appearing in Proposition 1.

Thus, in the considered game, perceived disagreement aversion on the side of S arises endogenously from S's incentive to convince R in stage 1 that S's effort allocation in stage 2 will be in line with R's view on the optimal effort allocation.[29]

3.3.2 Competing for authority

We here consider a game in which players compete for authority. The game has four stages. In stage 1, S can disclose her signal to R if she holds one. In stage 2, players engage in a Tullock contest to determine the assignment of authority (relevant for stage 4 later). Players simultaneously choose efforts and given efforts levels e_i, e_j, player i wins with probability $\frac{e_i}{e_i+e_j}$. In stage 3, S has a second opportunity to disclose her signal if she did not disclose it in stage 1 (i.e. S is not able to commit to non-disclosure in stage 1). In stage 4, the winner of the contest (who was selected in stage 2) picks an action $a \in \mathbb{R}$. Players' utility function is $-(\omega - a)^2 - \mu e_i$, where ω is the state of the world as before, and a is an action picked by the player who obtains the final decision authority.

[29] A previous version of this paper analyses an investment game which can be seen as a more general version of this game. S, a privately informed entrepreneur, seeks funding for a project from a risk-averse investor (R). For fixed beliefs about the state, S's expected profit is increasing in the amount invested by R. As a result, the investor's expected payoff in equilibrium becomes a function of the uncertainty about the state and belief disagreement. We find that maximal disclosure takes place under moderate prior misalignment (analysis available upon request). Specifically, the disclosure optimal R-prior is located strictly between β_S and $1 - \beta_S$.

We characterize conditions under which there exists an equilibrium with full disclosure (FD) already in stage 1, so that both the Tullock contest and the final action choice happen under full information. We first provide a necessary condition and then provide a sufficient condition. We introduce the following objects:

$$\widehat{\beta}_S^*(\beta_R, p) = \frac{\beta_R(1-p)(1+\beta_R(2p-1))}{\beta_R(1-\beta_R)(1-2p)^2 + 2(1-p)p},$$

$$\widehat{\beta}_S^{**}(\beta_R, p) = \frac{\beta_R p(1-\beta_R(2p-1))}{\beta_R(1-\beta_R)(1-2p)^2 + 2(1-p)p},$$

$$I_{outcome}(\beta_R, p) = \left[\widehat{\beta}_S^*(\beta_R, p), \widehat{\beta}_S^{**}(\beta_R, p)\right],$$

$$I_{disagreement}(\beta_R, p) = \left[\beta_S^*(\beta_R, p), \beta_S^{**}(\beta_R, p)\right],$$

where $\beta_S^*()$ and $\beta_S^{**}()$ were defined in section 2.1.

Proposition 11 *There exists an equilibrium featuring full disclosure in stage 1:*

a) only if β_S belongs to $I_{disagreement}(\beta_R, p)$,

b) if β_S belongs to the intersection of $I_{disagreement}(\beta_R, p)$ and $I_{outcome}(\beta_R, p)$.

Figure 5 shows the intervals $I_{outcome}$ and $I_{disagreement}$, assuming $\beta_R = 0.3$. The interval $I_{outcome}$ is located between the two dashed curves and always contains β_R. The interval $I_{disagreement}$ (which is the one appearing in Proposition 1) is located between the two continuous curves and always contains $1 - \beta_R$. For any given p, FD in period 1 requires that β_R is between the two continuous curves. Second, there exists an equilibrium with FD in period 1 if β_R is located between the two continuous curves as well as between the two dashed curves. We see that for p not too high, full disclosure in stage 1 is achievable only if β_S is moderately different from β_R. The main qualitative insight is that some prior misalignment can thus be essential to ensure an equilibrium featuring FD in stage 1.

The interval $I_{outcome}$ corresponds to the complete set of values of β_S for which S strictly favours disclosing no matter his signal if his concern is to minimize $-(\omega - a)^2$

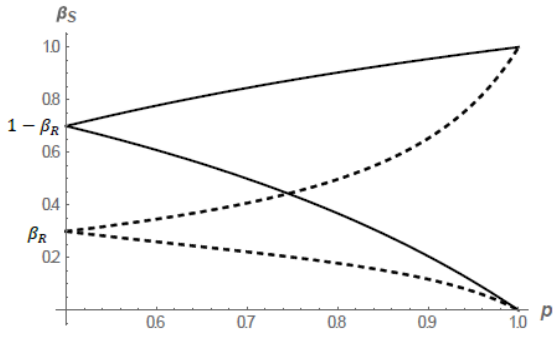

Figure 5: Disclosure-relevant intervals of β_S.

under the assumption that R has authority (and thus anticipating that R ultimately takes action $a = \widetilde{\beta}_R(d)$). The interval $I_{disagreement}$ is familiar from our main analysis. It corresponds to the values of β_S for which S strictly favours disclosing no matter his signal if his concern is to minimize perceived disagreement after disclosure.

In a putative equilibrium with FD in stage 1, S will disclose in stage 1 only if disclosure leads to weakly lower effort by R at the contest stage than non-disclosure. This, in turn, is true if and only if disclosure reduces perceived disagreement by the end of stage 1, which is equivalent to the requirement that β_S belongs to $I_{disagreement}(\beta_R, p)$. This explains Proposition 11.a. At the same time, full disclosure is optimal from the perspective of stage 3 if β_S belongs to $I_{outcome}(\beta_R, p)$. Hence, a sufficient condition for the equilibrium with FD in stage 1 is that β_S belongs to the intersection of $I_{disagreement}(\beta_R, p)$ and $I_{outcome}(\beta_R, p)$ (Proposition 11.b). Note finally that if β_S belongs to $I_{disagreement}(\beta_R, p)$ but not to $I_{outcome}(\beta_R, p)$, it is unclear whether or not there exists an equilibrium with FD in period 1. S indeed potentially faces a trade-off. While full disclosure minimizes the effort level of R at the contest stage, it suboptimally inflects R's action choice (which hurts S if R is ultimately assigned authority).

3.4 Joint observation of public signals

We here study the following simple game of voluntary and costly collective exposure to a public signal. The state ω belongs to $\{0,1\}$. Both players' utility function contains the loss from perceived disagreement as in (1), minus an extra i.i.d. cost of participation drawn from the uniform distribution on $[0,1]$. In stage 1, each player decides whether or not to participate after privately observing her cost c_i of participating. In stage 2, if both have decided to participate, players incur the participation cost and both observe the same randomly drawn public binary signal taken from $\{0,1\}$ which is identical to the state with probability p. If at least one of the agents has opted against participating, players incur no cost and no signal is observed. We call agents x and y, where agent $k \in \{x,y\}$ assigns prior probability k to state 1. Note that the environment is essentially non-strategic: Each player faces a simple decision problem and prefers to participate if and only if the expected reduction in perceived disagreement, conditional on joint observation of the signal, is larger than the private cost c_i of participating.

The following expression measures the ex post difference in beliefs conditional on a given public signal:

$$D_i(x,y,p) = \left|P(\omega = 1 \,|\sigma = i, x\,) - P(\omega = 1 \,|\sigma = i, y)\right|, \text{ for } i \in \{0,1\}\,.$$

From the perspective of agent $k \in \{x,y\}$, the expected posterior difference in beliefs conditional on joint exposure to a signal of quality p is thus given by:

$$\Lambda^k(x,y,p) = P(\sigma = 0\,|k)D_0(x,y,p) + P(\sigma = 1\,|k)D_1(x,y,p).$$

Note that $\Lambda^k(x,y,\frac{1}{2})$ is simply the prior disagreement. The value of a signal of quality p to player $k \in \{x,y\}$ is thus:

$$V^k(x,y,p) = \Lambda^k\left(x,y,\frac{1}{2}\right) - \Lambda^k(x,y,p).$$

Clearly, player k decides to participate if and only if $c_k \leq V^k(x,y,p)$. We obtain the

following characterization of the value of participating for each player.

Proposition 12 *1. For given x and $p > \frac{1}{2}$, $V^x(x,y,p) \geq 0$ for any y, while $V^x(x,y,p) = 0$ if and only if $y \in \{0,x,1\}$.*

2. For given x, $V^x(x,y,p)$ is single peaked in y on $(0,x)$ and on $(x,1)$.

3. For given $x \geq 1/2$, $V^x(x,y,p)$ reaches its maximum for $y = y^* \in (0,1/2)$. For given $x < 1/2$, $V^x(x,y,p)$ reaches its maximum for $y = y^* \in (1/2,1)$

Point 1 states that as long as $y \notin \{0,x,1\}$, an informative public signal strictly reduces the expected value of ex post disagreement from the ex ante perspective of both players.[30]

Note that the marginal value of participating is trivially 0 if parties share the same prior (in which case there is no disagreement both before and after the signal), or if the prior of one party equals 0 or 1 (in which case the latter party does not update). Point 2 states that a player's willingness to participate is maximized when her opponent has a moderately different prior. Intuitively, some degree of prior disagreement gives sufficient scope for disagreement reduction, and hence stimulates signal acquisition. At the same time, too extreme priors lead to stickiness of beliefs. Point 3 states that player k's optimal conversation partner (i.e. the partner maximizing k's participation incentive) is always biased in the opposite direction. Figure 6 illustrates this result by showing the expected value of a joint signal as a function of the opponent's prior for $x = 0.3$ and $p = 0.8$.

Next, consider a social planner who designs a two-members committee with the objective of maximizing the probability that a signal is acquired by the committee. We can show that this probability is maximized if the experts have priors that are symmetric around $\frac{1}{2}$ (i.e. oppositely biased) and non-radical.

[30]Note that this property does not generalize to continuous state environments. For instance, in the normal-normal setup considered earlier, an informative experiment can increase a player's ex ante expectation of ex post perceived disagreement, as compared to prior disagreement. Indeed, consider the case of identical prior means and different variances. In this case, a signal leads with probability one to a strictly positive ex post disagreement that is larger than prior disagreement of 0. In consequence, the value of jointly observing a signal is negative for both parties.

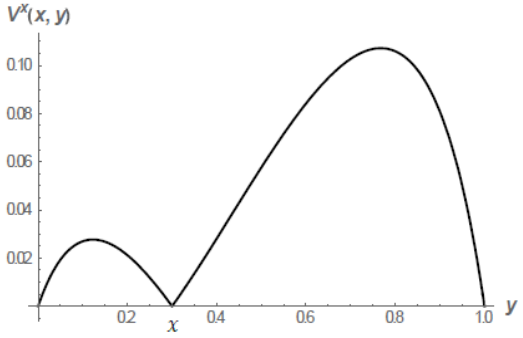

Figure 6: Expected value of a joint signal as a function of the opponent's prior.

Proposition 13 *For given p, there is a unique pair $\{x^*, y^*\}$ maximizing the probability of signal acquisition. For this pair, it holds true that $y^* = 1 - x^*$ and $x^* \notin \{0, \frac{1}{2}, 1\}$.*

For an intuition, consider the case when $V^x(x, y, p) \leq V^y(x, y, p)$. Since both players should agree to participate while one can show that V^k is a linear function of k for $k \in \{x, y\}$, this constraint should be binding at the optimum so that $V^x(x, y, p) = V^y(x, y, p)$. In turn, for non-identical priors such symmetry is achieved if and only if $y^* = 1 - x^*$. Finally, the priors should not be too extreme to ensure sufficient belief updating after a signal.

4 Conclusion

This paper introduces a new type of belief-dependent preferences reflecting an aversion to perceived disagreement. Our analysis has identified a range of implications for important instances of strategic communication and social learning. Disagreement aversion often leads to biases in information disclosure, in which case selective disclosure is biased towards the prior mean of the most confident player. Such disclosure bias may in turn be counterproductive from an ex ante perspective, in terms of minimizing ex post perceived disagreement or actual disagreement in beliefs. Generally, more similar prior variances beneficially affect disclosure while some heterogeneity in

prior means is helpful. If matching of informed and uninformed parties is endogenous, informed parties unfortunately prefer to interact with parties whose prior is similar, leading to incomplete disclosure featuring confirmatory bias. Finally, disagreement aversion can stimulate social learning in heterogeneous groups.

Our results provide a plausible explanation for stylized facts such as echo chambers and increasing social polarization. Further work building on the assumption of disagreement-aversion might provide more insight into the causes and consequences of contemporary societal patterns of belief heterogeneity.

References

Acemoglu, D., V. Chernozhukov, and M. Yildiz (2016). Fragility of asymptotic agreement under Bayesian learning. *Theoretical Economics 11*(1), 187–225.

Acemoglu, D., V. Chernozhukov, M. Yildiz, et al. (2007). Learning and Disagreement in an Uncertain World. Technical report, Collegio Carlo Alberto.

Andreoni, J. and T. Mylovanov (2012). Diverging opinions. *American Economic Journal: Microeconomics 4*(1), 209–232.

Asch, S. E. (1955). Opinions and social pressure. *Readings about the social animal 193*, 17–26.

Aumann, R. J. (1976). Agreeing to disagree. *The annals of statistics 4*(6), 1236–1239.

Austen-Smith, D. and T. J. Feddersen (2006). Deliberation, preference uncertainty, and voting rules. *American political science review 100*(2), 209–217.

Baliga, S., E. Hanany, and P. Klibanoff (2013). Polarization and ambiguity. *The American Economic Review 103*(7), 3071–3083.

Banerjee, A. and R. Somanathan (2001). A simple model of voice. *The Quarterly Journal of Economics 116*(1), 189–227.

Battigalli, P. and M. Dufwenberg (2007). Guilt in games. *The American economic review 97*(2), 170–176.

Battigalli, P. and M. Dufwenberg (2009). Dynamic psychological games. *Journal of Economic Theory 144*(1), 1–35.

Bénabou, R. (2012). Groupthink: Collective delusions in organizations and markets. *The Review of Economic Studies 80*, rds030.

Benabou, R. and G. Laroque (1992). Using privileged information to manipulate markets: Insiders, gurus, and credibility. *The Quarterly Journal of Economics 107*(3), 921–958.

Buechel, B., T. Hellmann, and S. Klößner (2015). Opinion dynamics and wisdom under conformity. *Journal of Economic Dynamics and Control 52*, 240–257.

Bursztyn, L., G. Egorov, and S. Fiorin (2017). From extreme to mainstream: How social norms unravel. Technical report, National Bureau of Economic Research.

Che, Y.-K. and N. Kartik (2009). Opinions as incentives. *Journal of Political Economy 117*(5), 815–860.

Coughlan, P. J. (2000). In defense of unanimous jury verdicts: Mistrials, communication, and strategic voting. *American Political science review 94*(2), 375–393.

Dandekar, P., A. Goel, and D. T. Lee (2013). Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences 110*(15), 5791–5796.

Deutsch, M. and H. B. Gerard (1955). A study of normative and informational social influences upon individual judgment. *The journal of abnormal and social psychology 51*(3), 629.

Dixit, A. K. and J. W. Weibull (2007). Political polarization. *Proceedings of the National Academy of Sciences 104*(18), 7351–7356.

Domínguez, D., F. Juan, S. A. Taing, and P. Molenberghs (2016). Why do some find it hard to disagree? An fMRI study. *Frontiers in human neuroscience 9*, 718.

Dufwenberg, M. and G. Kirchsteiger (2004). A theory of sequential reciprocity. *Games and economic behavior 47*(2), 268–298.

Ely, J., A. Frankel, and E. Kamenica (2015). Suspense and surprise. *Journal of Political Economy 123*(1), 215–260.

Ely, J. C. and J. Välimäki (2003). Bad reputation. *The Quarterly Journal of Economics 118*(3), 785–814.

Estlund, D. M. (2009). *Democratic authority: A philosophical framework*. Princeton University Press.

Feddersen, T. and W. Pesendorfer (1998). Convicting the innocent: The inferiority of unanimous jury verdicts under strategic voting. *American Political science review 92*(1), 23–35.

Festinger, L. (1950). Informal social communication. *Psychological review 57*(5), 271.

Festinger, L. (1957). *A theory of cognitive dissonance*. Evanston, Il: Row, Peterson.

Geanakoplos, J., D. Pearce, and E. Stacchetti (1989). Psychological Games and Sequential Rationality. *Games and Economic Behavior 1*, 60–79.

Gentzkow, M. and J. M. Shapiro (2006). Media bias and reputation. *Journal of political Economy 114*(2), 280–316.

Glaeser, E. L. and C. R. Sunstein (2009). Extremism and social learning. *Journal of Legal Analysis 1*(1), 263–324.

Goffman, E. (1959). *The presentation of self in everyday life.* Garden City, NY: Doubleday Anchor Books.

Golman, R., G. Loewenstein, K. O. Moene, and L. Zarri (2016). The preference for belief consonance. *The Journal of Economic Perspectives 30*(3), 165–187.

Golub, B. and M. O. Jackson (2012). How homophily affects the speed of learning and best-response dynamics. *The Quarterly Journal of Economics 127*(3), 1287–1338.

Grossman, S. J. (1981). The Informational Role of Warranties and Private Disclosure about Product Quality. *The Journal of Law & Economics 24*(3), 461–483.

Homans, G. C. (1961). Human behavior: Its elementary forms.

Huckfeldt, R., P. E. Johnson, and J. Sprague (2004). *Political disagreement: The survival of diverse opinions within communication networks.* Cambridge University Press.

Huston, T. L. and G. Levinger (1978). Interpersonal attraction and relationships. *Annual review of psychology 29*(1), 115–156.

Jung, W.-O. and Y. K. Kwon (1988). Disclosure When the Market Is Unsure of Information Endowment of Managers. *Journal of Accounting Research 26*(1), 146–153.

Kartik, N., F. X. Lee, and W. Suen (2015). Does competition promote disclosure. Technical report.

Landemore, H. and J. Elster (2012). *Collective wisdom: Principles and mechanisms.* Cambridge University Press.

Lazarsfeld, P. F. and R. K. Merton (1954). Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society 18*(1), 18–66.

Levy, G. (2007). Decision making in committees: Transparency, reputation, and voting rules. *American economic review 97*(1), 150–168.

Levy, G. and R. Razin (2015). Correlation neglect, voting behavior, and information aggregation. *American Economic Review 105*(4), 1634–45.

Loury, G. C. (1994). Self-censorship in public discourse: a theory of "political correctness" and related phenomena. *Rationality and Society 6*(4), 428–461.

Milgrom, P. and J. Roberts (1986). Relying on the Information of Interested Parties. *The RAND Journal of Economics 17*(1), 18–32.

Milgrom, P. R. (1981). Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics 12*(2), 380–391.

Morris, S. (1995). The common prior assumption in economic theory. *Economics & Philosophy 11*(2), 227–253.

Morris, S. (2001). Political correctness. *Journal of political Economy 109*(2), 231–265.

Mutz, D. C. (2006). *Hearing the other side: Deliberative versus participatory democracy*. Cambridge University Press.

Newcomb, T. M. (1961). *The acquaintance process*. Holt, Rinehart & Winston.

Ottaviani, M. and P. N. Sørensen (2006a). Reputational cheap talk. *The Rand journal of economics 37*(1), 155–175.

Ottaviani, M. and P. N. Sørensen (2006b). The strategy of professional forecasting. *Journal of Financial Economics 81*(2), 441–466.

Prendergast, C. (1993). A theory of "yes men". *The American Economic Review 83*(4), 757–770.

Prentice, D. A. and D. 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.

Rabin, M. (1993). Incorporating Fairness Into Game Theory and Economics. *American Economic Review 83*, 1281–1302.

Rosenberg, M. (1954). Some determinants of political apathy. *Public Opinion Quarterly 18*(4), 349–366.

Sethi, R. and M. Yildiz (2012). Public Disagreement. *American Economic Journal. Microeconomics 4*(3), 57.

Sethi, R. and M. Yildiz (2016). Communication with unknown perspectives. *Econometrica 84*(6), 2029–2069.

Shin, H. S. (1994a). The burden of proof in a game of persuasion. *Journal of Economic Theory 64*(1), 253–264.

Shin, H. S. (1994b). News management and the value of firms. *The RAND Journal of Economics 25*(1), 58–71.

Sobel, J. (1985). A Theory of Credibility. *Review of Economic Studies 52*, 557–573.

Sobel, J. (2013). Giving and receiving advice. In M. Acemoglu, D. Arellano and E. Dekel (Eds.), *Advances in Economics and Econometrics: Tenth World Congress*, pp. 305–341. New York: Cambridge University Press.

Sunstein, C. R. (2007). *Republic. com 2.0*. Princeton University Press.

Visser, B. and O. H. Swank (2007). On committees of experts. *The Quarterly Journal of Economics 122*(1), 337–372.

Vives, X. (2010). *Information and learning in markets: the impact of market microstructure*. Princeton University Press.