Disliking to disagree*

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

August 23, 2018

Abstract

Abundant evidence from social psychology suggests that people dislike openly disagreeing with each other. We study the implications of perceived disagreement aversion in a strategic disclosure setting where a sender faces a receiver with a different prior. With a binary state, excluding the knife-edge case of identical priors, full disclosure of the sender's information is feasible only if the receiver's prior is close enough to the symmetric of the sender's prior. If full disclosure is infeasible, only information congruent with the prior bias of the most extreme player is fully disclosed. Minimizing perceived disagreement can paradoxically be counterproductive from an ex ante perspective. Disagreement averse players prefer to be matched with players whose priors are similar, thus providing the basis for echo chambers under endogenous sorting. The effect of prior heterogeneity under unknown priors echoes that observed under known priors. Perceived disagreement aversion arises endogenously within simple games of delegation, relation-specific investment and competitive authority assignment. Finally, in committees featuring disagreement averse players, moderate prior heterogeneity encourages public information acquisition

Keywords: strategic disclosure, psychological games, disagreement aversion

JEL classification: D81, D83, D91

*We thank participants and discussants at the following workshops and seminars for helpful comments and suggestions: Toulouse IAS 2018 deliberation conference, Paris-Cergy IAS 2018 deliberation conference, Warwick Dr@w seminar 2018, Psychological Game theory 2017 conference in Norwich, Norwich 2017 Economic Theory workshop, 2017 Cologne behavioral economics brown bag seminar.

†University of Bonn. E-mail: fhoffmann@uni-bonn.de.

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

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

Citizens' political behavior is heavily influenced by the information available to them, which is a key reason why political parties, lobbies and governments often devote significant amounts of resources to influencing centralized information flows (campaigns, propaganda, censorship). But citizens also obtain information in a decentralized fashion from talking to each other. The latter channel has arguably gained in relative importance in the digital era (Internet, social media).[1]

Decentralized information exchange within social networks however exhibits many forms of bias. 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*.

Two aspects appear to play an important role in generating biases, namely the tendency to avoid open conflict of opinion and the heterogeneity in prior beliefs across individuals. This paper focuses on the consequences of these two features for central forms of social learning. A main source of tension is that when priors differ, while any informative experiment on average reduces disagreement, particular signal realizations can increase disagreement. While our main focus is on strategic information disclosure, we also shed light on the choice of interaction partners and the acquisition of public information in groups. Central questions addressed are as follows. How is the informativeness of equi-

[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."*

librium disclosure affected by prior disagreement and respective prior biases? Can the practice of avoiding disagreement be counterproductive from an ex ante perspective? Do people prefer to be matched with individuals who have similar priors? Which individual matches of individuals give rise to the highest incentive to acquire information?

Mutz (2006) reviews a number of studies showing that Americans avoid discussing politics with non like-minded people for fear of creating tensions.[4] A large body of experimental and empirical evidence documents that individuals tend to state opinions that conform to what they believe others think. 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 (before it became apparent that such views were shared by a large fraction of the population). Prentice and Miller (1993) established that students refrained from expressing dissent with campus alcohol practices based on the erroneous conviction that they held a potentially stigmatizing minority view. 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. Importantly, Deutsch and Gerard (1955) showed that this effect is weaker if subjects report their judgment privately, so that others' *perceived disagreement* is unaffected.

Disagreement aversion has many potential causes[5]. 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. 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 dif-

[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 descried 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]See Golman et al. (2016) for a general review of what the authors term a preference for *belief consonance*.

ferent labor market organisations.

Heterogeneous prior beliefs are an integral part of many social situations. Instances range from views on general political questions (climate change, immigration, free trade, religion and its consequences) to how to manage a firm or optimize an investment portfolio. A key underlying cause is that people have different personal histories (experiences, socialization, education).[6]

The main section of the paper (section 1) examines a simple game of disclosure by a potentially informed sender (S) who is averse to disagreement as perceived by an uninformed receiver (R). The state of the world is binary (0 or 1) and S and R have different publicly observed prior beliefs β_S and β_R that the state is 1. A binary informative signal $\sigma \in \{0, 1\}$ of commonly known precision p is available to S with some commonly known probability φ.

Our equilibrium characterization exhibits the following key properties. First, except under knife-edge conditions there always exists a unique equilibrium. Second, full disclosure is not always an equilibrium outcome. Third, increasing the difference in priors can imply better information transmission: For given p, β_R and excluding the knife-edge case of $\beta_S = \beta_R$, full disclosure is feasible only if S's prior is close enough to $1 - \beta_R$. Communication thus improves in the degree of symmetry of priors around $\frac{1}{2}$: Being similarly extreme is beneficial conditional on biases being opposed. Fourth, better information quality is always helpful: The higher p, the larger the set of values of β_S for which full disclosure is feasible. Fifth, if disclosure is partial, S only reveals information congruent with the most extreme player's prior bias. As we discuss, this can generate echo chambers like dynamics (i.e. confirmatory information bias) in a random matching setup.

Subsection 1.2 takes an ex ante perspective on equilibrium and disagreement. We show that the practice of avoiding perceived disagreement can backfire from an ex ante perspective. In other words, in some instances political correctness involves hidden costs. In the eyes of S, (ex ante) expected perceived disagreement can be higher in equilibrium than it would be under full disclosure, implying that S would prefer to commit to full disclosure. Furthermore, in the eyes of a third party with a prior potentially different

[6] 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.

from S's and R's and who cares about reducing *actual* disagreement between S and R, the equilibrium outcome with a perceived disagreement-averse sender can be worse than full disclosure. Subsection 1.3 considers disclosure under uncertainty about priors, which is in many contexts very realistic. We obtain results concerning the positive implications of prior heterogeneity which echo those obtained under known priors. Subsection 1.4 shows that if people can choose whom to be matched with, they prefer individuals whose prior is similar. Combining this feature with our equilibrium characterization, we conclude that selective matching further reinforces echo chamber dynamics as compared to the benchmark case of random matching.

Section 2 examines how our findings extend in a variety of fundamental directions. Subsection 2.1 tests the robustness of our results to an information structure featuring continuous signals satisfying the MLRP property. The main qualitative features of our characterization survive. Subsection 2.2 identifies a variety of games in which a disclosure stage is followed by one or several stages of decision making that generate material payoffs for both parties (e.g. delegated decision making, relation-specific investment, competition for authority). We find that in equilibrium, the privately informed party acts *as if* disagreement averse at the disclosure stage in his quest to strategically influence subsequent decision making. In consequence, the same informational biases and positive implications of heterogeneity arise in such contexts as in our main characterization. Subsection 2.3 assumes that R is interested in learning the state and examines implications of our preceding analysis. Finally, subsection 2.4 considers 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 the same underlying problem of learning in groups with heterogenous prior beliefs. We find that moderate disagreement in prior beliefs optimally incentivizes information acquisition, in a way that echoes our main findings.

Our theory offers a putative explanation of the following two stylized facts (call these A and B). First, many citizens are exposed disproportionately to information that confirms their worldview and thus evolve within so-called echo chambers. Second, social psychologists have documented 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.[7] Our tentative explanation of these facts rests on rationality, heterogeneous priors and aversion to (perceived) disagreement. First, our model predicts confirmatory bias given like-minded matches. Assume intermediate information precision and consider matches of individuals with very similar priors (and thus similar prior biases). For such pairs, only information congruent with the shared bias will be disclosed in equilibrium. Second, our model predicts a preference for interacting with people with similar priors, in (correct) anticipation of subsequent equilibrium disclosure. It follows that if people can choose whom to interact with, assortative matching will take place.

Can the same stylized facts be explained by other theories and if so, to what extent is our theory more compelling? An alternative theory 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 the association of A and B. Yet another theory is to simply assume that 1) people naturally associate with others who have similar worldviews and that 2) people with similar priors share the same confirmation bias. In consequence, the facts that people share naturally tend to be confirming facts. In this theory, it is unclear why rational people would associate with others who have the same confirmation bias as themselves,

[7] 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 (2006), 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.*".

if learning the truth is their ultimate goal. Yet another theory is to retain 1) and assume that 2) people are naive about the information held and provided by others with similar beliefs: The reason they have similar beliefs is that they in fact have learned the same type of information. Their worldviews are in fact posteriors and not priors. This is essentially an instance of correlation neglect and its corollary, namely overweighting of others' information, as explored in Levy and Razin (2015, 2016) and Glaeser and Sunstein (2009). The theory assumes an element of bounded rationality, which is not the case of our theory.

Literature review In its foundations, our paper relates to a literature studying how public information relates to disagreement in beliefs. A much studied phenomenon is 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). Under certain conditions, disagreement in beliefs may persist in the long run, i.e. asymptotically (Acemoglu et al., 2016; Andreoni and Mylovanov, 2012).[8] 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.

An extensive body of research dating back to Crawford and Sobel (1982) and Milgrom (1981) studies strategic information transmission between an informed sender (S) and an uninformed receiver (R), either in the form of cheap talk or in the form of disclosure of verifiable signals.[9] These models typically involve a difference in players' preferences over R's action conditional on the state. Newer papers study the case of different prior beliefs, often featuring identical preferences given the state. Banerjee and Somanathan (2001) and Kartik et al. (2015) study disclosure by multiple senders. In the first study, which features privately known priors, only experts with extreme priors disclose information, which on average has a moderating effect on R's actions. In the second study, the

[8]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).

[9]See in particular Dye (1985) and Shin (1994a,b, 2003) for the literature on disclosure. See Sobel (2013) for a general review of the literature on strategic information transmission.

authors identify cases where competition between senders promotes information revelation. Che and Kartik (2009) examines the effect of prior belief misalignment on S's incentives to acquire costly information. Prior misalignment hurts disclosure but increases S's effort, so that R may ultimately benefit from more misalignment. In one of our extensions, we consider a game in which an uninformed investor can decide how much to invest in a project managed by an exogenously informed entrepreneur. We find that moderate prior misalignment encourages disclosure by the entrepreneur.

In the above papers, S simply wants R's first-order beliefs to be close to some state dependent or independent bliss-point. In our paper, S effectively has preferences over second-order beliefs of R: She wants R to believe that her own first-order beliefs are close to those of S, i.e. cares about R's *perceived* disagreement. In consequence, S might for example want to conceal a signal that brings R's first-order beliefs closer to hers (i.e. reduces the actual disagreement) if this reduces perceived disagreement.

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, 2003) studies a sender-receiver game with an endogenous reputational concern of the sender for being perceived as unbiased, which leads to distorted communication. Loury (1994) offers a stimulating discussion of self-censorhip and political correctness in public discourse stemming from such concerns. In Gentzkow and Shapiro (2006), S wishes to signal a high quality of her information to R, who ultimately observes the actual state.[10] This leads S to bias her message towards R's prior belief. Similarly, in our setup if S's prior is more

[10]The models in Ottaviani and Sørensen (2006a) and Ottaviani and Sørensen (2006b) embed a similar mechanism resulting in S's reporting conforming to her own prior. Visser and Swank (2007) studies 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.

extreme than R's, S omits signals which contradict R's prior. The motivation is however very different: S wants to mitigate R's perception of ex-post disagreement (the quality of S's information being known). This same objective will as a matter of fact lead S to omit signals that confirm R's prior if R's prior is less extreme than S's.

Our paper 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). In contrast to ours, many applied models focus on preferences which depend on the interplay between beliefs and material payoffs, as in models of reciprocity (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004) or guilt aversion (Battigalli and Dufwenberg, 2007). Some models also study pure belief-based preferences: For instance, Ely et al. (2015) consider the behavior of a principal who wishes the beliefs of an agent to follow a specific time path exhibiting suspense or surprises.

Finally, our paper relates to a rich theoretical and empirical literature in social psychology 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). Our paper also relates to a current research agenda in political theory on deliberative and so-called epistemic democracy (see Estlund, 2009; Landemore and Elster, 2012; Sunstein, 2007, 2018; Mutz, 2006; Huckfeldt et al., 2004). The agenda evaluates democratic institutions and practices in terms of their truth-tracking properties, which intimately depend on citizens' incentive and ability to share their information with each other. The latter agenda itself relates closely to a strand of literature in political economy which studies information aggregation through voting and debate mechanisms (Austen-Smith and Banks, 1996; Feddersen and Pesendorfer, 1998; Coughlan, 2000; Austen-Smith and Feddersen, 2006; Mathis, 2011), tracing its origins to Condorcet's seminal work on majority voting.

All proofs, unless explicitly stated otherwise, are relegated to the Technical Appendix.

1 Main analysis

1.1 The disclosure game

There are two agents - the sender (S) and the receiver (R) and a state of Nature $\omega \in \{0,1\}$. Player $i \in S,R$ assigns prior probability $\beta_i \in (0,1)$ to $\omega = 1$. Let $\alpha_i = 1 - \beta_i$, for $i = S,R$. Priors are common knowledge. 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 $\sigma \neq \varnothing$. Player S can disclose the obtained signal to R or not. Denote S's disclosed information by d, where $d \in \{0,1,\varnothing\}$, where $\varnothing$ stands for no disclosure. R simply observes S's signal if disclosed and subsequently updates beliefs. Let $\widetilde{\beta}_i$ denote i's posterior probability assigned to $\omega = 1$ given obtained information. In particular, $\widetilde{\beta}_R(d)$ is the posterior probability assigned by R to state 1 given that S discloses d. Accordingly, $E_R[\widetilde{\beta}_S\,|d\,]$ is the expected value of S's posterior given d, in the eyes of R.

S is averse to perceived disagreement on the part of R, i.e. wants to minimize R's ex post perception of disagreement. S's utility function is given as follows:

$$U_S(E_R[\widetilde{\beta}_S\,|d\,],\widetilde{\beta}_R(d)) = -\left|E_R[\widetilde{\beta}_S\,|d\,] - \widetilde{\beta}_R(d)\right|. \tag{1}$$

In other words, S's utility is maximized if R *thinks* that S holds the same posterior belief as she. Note that S's *actual* posterior belief does not enter S's utility function. R's preferences are left unspecified, this player being entirely passive.[11]

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 disclosure

[11] Note that we could have assumed instead that S minimizes $E_R\,[\,|\widetilde{\alpha}_S - \widetilde{\alpha}_R|\,|\,d]$, in which case the sender would experience disutility even if R expects her belief to be the same as S's on average. The idea of our assumption is that S only cares about not being perceived as biased in a specific direction relative to R.

(called FD), disclosure of only 1-signals or only 0-signals (called D1 or D0). 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 hold a *stronger* or *more extreme* prior than j.

1.2 Equilibrium characterization

As our next proposition shows, S's optimal disclosure strategy depends on the relation of prior beliefs between the players, 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 $\beta_S^*(\beta_R, \frac{1}{2}) = 0$ and $\beta_S^{**}(\beta_R, \frac{1}{2}) = 1$.

Proposition 1 *1. If $\beta_S = \beta_R$, then the FD-, D0-, D1- and ND equilibria exist.*

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

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

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

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

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

e) There exists no ND-equilibrium.

Figure 1 below provides an illustration of our characterization for $\beta_R = .3$. The thick curves correspond to $\beta_S^*(.3, p)$ and $\beta_S^{**}(.3, p)$. Strictly between the two thick curves, 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.

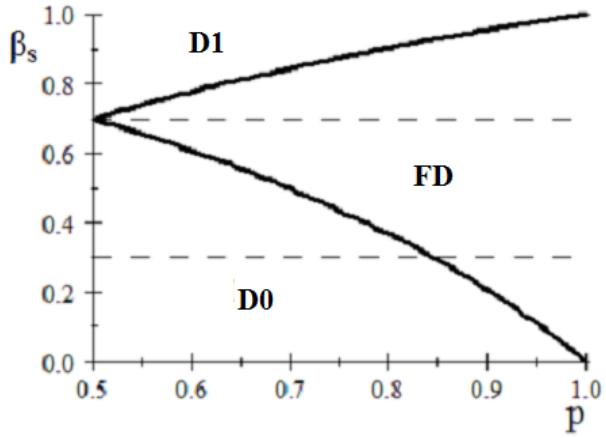

Figure 1: Equilibrium characterization.

Proposition 1 leads to the following corollary.

Corollary 1 *a) FD is the unique equilibrium if* $\beta_S = 1 - \beta_R$.

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

c) If equilibrium features partial disclosure, the signal that is disclosed is the one that is congruent with the bias of the player whose prior is the most extreme.

Summarizing, our characterization exhibits the following key properties:

1. Except under knife-edge conditions, our statement guarantees a unique equilibrium.

2. Unless $\alpha_S = \alpha_R$, there exists no ND-equilibrium. The reason is that for any p and $\alpha_S \neq \alpha_R$, the disclosure of at least one type of signal (either 0 or 1) leads to a strict decrease in disagreement w.r.t. prior disagreement. This follows from the fundamental statistical property that from an *ex-ante* perspective an informative signal

always reduces disagreement, by moving everyone's beliefs towards the truth in expectation (see Kartik et al. (2015) and Lemma II.A in Appendix II).

3. Full disclosure is not always feasible. The intuition comes from contemplating the fact that updating has two dimensions: The direction of belief updating and the intensity of belief updating. In our setup, players both update in the same direction after any given signal (no polarization), but they update with different intensities. An extreme player in particular heavily discounts a signal contradicting her prior bias (she considers it wrong with high probability) while a moderate player does not. In consequence, the difference in updating intensities for a given signal can be large enough to make posteriors more different than priors. The continuous curve in Figure 2 below shows the posterior probability attributed to state 0 given a 0-signal as a function of the prior β, for signal precision $p = .85$. For a given β, the intensity of belief updating corresponds to the distance between the diagonal line and the curve. The latter is plotted in the thick curve and is a hump shaped function of α. We see that extreme types update very little, while the maximum updating intensity arises for a prior moderately biased against the observed signal.

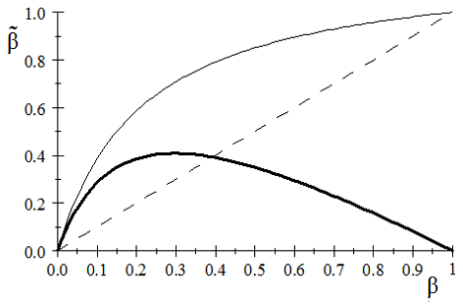

Figure 2: Intensity of belief updating given a 0-signal as a function of α.

4. Point a) of Corollary 1 implies that more prior misalignment, if not too extreme, can generate more disclosure, the disclosure-optimal sender prior being $1 - \beta_R$. The disclosure-optimal sender prior can thus be very different from R's prior but is exactly as extreme as R's. The technical intuition for the result is as follows. As noted

above, for any β_S, β_R, p at least one signal (either 0 or 1) leads to a decrease in actual disagreement w.r.t. the status quo. Next, note that if $\beta_S = 1 - \beta_R$, the effect of a 1-signal on disagreement is equivalent to the effect of a 0-signal, since priors are completely symmetric around $\frac{1}{2}$. I.e.

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

Since a disclosure of at least one type of signal must reduce disagreement, the other type of signal must achieve the same. Hence, full disclosure is achievable for any p for $\beta_S = 1 - \beta_R$. Note furthermore that updating prior β_S^* with a 0-signal or instead β_S^{**} with a 1-signal yields $1 - \beta_R$.

5. Point b) 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 policy.

6. Concerning Point c) of Corollary 1, consider the case where the two players have opposite prior biases and let the most extreme player be very extreme and the other player be very moderate (with prior 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 extremist's bias moves the belief of the moderate player closer to (away from) the extremist's prior.

   Within a simple random matching setup, Point c) of Corollary 1 naturally implies that the more R's prior is biased towards the wrong state, the less likely she is to be exposed to the truth. Assume that the true state is $\omega = 0$ and that R, whose prior β_R is publicly observed, faces a sender whose publicly observed prior is randomly drawn from the uniform distribution on $[0, 1]$. In such a setup, R is less likely to be exposed to a correct signal of 0, the higher β_R. Indeed, by Proposition 1 the ex-ante

probability of R being exposed to a 0-signal is

$$\Pr[\sigma = 1]\Pr[\beta_S < \beta_S^{**}(p)] = p\beta_S^{**}(p) = \frac{p^2(1-\beta_R)}{\beta_R + p(1-2\beta_R)},$$

which is strictly decreasing in β_R. Within a dynamic version of the above random matching scenario where R repeatedly plays the same one-shot disclosure game against short-sighted senders, perceived disagreement aversion on the part of senders thus slows down R's learning of the true state (i.e. causes inertia in beliefs) if the state is not congruent with R's extreme prior bias. Note that R's learning is only slowed down as opposed to entirely impeded, as R acknowledges that no disclosure by S does not necessarily imply that she holds no information.

1.3 The hidden cost of political correctness

Can S's attempt to minimize perceived disagreement be counter-productive from an ex ante perspective, thereby inducing what could be termed a hidden cost of political correctness? In what follows, 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 instead cares about actual disagreement (which might be a proxy for potentially costly social conflict or polarization).

First, from S's perspective, can the (ex ante) expected perceived disagreement be higher in a (partial disclosure) equilibrium than it would be under full disclosure? In such a case, S would prefer to commit to full disclosure. This question is answered in our next Proposition.

Proposition 2 *1. Let parameters be s.t. D1 is the unique equilibrium. Ex ante, S would strictly prefer to commit to full disclosure if $\beta_S > \beta_R$. If $\beta_S < \beta_R$, she instead ex ante strictly prefers the D1 equilibrium over FD.*

2. Let parameters be s.t. D0 is the unique equilibrium. Ex ante, S would strictly prefer to commit to full disclosure if $\beta_S < \beta_R$. If $\beta_S > \beta_R$, she instead ex ante strictly prefers the D0 equilibrium over FD.

Proof. See Appendix I. ■

S would thus ex ante prefer to commit to full disclosure if she is the most extreme player (which always holds in D1 (D0) if $\beta_S > \beta_R$ ($\beta_S < \beta_R$)). 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 she 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 extreme party. Recall that in this case, S omits signals contradicting her bias in a partial disclosure equilibrium (see Corollary 1.c). At the same time, R places a higher weight on the state corresponding to the omitted signal than does S, leading R to overweight (in S's eyes) the probability that such a signal is held (and omitted) by S, thereby inflating perceived disagreement after a non-disclosure. Instead, on the equilibrium path of the full disclosure equilibrium, R's prior does not affect her ex post perception of S's posterior (which then becomes common knowledge).

A second key question is whether from the perspective of a third party (TP) endowed with a prior $\widehat{\beta}$, the ex ante *actual* disagreement can be higher in equilibrium than it would be under FD. I.e., would TP prefer a truthful sender or a disagreement-averse sender if aiming at minimizing the expected 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 3 *Let parameters be s.t. 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 simul-

taneously:

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 it is sufficiently extreme.*

Part 1 of the proposition finds that the equilibrium concealment of information can indeed be counterproductive while Part 2 instead 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 always observes the original signal whatever the 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 is that 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. that 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 where β_S and β_R are on the same side of $\widehat{\beta}$, but S is further away and is sufficiently extreme (i.e., close to a boundary). An instance of this is the case of $\widehat{\beta} < \beta_R < \beta_S \approx 1$. Here, both players' beliefs decrease. At the same time, 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.

1.4 Strangers' talk

Conversations often take place between parties who do not exactly know each others' priors but who might hold some relevant information concerning each other's 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. In what follows, assume that priors are privately known.

Proposition 4 *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 (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. Under such distributional assumptions, 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, 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 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. For example, for $p = 0.7$, if β_S and β_R are both distributed according to a 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 changed to a truncated normal with mean $1/4$ and standard deviation $\sigma = 0.3$. 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 $\frac{1}{2}$. Figure 3 below provides examples of profiles of distributions of prior beliefs, complemented by a description (in bold) of the implied

equilibrium disclosure.

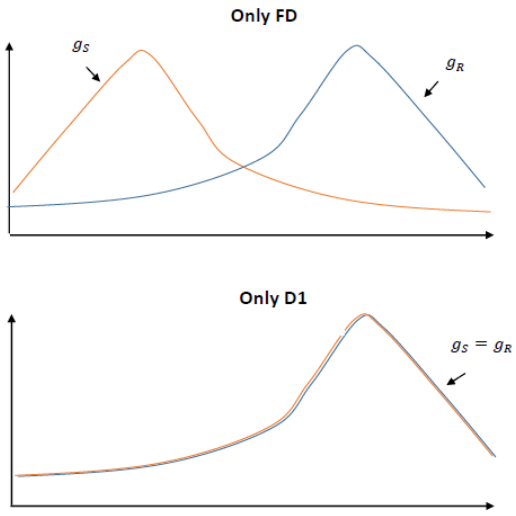

Figure 3: Distributions of possible priors.

1.5 Endogenous matching

We considered random matching in our discussion of Proposition 1 and found that extreme people were likely to be exposed only to confirmatory information over time. In reality, individuals often choose their conversation partners and we now explore this possibility within the context of our model. We find that voluntary matching further increases the prospect of echo chambers: Individuals select similar matching partners, which in turn induces partial and confirmatory disclosure.

Suppose a large population of senders and receivers, all being (perceived) disagreement averse. Senders and receivers are randomly matched and observe each others' priors. A match becomes *active* if and only if both players accept it. If the match does not become active, both players obtain a payoff of 0, which represents their outside option. If the match becomes active, the standard disclosure game introduced in section 1 ensues and each player's final payoff equals $B > 0$ minus the ex post perceived disagreement after the disclosure stage. The interpretation of the payoffs is that psychological payoffs

only arise once people explicitly decide to become involved in conversation. For simplicity, assume that acceptance decisions are made before S's information is realized. Note that if the disclosure subgame has multiple equilibria, then these all yield the same expected ex post perceived disagreement for any given player. Let $E_i[\Delta|\beta_i, \beta_j]$ denote i's expectation of j's ex post perceived disagreement given β_i, β_j. It follows that i will accept a match with j if and only if

$$B \geq E_i[\Delta|\beta_i, \beta_j]. \tag{2}$$

Proposition 5 *$E_i[\Delta|\beta_i, \beta_j]$ is continuous and V-shaped with respect to β_j, reaching its minimum of 0 at $\beta_i = \beta_j$.*

Figure 4 illustrates the above proposition. We assume $p = 0.9$, $\varphi = 0.6$, $\beta_S = 0.7$ and $B = 0.1$. The thick curve shows $E_S[\Delta|\beta_S, \beta_R]$ as a function of β_R. For $B = 0.1$ (represented by the horizontal green line), only values of β_R situated between the two vertical dotted lines satisfy (2). This is formalized in the following corollary.

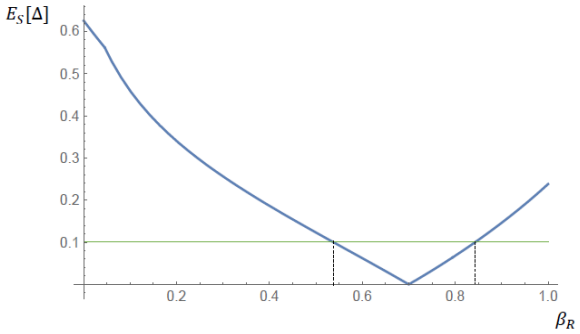

Figure 4: $E_S[\Delta|\beta_S, \beta_R]$ as a function of β_R

Corollary 2 *Given B, p, there are thresholds $\underline{\beta}_j(p, B, \beta_i) < \beta_i < \overline{\beta}_j(p, B, \beta_i)$ such that player i accepts a match with j if and only if $\beta_j \in I(p, B, \beta_i) = \left[\underline{\beta}_j(p, B, \beta_i), \overline{\beta}_j(p, B, \beta_i)\right]$. If $B < B'$, then $I(p, B, \beta_i) \subset I(p, B', \beta_i)$. Also, $\lim_{B \to 0} I(p, B, \beta_i) = \beta_i$.*

Proposition 1 and this corollary imply that the prospect of confirmatory information bias is strengthened with respect to the case of random matching, if we consider societies that are polarized in the sense that most people's prior is clearly away from $\frac{1}{2}$ and there

is significant mass on both sides of $\frac{1}{2}$. To see this, consider the following scenario. Fix p and let B be very small, and suppose that the population mass of prior values is equally shared between $\left(0, \underline{\beta}\right)$ and $\left(\overline{\beta}, 1\right)$, where $\underline{\beta} < \frac{1}{2} < \overline{\beta} = 1 - \underline{\beta}$. Consider a receiver i with $\beta_i > \overline{\beta}$ and and let us compare outcomes under respectively random and voluntary matching as described here. Assume finally that $\omega = 0$ so that a 0-signal contradicts i's prior bias.

Under random matching, with a probability bounded below by $\frac{1}{2}\left(\frac{1-\beta_i}{\underline{\beta}}\right)$, i (who satisfies $\beta_i > \overline{\beta}$) is matched in such a way that the implied equilibrium features either full disclosure or partial disclosure of the D0 type. As long as β_i is not extremely high, there is thus a significant probability that i encounters a contradicting 0-signal. The obtained probability bound is obtained as follows. Note first that i is matched half of the time with a sender satisfying $\beta_S \in (0, \underline{\beta})$. Second, given $\beta_S \in (0, \underline{\beta})$ the conditional probability that $\beta_S < 1 - \beta_i$ (yielding an FD or a D0 equilibrium) is $\frac{1-\beta_i}{\underline{\beta}}$.

Voluntary matching yields a very different picture. For B very low, a sender with prior β_S will accept a match with a receiver whose prior is β_R only if β_R is situated within a very close neighborhood of β_S. As a result, any active match in which the receiver i (with $\beta_i > \overline{\beta}$) is involved will involve a sender satisfying $\beta_S > \overline{\beta}$. Fixing p, for β_R large enough it furthermore holds true that $\beta_R \notin [\beta_S^*(\beta_R), \beta_S^{**}(\beta_R)]$. We may conclude that fixing p, for B very small and $\overline{\beta}$ large enough (see exact condition below), any active match involving a receiver i satisfying $\beta_i > \overline{\beta}$ will induce partial disclosure of the D1 type, implying that player i (who is biased towards state 1) only ever encounters confirmatory evidence no matter the state. Fixing p, the exact condition on $\overline{\beta}$ is easily shown to be $\overline{\beta} \geq \frac{1}{2p-1}\left(p - \sqrt{-p\,(p-1)}\right)$, which decreases as p decreases. For $p = \frac{3}{4}$, the requirement is $\overline{\beta} \geq 0.634$, which shows that weak societal polarization is already sufficient to create echo chamber dynamics under voluntary matching.

2 Extensions

In what follows, we consider a variety of key extensions and variations of our main setup. We first consider a more general information structure with continuous signals. The sec-

ond subsection finds that perceived disagreement aversion arises endogenously in a variety of simple strategic situations. Finally, we consider disagreement aversion within a game of costly collective acquisition of public signals.

2.1 Continuous signals

We now show that our characterization carries over qualitatively to the case of an information structure with continuous signals satisfying the marginal likelihood ratio property (MLRP). Assume that S's signal s is drawn from an interval $[\underline{s}, \bar{s}]$. 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. Assume furthermore that the extreme signal $\underline{s}$ ($\bar{s}$) is sufficiently small (large). 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 decreasing in s. Note that there exists a threshold signal $\tilde{s} \in (\underline{s}, \bar{s})$ such that whatever $\beta_i \in (0,1)$, it holds true that $\widetilde{\beta}_i(s) \lesseqgtr \beta_i$ 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 *continuous signals environment*. We call *simple disclosure equilibrium* (SDE) an equilibrium featuring two thresholds $\underline{s} < s_1 < s_2 < \bar{s}$ 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 *Assume the continuous signals environment:*

1. If $\beta_S \in \{\beta_R, 1-\beta_R\}$ then there exists FD. If $\beta_S \notin \{\beta_R, 1-\beta_R\}$, then the unique equilibrium is SDE.

2. All signals congruent with the bias of the player with the most extreme prior are disclosed.

3. When φ increases, the equilibrium becomes strictly more Blackwell informative.

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 most extreme player are fully revealed. Furthermore, if $\beta_S = 1 - \beta_R$, equilibrium features full disclosure, 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 carry over essentially identically to the continuous signals setup.

Proposition 7 *Assume the continuous signals environment:*

1. Let parameters be s.t. the equilibrium non-disclosure interval contains signals indicating state 0. S would strictly prefer to commit to full disclosure ex ante if $\beta_S > \beta_R$. If $\beta_S < \beta_R$, she ex ante strictly prefers any equilibrium over full disclosure.

2. Let parameters be s.t. the equilibrium non-disclosure interval contains signals indicating state 1. S would strictly prefer to commit to full disclosure ex ante if $\beta_S < \beta_R$. If $\beta_S > \beta_R$, she ex ante strictly prefers any equilibrium over full disclosure.

Proposition 8 *Assume the continuous signals environment. All the statements in Proposition 3 apply.*

2.2 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 a decision which is payoff-relevant to both S and R and which depends on players' first- and second-order beliefs. We consider four setups matching this description in what follows and find that in all of these, S is de facto averse to perceived disagreement at the initial disclosure stage and acts accordingly. The latter preference thus arises endogenously implying in turn the types of informational biases characterized in our main analysis. In all setups considered, the underlying environment is as in the main

section. Priors are commonly known, S is known to hold a binary signal of precision p with probability φ and the underlying state space is $\{0,1\}$.

2.2.1 Delegated decision making

An uninformed principal (R) faces a potentially informed agent (S), both being risk neutral. The principal faces a *problem* and there are two potential approaches for tackling it, named 0 and 1. One and only one of these actually can solve the problem, but it is a priori unknown which it is. We call the good approach (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. 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 $k = 1(0)$ indicates success (failure), $I(1) = 1$ and $I(0) = 0$. Conditional on S being hired and outcome k being secured, the payoff of R is thus $I(k)\tau - c$.

If S is hired, the contract proposed by R specifies a reward of 1 if the agent tackles the problem successfully (this outcome being observable). S has in total a unit of work time available and decides freely how much time to dedicate to each approach if hired. She incurs a cost $-\frac{1}{2}e_r^2$ of working e_r units of time on project $r \in \{1,2\}$. 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, 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 9 *There exists an FD-equilibrium if and only if there exists an FD-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 $d \in \{0, 1, \varnothing\}$. If she decides to hire, the principal expects to obtain the payoff of $\tau\Pi(\widetilde{\beta}_S(d), \widetilde{\beta}_R(d))$, where , given that posteriors are common knowledge after disclosure in a putative FD-equilibrium:

$$\Pi(\widetilde{\beta}_S(d), \widetilde{\beta}_R(d)) = \widetilde{\beta}_R(d)\widetilde{\beta}_S(d) + (1 - \widetilde{\beta}_R(d))(1 - \widetilde{\beta}_S(d)).$$

The principal thus hires if and only if c is smaller than the above (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 are reached: $\widetilde{\beta}_R(d) = \widetilde{\beta}_S(d) = 0$ or $\widetilde{\beta}_R(d) = \widetilde{\beta}_S(d) = 1$. In other words, S exhibits a form of disagreement aversion at the disclosure stage, in attempting to maximizing the probability of being hired.

We now examine the disclosure choice of the agent if she holds a signal $\sigma \in \{0, 1\}$. Let:

$$\Delta_\sigma(\beta_S, \beta_R) = \Pi(\widetilde{\beta}_R(\sigma), \widetilde{\beta}_S(\sigma)) - \Pi(\beta_R, \beta_S),\ \sigma \in \{0, 1\}.$$

Note that $\Delta_\sigma(\beta_S, \beta_R)\tau$ is thus the change in R's subjective expected payoff from hiring occasioned 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 $\Delta_\sigma(\beta_S, \beta_R) \geq 0$. In words, S discloses her signal only if the disclosure increases the probability that she is hired (and thereby obtains a positive utility). Now, it is easily shown that $\Delta_0(\beta_S, \beta_R)$ and $\Delta_1(p, \beta_S, \beta_R)$ 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.

2.2.2 Relationship-specific investment

An entrepreneur (S) seeks funding for a project from an uninformed investor (R). The game has four stages. In stage 1, S can disclose to R the signal available to her. In stage 2, R decides how much to invest in S's project, the cost of investing amount t being $\frac{1}{2}t^2$. In stage 3, S chooses an action $a \in \mathbb{R}$ pertaining to the project. In stage 4, payoffs are realized. The profit Π generated by the project given state of the world ω, and actions a_S, t is $t\gamma\left[1-(\omega - a_S)^2\right]$. A share λ of generated profits goes to R and the remainder goes to S. The utility function of R and S are thus respectively $\lambda t\gamma\left[1-(\omega-a_S)^2\right] - \frac{1}{2}t^2$ and $(1-\lambda)t\gamma\left[1-(\omega-a_S)^2\right]$.

At stage 4, given belief β_S, S's optimal action is trivially $a = \beta_S$. In the eyes of R, given her first order belief $\widetilde{\beta}_R$ and her second order beliefs about S's final beliefs $\widetilde{\beta}_S$, the expected profit $E_R\Pi$ generated by the investment equals $t\gamma$ times

$$\begin{aligned} & 1 - \left[E_R(1-\widetilde{\beta}_R)(\widetilde{\beta}_S)^2 + \widetilde{\beta}_R(1-\widetilde{\beta}_S)^2\right] \\ = \; & 1 - Var(\omega \,|\widetilde{\beta}_R) - E_R\left[\left(\widetilde{\beta}_R - \widetilde{\beta}_S\right)^2\right]. \end{aligned}$$

The expected profit of the project in the eyes of R is thus affected by two aspects: The uncertainty about ω (as captured by its variance) and R's perception of belief disagreement. It is also easily seen that the investor invests more, the larger her perception of the expected profit. The optimal amount of investment by R is pinned down by the simple

FOC

$$\frac{\partial\left(\lambda\gamma tE_R\left[1-(\omega-a_S)^2\right]-\frac{1}{2}t^2\right)}{\partial t}=\lambda\gamma E_R\left[1-(\omega-a_S)^2\right]-t=0,$$

which yields $t^*=\lambda\gamma E_R\left[1-(\omega-a_S)^2\right]$. The following Proposition identifies conditions under which a full disclosure (FD) equilibrium exists.

Proposition 10 *An FD equilibrium exists if and only if* $\beta_R\in\left[\beta_R^*\left(p,\beta_S\right),\beta_R^{**}\left(p,\beta_S\right)\right]$, *where* $\beta_R^*\left(p,\beta_S\right)$ *and* $\beta_R^{**}\left(p,\beta_S\right)$ *satisfy the following conditions:* $\beta_R^*\left(1,\beta_S\right)=0$, $\beta_R^{**}\left(1,\beta_S\right)=1$,

$$\beta_R^*\left(\frac{1}{2},\beta_S\right)=\beta_R^{**}\left(\frac{1}{2},\beta_S\right)\in\left(\min\left\{\frac{1}{2},1-\beta_S\right\},\max\left\{\frac{1}{2},1-\beta_S\right\}\right),$$

β_R^* and β_R^{**} are continuous in p. Finally, β_R^* (resp. β_R^{**}) is decreasing (resp. increasing) in p.

The exact formulas for $\beta_R^*\left(p,\beta_S\right)$ and $\beta_R^{**}\left(p,\beta_S\right)$ are omitted and available from the authors. The main feature of the above result is that for a given informed party prior β_S, the disclosure-optimal investor prior $\widehat{\beta}_R\left(\beta_S\right)$ is typically significantly different from β_S, where $\widehat{\beta}_R\left(\beta_S\right)$ is the investor prior for which FD is feasible under the lowest signal precision. This echoes the result obtained in Proposition 1. The following two figures provide a graphical illustration. In Figure 5, assuming $\beta_S=.3$, $\widehat{\beta}_R\left(.3\right)=0.537$ and FD is feasible only if β_R is located between the two continuous curves. In Figure 6, $\widehat{\beta}_R\left(\beta_S\right)$ is plotted (continuous curve) for every possible value of β_S.

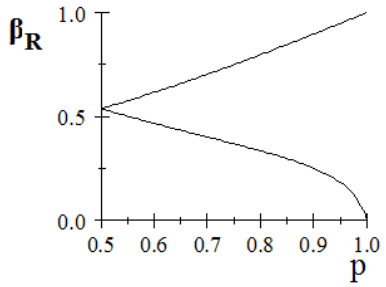

Figure 5: Investor (R) priors compatible with FD given p and $\beta_S=.3$.

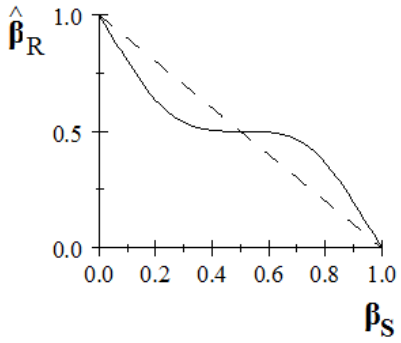

Figure 6: Disclosure optimal investor prior as function β_S.

The proof of the result is as follows. Note that the entrepreneur (S) trivially always benefits from higher investment by R. At the disclosure stage, the objective of S is thus to maximize $E_R\Pi$. Denote by $E_R[\Pi|d, FD]$ the expected profit conditional on disclosure d in a putative FD equilibrium. An FD equilibrium exists if and only if $E_R[\Pi|d=\emptyset] \leq E_R[\Pi|d=\sigma], \sigma \in \{0,1\}$, i.e.

$$-Var(\omega|\beta_R) - (\beta_R - \beta_S)^2 < -Var(\omega\left|\widetilde{\beta}_R(\sigma)\right.) - E_R\left(\widetilde{\beta}_R(\sigma) - \widetilde{\beta}_S(\sigma)\right)^2, \sigma \in \{0,1\}$$

which in turn rewrites as

$$Var(\omega\left|\widetilde{\beta}_R(\sigma)\right.) - Var(\omega|\beta_R) < (\beta_R - \beta_S)^2 - E_R\left(\widetilde{\beta}_R(\sigma) - \widetilde{\beta}_S(\sigma)\right)^2, \sigma \in \{0,1\}.$$

The LHS (RHS) expression is the change in the uncertainty level (perceived disagreement) implied by a disclosure of σ. S is thus willing to disclose a signal that generates uncertainty as long as it reduces perceived disagreement more than it increases uncertainty. S's aversion to perceived disagreement is the reason why some prior misalignment helps disclosure.

An important implication of the above is that there will be cases where the investor benefits from an increase in the belief misalignment of the entrepreneur. In such cases, though a more biased entrepreneur makes worse decisions conditional on given information, he provides the investor with strictly more information (FD instead of partial disclosure), which more than makes up for the first adverse effect.

2.2.3 Competing for authority

We here consider a game in which players compete for authority. Players' utility function is $-(\omega - a)^2 - \mu e_i$, where a is an action picked by the player allocated the final decision authority and e_i is the effort exerted by player i in a contest that determines the final allocation of authority among S and R. 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. Players simultaneously choose efforts and the winning probability of i given efforts levels e_i, e_j is $\frac{e_i}{e_i+e_j}$. In stage 3, S has a second opportunity to disclose her signal if she did not disclose it at stage 1. In stage 4, the winner of the contest picks an action $a \in \mathbb{R}$. 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:

$$\beta_S^{*,outcome}(\beta_R, p) = -\frac{-\beta_R + 2p^2\beta_R^2 + p\beta_R + \beta_R^2 - 3p\beta_R^2}{2p + \beta_R - 4p^2\beta_R^2 - 4p\beta_R - \beta_R^2 + 4p\beta_R^2 + 4p^2\beta_R - 2p^2},$$

$$\beta_S^{**,outcome}(\beta_R, p) = \frac{-2p^2\beta_R^2 + p\beta_R + p\beta_R^2}{2p + \beta_R - 4p^2\beta_R^2 - 4p\beta_R - \beta_R^2 + 4p\beta_R^2 + 4p^2\beta_R - 2p^2},$$

$$\beta_S^{*,disagreement}(\beta_R, p) = \frac{(1-p)(1-\beta_R)}{1-p+(\beta_R)(2p-1)},$$

$$\beta_S^{**,disagreement}(\beta_R, p) = \frac{p(1-\beta_R)}{(\beta_R)+p(1-2(\beta_R))},$$

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

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 7 below shows the intervals $I_{outcome}$ and $I_{disagreement}$, assuming $\beta_R = .3$. 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 moderately high (slightly larger than .75), full disclosure in stage 1 is achievable if and 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 her signal, if her concern is to minimize $-(\omega - a)^2$ under the assumption that R has authority. The interval $I_{disagreement}$ corresponds to the values of β_S for which S strictly favours disclosing no matter her signal, if her concern is to minimize perceived disagreement after disclosure. In a putative equilibrium with FD in stage 1, S will disclose in stage one only if disclosure leads to lower effort by R at the contest stage than not disclosing in stage 1. This, in turn, is true if and only if disclosing reduces perceived disagreement, which is equivalent to the requirement that β_S belongs to $I_{disagreement}(\beta_R, p)$. Recall that when it comes to affecting R's action if R is assigned authority, S can always revert a non-disclosure in stage 1 by disclosing in stage 3. It follows immediately from the above that an equilibrium with FD in stage 1 exists if β_S belongs to the intersection of $I_{disagreement}(\beta_R, p)$ and $I_{outcome}(\beta_R, p)$. Note finally that if β_S belongs to $I_{disagreement}(\beta_R, p)$ but not to $I_{disagreement}(\beta_R, p)$, 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 might help minimize the effort level of R at the contest stage, it does optimally inflect R's action choice.

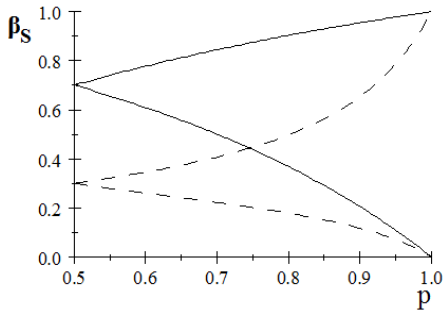

Figure 7: Disclosure boundaries.

2.2.4 Collective decision-making by compromise

Consider the following simple game of decision making by compromise. The game again consists of two stages, with stage 1 being the disclosure game studied in the main section. In stage 2 (policy stage), each agent submits a proposal $x_i \in \mathbb{R}$ (e.g. a draft of a law). The final policy x that is implemented is the compromise $x = \frac{1}{2}(x_S + x_R)$. There are two states 0 and 1. Let β_i denote the probability that i attaches to state 1 at the beginning of stage 2. Agent's i policy-related utility given final policy x and belief β_i is $-(\beta_i - x)^2$, so that i's ideal policy equals β_i. Given β_i, agent i has a cost of submitting an untruthful proposal $x_i \neq \beta_i$ described by the lying cost function $c(\beta_i, x) = \frac{1}{2}(\beta_i - x_i)^2$. A moderate party is thus for example intrinsically reluctant to submit an extreme proposal just to get its way in negotiations. We now show that S's payoff in equilibrium, at the beginning of the policy proposal stage, is decreasing in R's perception of disagreement in beliefs. The reason being that perceived disagreement encourages R (and as a consequence also S) to strategically distort her proposal, thereby wastefully inflating lying costs. S's problem in stage 2 is:

$$\min_{x_S} \left\{ \left(\beta_S - \frac{x_S + x_R}{2} \right)^2 + \frac{1}{2}(\beta_S - x_S)^2 \right\},$$

which implies $x_S = \frac{4\beta_S - x_R}{3}$. Similarly, R solves

$$\min_{x_R} \left\{ E_R \left[\left(\beta_R - \frac{x_S + x_R}{2} \right)^2 \right] + \frac{1}{2} (\beta_R - x_R)^2 \right\},$$

implying $x_R = \frac{4\beta_R - E_R[x_S]}{3}$. In equilibrium, we thus have

$$x_S = \frac{8\beta_S - 3\beta_R + E_R[\beta_S]}{6}, x_R = \frac{3\beta_R - E_R[\beta_S]}{2}.$$

Plugging the above quantities into S's payoff function, we may conclude that S obtains the following expected payoff in stage 2, given the profile of beliefs $\{\beta_S, \beta_R, E_R[\beta_S]\}$:

$$-\frac{3}{72} (2(\beta_S - \beta_R) - (E_R[\beta_S] - \beta_R))^2.$$

S's expected payoff at the beginning of stage 2 is thus negatively affected by R's perceived disagreement $(E_R[\beta_S] - \beta_R)$. Note that actual disagreement also enters the payoff function, so that S now not only wants to reduce perceived ex-post disagreement but is also averse to misleading R. One can use backward induction to solve for S's equilibrium disclosure choice in stage 1. In Figure 8 below, we show the resulting equilibrium outcomes assuming $\alpha_S = .55$. FD is feasible above the solid black curve. Below (strictly), only either D0 or D1 is feasible.

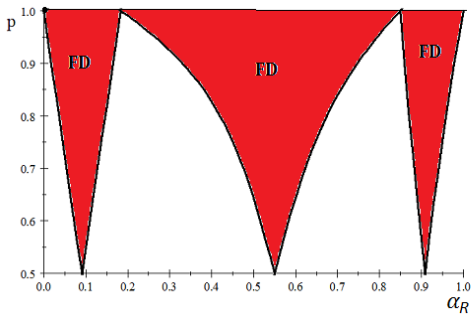

Figure 8: Partial equilibrium characterization.

It can be shown formally that for any α_R, p, the obtained characterization always exhibits the same qualitative features as in the present example. We call an equilibrium

featuring the full disclosure an FD-equilibrium. Define the following functions:

$$\begin{aligned}
\alpha_R^1(\alpha_S) &= \frac{1}{2} - \frac{1}{6}\sqrt{3}\sqrt{-4\alpha_S + 4(\alpha_S)^2 + 3}, \\
\alpha_R^2(\alpha_S) &= \frac{\alpha_S}{3}, \\
\alpha_R^3(\alpha_S) &= \frac{\alpha_S}{3} + \frac{2}{3}, \\
\alpha_R^4(\alpha_S) &= \frac{1}{6}\sqrt{3}\sqrt{-4\alpha_S + 4(\alpha_S)^2 + 3} + \frac{1}{2}.
\end{aligned}$$

It can be shown that for any $\alpha_S \in (0,1)$,

$$0 < \alpha_R^1(\alpha_S) < \alpha_R^2(\alpha_S) < \alpha_S < \alpha_R^3(\alpha_S) < \alpha_R^4(\alpha_S) < 1.$$

We may now state the following.

Proposition 12 *Fix $p \in (\frac{1}{2}, 1)$.*

a) For α_R sufficiently close to $\alpha_R^1(\alpha_S)$ or α_S or $\alpha_R^4(\alpha_S)$, there exists an FD-equilibrium.

b) For α_R sufficiently close to $\alpha_R^2(\alpha_S)$ or to $\alpha_R^3(\alpha_S)$, there exists no FD-equilibrium.

A proof of the statement is available upon request. The above Proposition establishes a sense in which increasing the difference in priors can be (locally) beneficial, thereby echoing our main characterization.

2.3 When R wants to learn the state

Our previous analysis has not assumed any explicit preferences of R. A natural assumption is that R wants to learn the state so as to pick an action that matches it as precisely as possible. In what follows, we accordingly assume that her utility function is $-(a-\omega)^2$, where $a \in [0,1]$ is the action chosen by R after S's disclosure. We first reexamine the simple disclosure game studied in Proposition 1 and characterize R's preference across possible sender priors. Second, we consider an extended version of the disclosure game in which R's prior is unknown while that of S is known, and R can make a cheap talk statement about her prior to S before the disclosure stage.

Proposition 13 *a) Given $\beta_R > \frac{1}{2}$, R strictly prefers D0-communication over D1-communication, and vice versa given $\beta_R < \frac{1}{2}$. b) Let $\beta_S \neq \beta_R$. R strictly prefers to face a sender with prior $\beta_S \in (\beta_S^*(\beta_R, p), \beta_S^{**}(\beta_R, p))$ than a sender with prior $\beta_S \notin (\beta_S^*(\beta_R, p), \beta_S^{**}(\beta_R, p))$. Given $\beta_R > \frac{1}{2}$, R strictly prefers to face a sender with prior $\beta_S \in (0, \beta_S^*(\beta_R, p)]$ than a sender with prior $[\beta_S^{**}(\beta_R, p), 1)$, and vice versa given $\beta_R < \frac{1}{2}$.*

Point a) shows that among two partially revealing experiments (D0 or D1), R prefers the one leading to disclosure of the signal that she considers ex ante the least likely. Point b) is a corollary of point a) and Proposition 1. For a fixed p and two senders with different priors, R will often strictly prefer to face the one with the most distant prior from her own. This will also be true if none of the two senders is compatible FD. If for example $\beta_R = .7$ and $p = .75$, then R prefers to face a sender with prior $\beta_S < \beta_S^*(.7, .75) = 0.437\,5$ to any sender with prior $\beta_S > \beta_S^{**}(.7, .75) = 0.875$.

While we established in our main section that uncertainty about prior bias is under some natural condition helpful, one might worry that it may be eliminated by communication about biases prior to disclosure. Players might be stuck in an equilibrium in which credible communication about priors gives rise to partial revelation at the disclosure stage. In what follows, assume that S's prior is known and equal to $\frac{1}{2}$ while R's prior is privately observed but drawn from a known distribution. The disclosure stage is preceded by a communication stage in which R sends a cheap talk message taken from the set $[0, 1]$, potentially providing information about her prior. We call the hereby constructed game the extended disclosure game. Note that given β_S and p, there are three sets of values of β_R (call these $\chi_{D0}, \chi_{D1}, \chi_{FD}$) for which the most informative equilibrium is respectively the D0, D1 and FD equilibrium. We call *essentially truthful* an equilibrium in which R truthfully reveals to which of these three sets β_R belongs by sending m_i with probability one whenever $\beta_R \in \chi_i$, for $i \in \{D0, D1, FD\}$, and the corresponding disclosure is subsequently used. Our next Proposition presents a negative result.

Proposition 14 *Consider the extended disclosure game. Suppose that at least two of the sets $\chi_{D0}, \chi_{D1}, \chi_{FD}$ have strictly positive probability mass. There exists no equilibrium featuring essentially truthful communication by R.*

The intuition for the above is as follows. First, if the distribution of R's prior contains a value such that full disclosure is incentive compatible for S, any R-type would trivially want to announce this prior value in a putative equilibrium featuring essentially truthful communication. Consider now the case where the set of possible priors only contains values that imply respectively D0 and D1 at the disclosure stage. As already noted, R's preference over partially revealing experiments reverses her prior bias. Note furthermore that given $\beta_S = \frac{1}{2}$, R is always more extreme than S so that partial disclosure after essentially truthful communication by R implies that S only discloses signals congruent with R's prior bias. R thus trivially has an incentive to deviate in stating her bias.

2.4 Joint observation of public signals

2.4.1 Basic setup and result

We here study the following simple game of voluntary and costly collective exposure to a public signal. 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 a randomly drawn public binary signal 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 ex-ante expected difference in beliefs con-

ditional on joint exposure to a signal of quality p is thus given by:

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

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 15 *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, $V^x(x,y,p)$ reaches its maximum for $y = y^* \in (0,1/2)$ if and only if $x \geq 1/2$.

Point 1 states that from the perspective of both players, an informative public signal reduces perceived disagreement in expectation (this is in line with Kartik et al. (2015)). Note that the marginal value of participating is trivially 0 if parties share the same prior, 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. Point 3 states that a player's optimal conversation partner (in the sense of maximizing the participation incentive) is always biased in the opposite direction.

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 symmetric (and non-radical) priors.

Proposition 16 *There is a unique pair $\{x^*, y^*\}$ maximizing the probability of signal acquisition. For this pair, it holds true that $y^* = 1 - x^*$ and $x^* \notin \{0, \frac{1}{2}, 1\}$.*

2.4.2 A dynamic matching game

Building on our basic exposure game, we provide a numerical analysis of a multi-period matching game. There are N agents, where N is very large. There are T periods, where T is sufficiently large. At $t = 0$, each agent's prior is randomly drawn from a uniform distribution on $[0, 1]$. In each period, agents are randomly matched in pairs. Period-t priors in each pair are observed. Agents have perfect recall of the history of signals that they have observed but do not observe other peoples' histories. Each participant decides whether to talk at fixed cost c per player. If (and only if) both members of a pair decide to talk, a signal of quality p is generated. Does such a simple learning process converge, and if so, to what distribution of beliefs?

We make two simplifying assumptions that embody forms of myopia. First, a player aims only at minimizing the perceived disagreement with the current (period-t) matching partner. Second, an agent, when observing the prior of the agent with whom she is matched at the beginning of period t, does not update her own prior on the basis of this other agent's prior. A fully rational player would instead do so: In a dynamic matching framework where agents' beliefs evolve over time as a function of the information to which they are exposed, the belief (i.e. the prior) of an agent at the beginning of period t contains information about the history of signals that this person has been exposed to over time.

Figure 9, below, provides the results from a simulation of the game with the following parameter values: $N = 10^6$, $T = 200$, $p = .7$, $c = 0.04$. We set $\omega = 1$. The process converges to the asymptotic distribution of posteriors appearing in the figure. The share of pairs talking to each other converges to 0, and beyond some period there is no further

change in the distribution of beliefs.

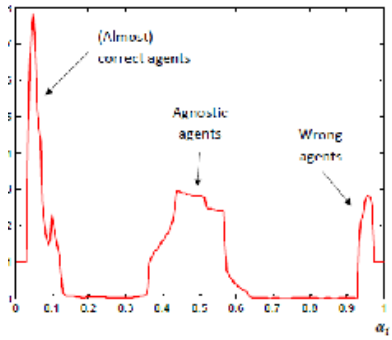

Figure 9: Asymptotic population distribution of beliefs.

In the asymptotic distribution, all agents are contained within three separate intervals featuring beliefs respectively close to 0, close to $\frac{1}{2}$ and close to 1. A first property is that virtually no one converges to the true belief of 0. Second, a large share of people are stuck with beliefs close to 1, i.e. to the wrong state. Finally, moderately biased types are to a large extent washed out. Society is thus arguably more polarized than at $t = 0$. The intuition for the stability of the asymptotic distribution is as follows. Nobody is willing to talk to extremists, who are too extreme to be convinced. And agnostic individuals do not want to talk to other agnostics. As a result, any pair formed by picking subjects from the three non-empty categories of agents is such that at least one agent is unwilling to talk. It follows that societal learning stops.

3 Conclusion

This paper introduces a new type of belief-dependent preferences reflecting an aversion to perceived disagreement. Our analysis has identified a range of important implications for key instances of information learning. A central finding is that larger differences in priors can imply better incentives for disclosure and joint information acquisition. Further work building on the assumption of disagreement-aversion might provide more insight into the causes and consequences of belief polarization in society.

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.

Austen-Smith, D. and J. S. Banks (1996). Information aggregation, rationality, and the Condorcet jury theorem. *American political science review 90*(1), 34–45.

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.

Crawford, V. P. and J. Sobel (1982). Strategic information transmission. *Econometrica: Journal of the Econometric Society 50*(6), 1431–1451.

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.

Dye, R. A. (1985). Disclosure of nonproprietary information. *Journal of accounting research 23*(1), 123–145.

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.

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.

Kartik, N., F. X. Lee, and W. Suen (2015). Does Competition Promote Disclosure?

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.

Levy, G. and R. Razin (2016). Correlation capacity, ambiguity and group shifts.

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

Mathis, J. (2011). Deliberation with evidence. *American Political Science Review 105*(3), 516–529.

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.

Shin, H. S. (2003). Disclosures and asset returns. *Econometrica 71*(1), 105–133.

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.

Sunstein, C. R. (2018). *# Republic: Divided democracy in the age of social media*. Princeton University Press.

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

4 Technical Appendix

4.1 Appendix I: Preliminaries

Throughout the proofs we use the following notation for the perceived disagreement under equilibrium of type $X = \{D0, D1, FD\}$ given the disclosed information $d = \{0, 1, \varnothing\}$:

$$\Delta^X(d) = \left| E_R[\widetilde{\beta}_S|d] - \widetilde{\beta}_R(d) \right|.$$

Besides, it is convenient to denote the highest and the lowest prior belief as, respectively

$$\begin{aligned} x &= \max\{\beta_S, \beta_R\}, \\ y &= \min\{\beta_S, \beta_R\}. \end{aligned}$$

In what follows, we use the following posterior beliefs, obtained by applying Bayes' rule.

In an FD equilibrium:

$$\begin{aligned} \widetilde{\beta}_i(1) &= \frac{\Pr[\sigma = 1|\omega = 1]\beta_i}{\Pr[\sigma = 1|\omega = 1]\beta_i + \Pr[\sigma = 1|\omega = 0](1-\beta_i)} = \frac{p\beta_i}{p\beta_i + (1-p)(1-\beta_i)}, \\ \widetilde{\beta}_i(0) &= \frac{\Pr[\sigma = 0|\omega = 1]\beta_i}{\Pr[\sigma = 0|\omega = 1]\beta_i + \Pr[\sigma = 0|\omega = 0](1-\beta_i)} = \frac{(1-p)\beta_i}{(1-p)\beta_i + p(1-\beta_i)}, \end{aligned}$$

In a D1 equilibrium:

$$\begin{aligned}
\widetilde{\beta}_R^{D1}(\varnothing) &= \Pr[\sigma=0|d=\varnothing,D1]\widetilde{\beta}_R(0)+\Pr[\sigma=\varnothing|d=\varnothing,D1]\beta_R \\
&= \frac{\Pr[\sigma=0]}{\Pr[\sigma=0]+\Pr[\sigma=\varnothing]}\widetilde{\beta}_R(0)+\frac{\Pr[\sigma=\varnothing]}{\Pr[\sigma=0]+\Pr[\sigma=\varnothing]}\beta_R \\
&= \frac{\varphi((1-p)\beta_R+p(1-\beta_R))}{\varphi((1-p)\beta_R+p(1-\beta_R))+(1-\varphi)}\widetilde{\beta}_R(0) \\
&\quad +\frac{1-\varphi}{\varphi((1-p)\beta_R+p(1-\beta_R))+(1-\varphi)}\beta_R, \\
E_R^{D1}[\widetilde{\beta}_S|\varnothing] &= \Pr[\sigma=0|d=\varnothing,D1]\widetilde{\beta}_S(0)+\Pr[\sigma=\varnothing|d=\varnothing,D1]\beta_S \\
&= \frac{\varphi((1-p)\beta_R+p(1-\beta_R))}{\varphi((1-p)\beta_R+p(1-\beta_R))+(1-\varphi)}\widetilde{\beta}_S(0) \\
&\quad +\frac{1-\varphi}{\varphi((1-p)\beta_R+p(1-\beta_R))+(1-\varphi)}\beta_S.
\end{aligned}$$

In a D0 equilibrium:

$$\begin{aligned}
\widetilde{\beta}_R^{D0}(\varnothing) &= \Pr[\sigma=1|d=\varnothing,D0]\widetilde{\beta}_R(1)+\Pr[\sigma=\varnothing|d=\varnothing,D0]\beta_R \\
&= \frac{\Pr[\sigma=1]}{\Pr[\sigma=1]+\Pr[\sigma=\varnothing]}\widetilde{\beta}_R(1)+\frac{\Pr[\sigma=\varnothing]}{\Pr[\sigma=1]+\Pr[\sigma=\varnothing]}\beta_R \\
&= \frac{\varphi(p\beta_R+(1-p)(1-\beta_R))}{\varphi(p\beta_R+(1-p)(1-\beta_R))+(1-\varphi)}\widetilde{\beta}_R(1) \\
&\quad +\frac{1-\varphi}{\varphi(p\beta_R+(1-p)(1-\beta_R))+(1-\varphi)}\beta_R, \\
E_R^{D0}[\widetilde{\beta}_S|\varnothing] &= \Pr[\sigma=1|d=\varnothing,D0]\widetilde{\beta}_S(1)+\Pr[\sigma=\varnothing|d=\varnothing,D0]\beta_S \\
&= \frac{\varphi(p\beta_R+(1-p)(1-\beta_R))}{\varphi(p\beta_R+(1-p)(1-\beta_R))+(1-\varphi)}\widetilde{\beta}_S(1) \\
&\quad +\frac{1-\varphi}{\varphi(p\beta_R+(1-p)(1-\beta_R))+(1-\varphi)}\beta_S.
\end{aligned}$$

4.2 Appendix II: Disclosure with binary signals (Proof of Proposition 1 and Corollary 1)

Proposition 1 follows from a set of Lemmas, which are stated and proved in what follows. The first lemma is an adaptation of Theorem 1 in Kartik et al. (2015).

Lemma II.A *Assume $\beta_S \neq \beta_R$. Then, under full disclosure, the expected disagreement is strictly reduced relative to the prior disagreement from S's ex-ante perspective.*

Proof. Assume without loss of generality that $\beta_S > \beta_R$. Then, the difference between the prior and the expected values of disagreement under full disclosure is

$$\begin{aligned}
(\beta_S - \beta_R) - E_S[\Delta^{FD}] &= (\beta_S - \beta_R) \\
&\quad -\varphi(\beta_S p + (1-\beta_S)(1-p))\Delta(1) - \varphi(\beta_S(1-p) + (1-\beta_S)p)\Delta(0) \\
&\quad -(1-\varphi)(\beta_S - \beta_R) \\
&= (\beta_S - \beta_R) - \varphi(\beta_S p + (1-\beta_S)(1-p)) \\
&\quad \times \left(\frac{\beta_S p}{\beta_S p + (1-\beta_S)(1-p)} - \frac{\beta_R p}{\beta_R p + (1-\beta_R)(1-p)} \right) \\
&\quad -\varphi(\beta_S(1-p) + (1-\beta_S)p) \\
&\quad \times \left(\frac{\beta_S(1-p)}{\beta_S(1-p) + (1-\beta_S)p} - \frac{\beta_R(1-p)}{\beta_R(1-p) + (1-\beta_R)p} \right) \\
&\quad -(1-\varphi)(\beta_S - \beta_R) \\
&= \varphi \frac{(\beta_S - \beta_R)(1-\beta_R)\beta_R(2p-1)^2}{(1-p+\beta_R(2p-1))(\beta_R + p(1-2\beta_R))} > 0.
\end{aligned}$$

Hence, S expects (before obtaining the signal) that the full disclosure strategy will reduce the perceived disagreement relative to prior disagreement. ■

Lemma II.B *There exists no equilibrium in which S always omits to disclose the signal unless $\beta_S = \beta_R$.*

Proof.

Step 1. Consider $\beta_S \neq \beta_R$. Assume by contradiction that there exists an equilibrium with no disclosure. Hence, S should prefer no disclosure over disclosure conditional on both signals. Since in this equilibrium we have $\Delta(d) = |\widetilde{\beta}_S(d) - \widetilde{\beta}_R(d)|$ for $d \in \{0,1\}$, and $\Delta(\varnothing) = |\beta_S - \beta_R|$, the S's incentive constraints are

$$\begin{aligned}
|\widetilde{\beta}_S(0) - \widetilde{\beta}_R(0)| &\geq |\beta_S - \beta_R|, \\
|\widetilde{\beta}_S(1) - \widetilde{\beta}_R(1)| &\geq |\beta_S - \beta_R|.
\end{aligned}$$

This implies that the expected signal, if always disclosed, does not reduce disagreement from the ex-ante perspective, which contradicts Lemma I.A.

Step 2. If $\beta_S = \beta_R$, then clearly $|\widetilde{\beta}_S(0) - \widetilde{\beta}_R(0)| = |\widetilde{\beta}_S(1) - \widetilde{\beta}_R(1)| = \left|\beta_S - \beta_R\right| = 0$.

Hence, the above incentive constraints are trivially satisfied and the equilibrium with no disclosure exists.■

Lemma II.C. *If $\beta_S = \beta_R$, then all pure-strategy equilibria exist.*

Proof. The existence of no disclosure equilibrium follows from Lemma I.B. The existence of FD follows from the fact that if $\beta_S = \beta_R$, then $\Delta(d) = 0$ for $d \in \{0,1\}$, hence the sender never has a strict incentive not to disclose the obtained signal.

Finally, let us prove the existence of equilibria where only $\sigma = \eta$, $\eta \in \{0,1\}$ is disclosed. Since in such equilibrium no disclosure signals to R either $\sigma = \eta$ or that S is truly uninformed, we have (denoting $\delta = \Pr[\sigma = \eta | d = \varnothing]$)

$$\begin{aligned}
\Delta(\varnothing) &= \left|E_R[\widetilde{\beta}_S|\varnothing] - \widetilde{\beta}_R(\varnothing)\right| \\
&= \left|\delta\widetilde{\beta}_S(\eta) + (1-\delta)\beta_S - \delta\widetilde{\beta}_R(\eta) - (1-\delta)\beta_R\right| \\
&= \left|\delta(\widetilde{\beta}_S(\eta) - \widetilde{\beta}_R(\eta)) + (1-\delta)(\beta_S - \beta_R)\right| = 0, \qquad (3)
\end{aligned}$$

where the last equality is due to $\beta_S = \beta_R$. Hence, S is indifferent between no disclosure and disclosure of any signal (which in turn also yields a null disagreement), which confirms the claim.■

Lemma II.D. *If $\beta_S \neq \beta_R$, then D0 exists if and only if $\beta_S \leq \beta_S^*(\beta_R)$.*

Proof. In D0 equilibrium the following S's incentive constraints should be satisfied:

$$\Delta^{D0}(0) \leq \Delta^{D0}(\varnothing) \leq \Delta^{D0}(1). \qquad (4)$$

Using (3), the second incentive constraint simplifies to

$$\begin{aligned}
\Delta^{D0}(\varnothing) - \Delta^{D0}(1) &\leq 0, && (5)\\
\left|\delta(\widetilde{\beta}_S(1) - \widetilde{\beta}_R(1)) + (1-\delta)(\beta_S - \beta_R)\right| - \left|\widetilde{\beta}_S(1) - \widetilde{\beta}_R(1)\right| &\leq 0, && (6)\\
(1-\delta)\left(x - y - \frac{xp}{xp + (1-x)(1-p)} + \frac{yp}{yp + (1-y)(1-p)}\right) &\leq 0, && (7)\\
\left(1 - \frac{\varphi(yp + (1-y)(1-p))}{\varphi(yp + (1-y)(1-p)) + (1-\varphi)}\right)&\\
\times\left(x - y - \frac{xp}{xp + (1-x)(1-p)} + \frac{yp}{yp + (1-y)(1-p)}\right) &\leq 0, && (8)\\
(x-y)(2p-1)(1-\varphi)&\\
\times\frac{x(1-p+y(2p-1)) - (1-p)(1-y)}{(1-p+x(2p-1))(1-p+y(2p-1))(1-\varphi(p-y(2p-1))} &\leq 0. && (9)
\end{aligned}$$

On the left-hand side of the last inequality, all terms are always positive except for the nominator, which is increasing in both x and y and is equal to 0 if and only if

$$x = \frac{(1-y)(1-p)}{1-p+y(2p-1)} \Leftrightarrow y = \frac{(1-x)(1-p)}{1-p+x(2p-1)}.$$

Thus, independently of whether $\beta_S = x$ or $\beta_S = y$ (i.e., of whether $\beta_S > \beta_R$ or $\beta_S < \beta_R$) we have

$$\Delta^{D0}(\varnothing) - \Delta^{D0}(1) \leq 0 \text{ if and only if } \beta_S \leq \frac{(1-\beta_R)(1-p)}{1-p+\beta_R(2p-1)} = \beta_S^*(\beta_R). \qquad (10)$$

Note further that $\Delta^{D0}(\varnothing) - \Delta^{D0}(1) \leq 0$ immediately implies $\Delta^{D0}(0) \leq \Delta^{D0}(\varnothing)$. Indeed, otherwise we would have $\Delta^{D0}(\varnothing) \leq \min\{\Delta^{D0}(1), \Delta^{D0}(0)\}$, i.e. for any $\eta \in \{0,1\}$

$$\begin{aligned}
\delta|\widetilde{\beta}_S(\eta) - \widetilde{\beta}_R(\eta)| + (1-\delta)|\beta_S - \beta_R| &\leq \left|\widetilde{\beta}_S(\eta) - \widetilde{\beta}_R(\eta)\right|,\\
|\beta_S - \beta_R| &\leq |\widetilde{\beta}_S(\eta) - \widetilde{\beta}_R(\eta)|
\end{aligned}$$

for any $\eta \in \{0,1\}$, which would imply that (a putative) full disclosure would not expectedly reduce disagreement which is a contradiction by Lemma I.A.

Thus, (4) holds if and only if $\beta_S \leq \beta_S^*(\beta_R)$.■

Lemma II.E. *If $\beta_S \neq \beta_R$, then D1 exists if and only if $\beta_S \geq \beta_S^{**}(\beta_R)$.*

Proof. In D1 equilibrium the following S's incentive constraints should be satisfied:

$$\Delta^{D1}(1) \leq \Delta^{D1}(\varnothing) \leq \Delta^{D1}(0). \tag{11}$$

Using (3), the second incentive constraint simplifies to

$$\begin{aligned}
\Delta^{D1}(\varnothing) - \Delta^{D1}(0) &\leq 0, && (12)\\
\left|\delta(\widetilde{\beta}_S(0) - \widetilde{\beta}_R(0)) + (1-\delta)(\beta_S - \beta_R)\right| - \left|\widetilde{\beta}_S(0) - \widetilde{\beta}_R(0)\right| &\leq 0, && (13)\\
(1-\delta)\left(x - y - \frac{x(1-p)}{x(1-p)+(1-x)p} + \frac{y(1-p)}{y(1-p)+(1-y)p}\right) &\leq 0, && (14)\\
\left(1 - \frac{\varphi(y(1-p)+(1-y)p)}{\varphi(y(1-p)+(1-y)p)+(1-\varphi)}\right)\\
\times\left(x - y - \frac{x(1-p)}{x(1-p)+(1-x)p} + \frac{y(1-p)}{y(1-p)+(1-y)p}\right) &\leq 0, && (15)\\
(x-y)(2p-1)(1-\varphi)\\
\times\frac{x(y(2p-1)-p)+p(1-y)}{(p-x(2p-1))(p-y(2p-1))(1-\varphi(1-p+y(2p-1)))} &\leq 0. && (16)
\end{aligned}$$

On the left-hand side of the last inequality, all terms are always positive except for the nominator, which is decreasing in both x and y and is equal to 0 if and only if

$$x = \frac{p(1-y)}{y+p(1-2y)} \Leftrightarrow y = \frac{p(1-x)}{x+p(1-2x)}.$$

Thus, independently of whether $\beta_S = x$ or $\beta_S = y$ (i.e., of whether $\beta_S > \beta_R$ or $\beta_S < \beta_R$) we have

$$\Delta^{D1}(\varnothing) - \Delta^{D1}(0) \leq 0 \text{ if and only if } \beta_S \geq \frac{p(1-\beta_R)}{\beta_R + p(1-2\beta_R)} = \beta_S^{**}(\beta_R). \tag{17}$$

Note further that $\Delta^{D1}(\varnothing) - \Delta^{D1}(0) \leq 0$ immediately implies $\Delta^{D1}(1) \leq \Delta^{D1}(\varnothing)$ by the same argument as in the proof of Lemma I.D. Thus, (11) holds if and only if $\beta_S \geq \beta_S^{**}(\beta_R)$.■

Lemma II.F *FD exists if and only if* $\beta_S \in [\beta_S^*(\beta_R), \beta_S^{**}(\beta_R)]$.

Proof. In FD equilibrium the following S's incentive constraints should be satisfied:

$$|\beta_S - \beta_R| \geq \Delta^{FD}(0), \tag{18}$$
$$|\beta_S - \beta_R| \geq \Delta^{FD}(1). \tag{19}$$

Note that the reverse inequality to (18) holds under the same conditions as (7), which is turn is equivalent to (5). Hence, by the proof of Lemma I.D $|\beta_S - \beta_R| \leq \Delta^{FD}(0)$ iff $\beta_S \leq \beta_S^*(\beta_R)$ (with $|\beta_S - \beta_R| = \Delta^{FD}(0)$ iff $\beta_S = \beta_S^*(\beta_R)$). Consequently, (18) holds if and only if $\beta_S \geq \beta_S^*(\beta_R)$. Analogously, from the proof of Lemma I.E we obtain that (19) holds if and only if $\beta_S \leq \beta_S^{**}(\beta_R)$. Hence, both constraints hold simultaneously if and only if $\beta_S \in [\beta_S^*(\beta_R), \beta_S^{**}(\beta_R)]$.■

Lemma II.G *If $\beta_S \neq \beta_R$, mixed strategy equilibria exist if and only if $\beta_S \in \{\beta_S^*(\beta_R), \beta_S^{**}(\beta_R)\}$.*

Proof. First, if $\beta_S \neq \beta_R$, there cannot be an equilibrium in which S omits to disclose with positive probability after both signals, since the actual (and hence perceived) disagreement should strictly decline at least after one signal by Lemma I.A. Consider the remaining case when mixing between disclosure and non-disclosure occurs only for one signal $\sigma^* \in \{0,1\}$. For the case of $\sigma^* = 1$, such an equilibrium requires the indifference condition $\Delta(\varnothing) - \Delta(1) = 0$. This is further equivalent to (where δ again denotes the probability of S having obtained the signal conditional on no disclosure):

$$\begin{aligned}
\left|\delta(\widetilde{\beta}_S(1) - \widetilde{\beta}_R(1)) + (1-\delta)(\beta_S - \beta_R)\right| - \left|\widetilde{\beta}_S(1) - \widetilde{\beta}_R(1)\right| &= 0, \\
(1-\delta)\left(x - y - \frac{xp}{xp + (1-x)(1-p)} + \frac{yp}{yp + (1-y)(1-p)}\right) &= 0, \qquad (20) \\
x - y - \frac{xp}{xp + (1-x)(1-p)} + \frac{yp}{yp + (1-y)(1-p)} &= 0, \\
x = \left\{y, \frac{(1-y)(1-p)}{1-p+y(2p-1)}\right\} & \\
\Leftrightarrow y = \left\{x, \frac{(1-x)(1-p)}{1-p+x(2p-1)}\right\}. & \qquad (21)
\end{aligned}$$

Thus, if $\beta_S \neq \beta_R$, then S is indifferent between sending 1 and no disclosure if and only if $\beta_S = \beta_S^*(\beta_R)$. The other equilibrium incentive constraint $\Delta(\varnothing) \geq \Delta(0)$ is then necessarily satisfied by Lemma I.A.

Analogously, for the case of $\sigma^* = 0$, the mixed-strategy equilibrium requires the indifference condition $\Delta(\varnothing) - \Delta(0) = 0$. This is further equivalent to

$$\begin{aligned}
\left|\delta(\widetilde{\beta}_S(0) - \widetilde{\beta}_R(0)) + (1-\delta)(\beta_S - \beta_R)\right| - \left|\widetilde{\beta}_S(0) - \widetilde{\beta}_R(0)\right| &= 0, \\
(1-\delta)\left(x - y - \frac{x(1-p)}{x(1-p)+(1-x)p} + \frac{y(1-p)}{y(1-p)+(1-y)p}\right) &= 0, \qquad (22) \\
\left(x - y - \frac{x(1-p)}{x(1-p)+(1-x)p} + \frac{y(1-p)}{y(1-p)+(1-y)p}\right) &= 0, \\
x = \left\{y, \frac{p(1-y)}{y+p(1-2y)}\right\} & \\
\Leftrightarrow y = \left\{x, \frac{p(1-x)}{x+p(1-2y)}\right\}. & \qquad (23)
\end{aligned}$$

This similarly leads to $\beta_S = \beta_S^{**}(\beta_R)$. ■

Proof of Corollary 1. Point a) follows from the fact that $\beta_S^*(\beta_R, p) < 1 - \beta_R < \beta_S^{**}(\beta_R, p)$, i.e. $\beta_S = 1 - \beta_R$ always satisfies the condition for the uniqueness of FD equilibrium according to Proposition 1. Point b) follows due to $\beta_S^*(\beta_R, p)$ ($\beta_S^{**}(\beta_R, p)$) being continuously decreasing (increasing) in p, being equal to 0 (1) if $p = 1$. Point c) follows again from the fact that $\beta_S^*(\beta_R, p) < 1 - \beta_R < \beta_S^{**}(\beta_R, p)$. Then, if R is biased towards 1 ($\beta_R < 1/2$), then by Proposition 1 the opposite 1-signal is disclosed only if $\beta_S \geq \beta_S^{**}(\beta_R) > 1 - \beta_R$, i.e., S is stronger biased towards 0. Analogously, if R is biased towards 1 ($\beta_R < 1/2$), then by Proposition 1 the opposite 0-signal is disclosed only if $\beta_S \leq \beta_S^*(\beta_R) < 1 - \beta_R$, i.e., S is stronger biased towards 1.

4.3 Proof of Proposition 5

Proof:

Step 1. First, note that $\Delta(0, \beta_S, \beta_R)$ and $\Delta(1, \beta_S, \beta_R)$ are V-shaped with respect to either β_S or β_R reaching its minimum at $\beta_S = \beta_R$. Indeed, since $\widetilde{\beta}_i(0)$ is increasing in β_i, it follows that $\Delta(0, \beta_i, \beta_j)$ decreases in β_i if $\beta_i < \beta_j$ and increases in β_i otherwise, being equal to 0 for $\beta_i = \beta_j$. The same argument applies to $\Delta(1, \beta_S, \beta_R)$.

Step 2. Let us show another auxiliary result that $E_S[\Delta^{D0}(\varnothing, \beta_S, \beta_R)]$ and $E_S[\Delta^{D1}(\varnothing, \beta_S, \beta_R)]$ are V-shaped with respect to β_R reaching its minimum at $\beta_S = \beta_R$. Consider $E_S[\Delta^{D1}(\varnothing, \beta_S, \beta_R)]$. Using the expressions from Appendix I, we have:

$$E_S[\Delta^{D1}(\varnothing, \beta_S, \beta_R)] = \frac{p(1-p\varphi) - \beta_S(2p-1)(1-\varphi)}{p - \beta_S(2p-1)} \frac{|\beta_S - \beta_R|}{1 - \varphi(1 - p + \beta_R(2p-1))}. \quad (24)$$

Taking the derivative with respect to β_R and simplifying we obtain (for $\beta_R \neq \beta_S$)

$$\frac{\partial E_S[\Delta^{D1}(\varnothing, \beta_S, \beta_R)]}{\partial \beta_R} = \text{sgn}[\beta_R - \beta_S] \frac{p(1-p\varphi) - \beta_S(2p-1)(1-\varphi)}{p - \beta_S(2p-1)} \frac{1 - \varphi(1 - p + \beta_S(2p-1))}{(1 - \varphi(1 - p + \beta_R(2p-1)))^2}$$

It is easy to verify that all terms on the right-hand side following the sign function are always positive. Hence, the sign of the derivative is determined by $\text{sgn}[\beta_R - \beta_S]$, which implies that function $E_S[\Delta^{D1}(\varnothing, \beta_S, \beta_R)]$ is V-shaped with respect to β_R, being kinked at $\beta_S = \beta_R$ where it is equal to 0 (see 24).

Consider $E_S[\Delta^{D0}(\varnothing, \beta_S, \beta_R)]$. Using the expressions from Appendix I, we have:

$$E_S[\Delta^{D0}(\varnothing, \beta_S, \beta_R)] = \frac{1 - p + \beta_S(2p-1)(1-\varphi) - (1-p)^2\varphi}{1 - p + \beta_S(2p-1)} \frac{|\beta_S - \beta_R|}{1 - \varphi(p - \beta_R(2p-1))}. \quad (25)$$

Taking the derivative with respect to β_R and simplifying we obtain (for $\beta_R \neq \beta_S$)

$$\frac{\partial E_S[\Delta^{D0}(\varnothing, \beta_S, \beta_R)]}{\partial \beta_R} = \text{sgn}[\beta_R - \beta_S] \frac{1 - p + \beta_S(2p-1)(1-\varphi) - (1-p)^2\varphi}{1 - p + \beta_S(2p-1)} \frac{1 - \varphi(p - \beta_S(2p-1))}{(1 - \varphi(p - \beta_R(2p-1)))^2}$$

It is easy to verify that all terms on the right-hand side following the sign function are always positive. Hence, the sign of the derivative is determined by $\text{sgn}[\beta_R - \beta_S]$, which again implies that function $E_S[\Delta^{D0}(\varnothing, \beta_S, \beta_R)]$ is V-shaped with respect to β_R, being kinked at $\beta_S = \beta_R$ where it is equal to 0 (see 25).

Step 3. Consider the ex-ante perceived disagreement from R's perspective. Note that for R the perceived disagreement is still

$$-\left|E_R[\widetilde{\beta}_S \,|d] - E_R E_S[\widetilde{\beta}_R(d)]\right| = -\left|E_R[\widetilde{\beta}_S \,|d] - \widetilde{\beta}_R(d)\right|$$

as in S's case, given that $\widetilde{\beta}_R(d)$ is common knowledge. Next note that from the ex-ante perspective conditioning on a future signal does not change the expectation, i.e.

$E_R[E_R[\widetilde{\beta}_S\,|d]] = E_R[\widetilde{\beta}_S]$ and $E_R[\widetilde{\beta}_R(d)] = E_R[\widetilde{\beta}_R]$. This implies, that the ex-ante perceived disagreement from R's perspective does not depend on the disclosure rule. Hence, it is sufficient to show that $E_i[\Delta|\beta_i, \beta_j]$ is V-shaped with respect to β_j just for FD (since the value of the expected disagreement is the same under any other disclosure rule). We have

$$E_R[\Delta^{FD}] = \Pr[\sigma_S = 1|\beta_R]\Delta(1, \beta_S, \beta_R) + \Pr[\sigma_S = 0|\beta_R]\Delta(0, \beta_S, \beta_R) + \Pr[\sigma_S = \varnothing|\beta_R]|\beta_S - \beta_R|.$$

Given that all terms are V-shaped with respect to β_S by Step 1 (with the minimum at $\beta_S = \beta_R$), the claim follows.

Step 4. Consider finally the perceived disagreement from S's perspective. Depending on the equilibrium disclosure strategy, we have

$$E_S[\Delta^{FD}] = \Pr[\sigma_S = 1|\beta_S]\Delta(1, \beta_S, \beta_R) + \Pr[\sigma_S = 0|\beta_S]\Delta(0, \beta_S, \beta_R) + \Pr[\sigma_S = \varnothing|\beta_S]|\beta_S - \beta_R|, \tag{26}$$

$$E_S[\Delta^{D0}] = \Pr[\sigma_S = 0|\beta_S]\Delta(0, \beta_S, \beta_R) + (1 - \Pr[\sigma_S = 0|\beta_S])\Delta^{D0}(\varnothing, \beta_S, \beta_R), \tag{27}$$

$$E_S[\Delta^{D1}] = \Pr[\sigma_S = 1|\beta_S]\Delta(1, \beta_S, \beta_R) + (1 - \Pr[\sigma_S = 1|\beta_S])\Delta^{D1}(\varnothing, \beta_S, \beta_R). \tag{28}$$

Note now that by Steps 1 and 2 all terms in the above three expressions are V-shaped with respect to β_R, reaching its minimum of 0 at $\beta_S = \beta_R$. Hence, it remains to show that the overall expected disagreement $E_S[\Delta]$ is continuous at the points where it switches from $E_S[\Delta^{D0}]$ to $E_S[\Delta^{FD}]$ and from $E_S[\Delta^{FD}]$ to $E_S[\Delta^{D1}]$, i.e. where $\beta_S = \beta^*(\beta_R)$ and $\beta_S = \beta^{**}(\beta_R)$ (see Proposition 1). By the proof of Lemma II.G, when $\beta_S = \beta^*(\beta_R)$, S is indifferent between disclosing $\sigma = 1$ and non-disclosure in D0, i.e.

$$\Delta(1, \beta^*(\beta_R), \beta_R) = \Delta^{D0}(\varnothing, \beta^*(\beta_R), \beta_R). \tag{29}$$

Note that by (3) $\Delta^{D0}(\varnothing, \beta^*(\beta_R), \beta_R)$ is a weighted average between $\Delta(1, \beta^*(\beta_R), \beta_R)$ and $|\beta^*(\beta_R) - \beta_R|$. Together with (29), this implies

$$\Delta(1, \beta^*(\beta_R), \beta_R) = |\beta^*(\beta_R) - \beta_R|. \tag{30}$$

This together (29), (26) and (27) imply that $E_S[\Delta^{FD}|\beta_S = \beta^*(\beta_R)] = E_S[\Delta^{D0}|\beta_S = \beta^*(\beta_R)]$, as well as to the expected disagreement under any mixing between disclosure and non-disclosure of $\sigma = 1$ (which is the only possible mixed-strategy equilibrium if $\beta_S = \beta^*(\beta_R)$ by the proof of Lemma II.G). Consequently, $E_S[\Delta]$ is continuous at $\beta_S = \beta^*(\beta_R)$. By the analogous argument, $E_S[\Delta]$ is continuous at $\beta_S = \beta^{**}(\beta_R)$. ■

5 Appendix III: Hidden cost of PC with binary signals

5.0.1 Proof of Proposition 2

Step 1 Consider the case $\beta_S > \beta_R$ in D0 equilibrium. The expected perceived disagreement for S is

$$E_S[\Delta^{D0}] = (\Pr[\sigma = 1] + \Pr[\sigma = \varnothing])(E_R^{D0}[\widetilde{\beta}_S|\varnothing] - \widetilde{\beta}_R^{D0}(\varnothing)) + \Pr[\sigma = 0](\widetilde{\beta}_S(0) - \widetilde{\beta}_R(0)).$$

At the same time, under full disclosure

$$E_S[\Delta^{FD}] = \Pr[\sigma = 1](\widetilde{\beta}_S(1) - \widetilde{\beta}_R(1)) + \Pr[\sigma = 0](\widetilde{\beta}_S(0) - \widetilde{\beta}_R(0)) + \Pr[\sigma = \varnothing](\beta_S - \beta_R).$$

Then, using the expressions from Appendix I

$$\begin{aligned}
& E_S[\Delta^{D0}] - E_S[\Delta^{FD}] \\
= \; & \Pr[\sigma = 1](E_R^{D0}[\widetilde{\beta}_S|\varnothing] - \widetilde{\beta}_R^{D0}(\varnothing) - (\widetilde{\beta}_S(1) - \widetilde{\beta}_R(1))) + \Pr[\sigma = \varnothing](E_R^{D0}[\widetilde{\beta}_S|\varnothing] - \widetilde{\beta}_R^{D0}(\varnothing) \\
& -(\beta_S - \beta_R)) \\
= \; & \varphi(\beta_S p + (1-\beta_S)(1-p)) \\
& \times \left(\begin{array}{c} \left(\frac{\varphi(\beta_R p + (1-\beta_R)(1-p))}{\varphi(\beta_R p + (1-\beta_R)(1-p)) + (1-\varphi)} - 1\right)\left(\widetilde{\beta}_S(1) - \widetilde{\beta}_R(1)\right) \\ + \left(\frac{(1-\varphi)}{\beta_R \varphi p + (1-\beta_R)\varphi(1-p) + (1-\varphi)}\right)(\beta_S - \beta_R) \end{array} \right) \\
& + (1-\varphi) \left(\begin{array}{c} \left(\frac{\varphi(\beta_R p + (1-\beta_R)(1-p))}{\varphi(\beta_R p + (1-\beta_R)(1-p)) + (1-\varphi)}\right)\left(\widetilde{\beta}_S(1) - \widetilde{\beta}_R(1)\right) \\ + \left(\frac{(1-\varphi)}{\beta_R \varphi p + (1-\beta_R)\varphi(1-p) + (1-\varphi)} - 1\right)(\beta_S - \beta_R) \end{array} \right) \\
= \; & \Phi_1 \Phi_2
\end{aligned}$$

where

$$\begin{aligned}
\Phi_1 &= \frac{(\beta_S - \beta_R)^2 (1-2p)^2 (1-\varphi)\varphi}{(\beta_R p + (1-\beta_R)(1-p))(\beta_S p + (1-\beta_S)(1-p))(1 - p\varphi + \beta_R \varphi(2p-1))} > 0, \\
\Phi_2 &= (\beta_R + \beta_S - 1)(1-p) + \beta_R \beta_S (2p-1).
\end{aligned}$$

Note that Φ_2 is an increasing function of β_S. At the same time, by Proposition 1, $\beta_S < \beta_S^*$ in D0 equilibrium. Consequently,

$$\begin{aligned}
\Phi_2(\beta_S) &< \Phi_2(\beta_S^*) = \left(\beta_R + \frac{(1-\beta_R)(1-p)}{1-p+\beta_R(2p-1)} - 1\right)(1-p) \\
+\beta_R \frac{(1-\beta_R)(1-p)}{1-p+\beta_R(2p-1)}(2p-1) &= 0.
\end{aligned}$$

Hence, $\Phi_1\Phi_2 < 0$ so that

$$E_S[\Delta^{D0}] - E_S[\Delta^{FD}] < 0,$$

i.e., the sender would ex-ante prefer D0 over FD.

Step 2 Consider the case $\beta_S > \beta_R$ in D1 equilibrium. The expected perceived disagreement in equilibrium for S is

$$E_S[\Delta^{D1}] = (\Pr[\sigma = 0] + \Pr[\sigma = \varnothing])(E_R^{D1}[\widetilde{\beta}_S|\varnothing] - \widetilde{\beta}_R^{D1}(\varnothing)) + \Pr[\sigma = 1](\widetilde{\beta}_S(1) - \widetilde{\beta}_R(1))$$

Then,

$$\begin{aligned}
& E_S[\Delta^{D1}] - E_S[\Delta^{FD}] \\
= \;& \varphi(\beta_S(1-p) + (1-\beta_S)p) \\
& \times \left(\begin{array}{c} \left(\frac{\varphi(\beta_R(1-p)+(1-\beta_R)p)}{\varphi(\beta_R(1-p)+(1-\beta_R)p)+(1-\varphi)} - 1\right)\left(\widetilde{\beta}_S(0) - \widetilde{\beta}_R(0)\right) \\ + \left(\frac{(1-\varphi)}{\varphi(\beta_R(1-p)+(1-\beta_R)p)+(1-\varphi)}\right)(\beta_S - \beta_R) \end{array} \right) \\
& + (1-\varphi) \left(\begin{array}{c} \left(\frac{\varphi(\beta_R(1-p)+(1-\beta_R)p)}{\varphi(\beta_R(1-p)+(1-\beta_R)p)+(1-\varphi)}\right)\left(\widetilde{\beta}_S(0) - \widetilde{\beta}_R(0)\right) \\ + \left(\frac{(1-\varphi)}{\varphi(\beta_R(1-p)+(1-\beta_R)p)+(1-\varphi)} - 1\right)(\beta_S - \beta_R) \end{array} \right) \\
= \;& \Phi_3\Phi_4,
\end{aligned}$$

where

$$\begin{aligned}
\Phi_3 &= -\frac{(\beta_S - \beta_R)^2(1-2p)^2(1-\varphi)\varphi}{(\beta_R(1-p) + (1-\beta_R)p)(\beta_S(1-p) + (1-\beta_S)p)} \frac{1}{1 - \varphi((1-\beta_R)(1-p) + \beta_R p)} < 0, \\
\Phi_4 &= p(1-\beta_R) - \beta_S(p(1-\beta_R) + \beta_R(1-p)).
\end{aligned}$$

Function Φ_4 is decreasing in β_S. At the same time, by Proposition 1 in D1-equilibrium we have $\beta_S > \beta_S^{**}$. Consequently,

$$\Phi_4(\beta_S) < \Phi_4(\beta_S^{**}) = p(1-\beta_R) - \frac{p(1-\beta_R)}{\beta_R + p(1-2\beta_R)}(p(1-\beta_R) + \beta_R(1-p)) = 0.$$

Hence, $\Phi_3\Phi_4 > 0$, i.e.

$$E_S[\Delta^{D1}] - E_S[\Delta^{FD}] > 0,$$

i.e., the sender would ex-ante prefer FD over D1.

Step 3 Consider the case $\beta_S < \beta_R$. Then, the expressions for disagreement from Steps 1 and 2 just switch signs so that

$$\begin{aligned} E_S[\Delta^{D0}] - E_S[\Delta^{FD}] &= -\Phi_1\Phi_2 > 0, \\ E_S[\Delta^{D1}] - E_S[\Delta^{FD}] &= -\Phi_3\Phi_4 < 0. \end{aligned}$$

Thus, the sender would ex-ante prefer FD over D0 and D1 over FD whenever D0 and D1 are the unique equilibria, respectively. ■

5.0.2 Proof of Proposition 3

Step 1 In Steps 1-4 below, we consider the case that $\beta_S > \beta_R$. Define as $\widetilde{\Theta}(\text{Partial},\widehat{\beta})$ and $\widetilde{\Theta}(\text{Full},\widehat{\beta})$ the expected actual disagreement under partial and full disclosure respectively, from the perspective of a third party endowed with prior $\widehat{\beta}$. Denote further by $\widetilde{\beta}_i(\iota,\text{Partial})$ and $\widetilde{\beta}_i(\iota,\text{Full})$ the posterior of player i conditional on obtained information ι under partial and full disclosure respectively. We have:

$$\begin{aligned} \widetilde{\Theta}(\text{Partial},\widehat{\beta}) &= E_{\widehat{\beta}}\left[\left|\widetilde{\beta}_S(\sigma,\text{Partial}) - \widetilde{\beta}_R(d,\text{Partial})\right|\right] \\ &\geq E_{\widehat{\beta}}\left[\widetilde{\beta}_S(\sigma,\text{Partial}) - \widetilde{\beta}_R(d,\text{Partial})\right] \\ &= E_{\widehat{\beta}}[\widetilde{\beta}_S(\sigma,\text{Partial})] - E_{\widehat{\beta}}[\widetilde{\beta}_R(d,\text{Partial})] \\ &= E_{\widehat{\beta}}\left[\widetilde{\beta}_S(\sigma,\text{Full})\right] - E_{\widehat{\beta}}\left[\widetilde{\beta}_R(d,\text{Partial})\right]. \end{aligned} \tag{31}$$

In the above, the equality $E_{\widehat{\beta}}[\widetilde{\beta}_S(\sigma,\text{Partial})] = E_{\widehat{\beta}}\left[\widetilde{\beta}_S(\sigma,\text{Full})\right]$ follows from the fact that S's expected posterior is the same under both full and partial disclosure. Note on the other hand that

$$\begin{aligned} \widetilde{\Theta}(\text{Full},\widehat{\beta}) &= E_{\widehat{\beta}}\left[\left|\widetilde{\beta}_S(\sigma,\text{Full}) - \widetilde{\beta}_R(d,\text{Full})\right|\right] \\ &= E_{\widehat{\beta}}\left[\widetilde{\beta}_S(\sigma,\text{Full})\right] - E_{\widehat{\beta}}\left[\widetilde{\beta}_R(d,\text{Full})\right]. \end{aligned} \tag{32}$$

It follows from the above that

$$\widetilde{\Theta}(\text{Partial},\widehat{\beta}) - \widetilde{\Theta}(\text{Full},\widehat{\beta}) \geq E_{\widehat{\beta}}\left[\widetilde{\beta}_R(d,\text{Full})\right] - E_{\widehat{\beta}}\left[\widetilde{\beta}_R(d,\text{Partial})\right]. \tag{33}$$

Step 2 We now show that $E_{\widehat{\beta}}\left[\widetilde{\beta}_R(d,\text{Full})\right] - E_{\widehat{\beta}}\left[\widetilde{\beta}_R(d,\text{Partial})\right] > 0$ if and only if $\widehat{\beta} > \beta_R$. Here we simply follow the analysis presented in Kartik et al. (2015) (the result is directly implied by their Theorem 1). One can verify that

$$\widetilde{\beta}_R(d) = \frac{\widehat{\beta}(d)\frac{\beta_R}{\widehat{\beta}}}{\widehat{\beta}(d)\frac{\beta_R}{\widehat{\beta}} + (1-\widehat{\beta}(d))\frac{1-\beta_R}{1-\widehat{\beta}}},$$

where $\widehat{\beta}(d)$ is the posterior belief of the receiver had she had a prior $\beta_R = \widehat{\beta}$. One can verify that the above function is concave in $\widehat{\beta}(d)$ if $\widehat{\beta} < \beta_R$ and convex if the opposite inequality holds. Blackwell (1953) has shown that a garbling increases (resp. reduces) an individual's expectation of any concave (resp. convex) function of his posterior. Then, since partial disclosure is a garbling of full disclosure,[12] we obtain that

$$E_{\widehat{\beta}}\left[\widetilde{\beta}_R(d,\text{Partial})\right] < (>) E_{\widehat{\beta}}\left[\widetilde{\beta}_R(d,\text{Full})\right] \text{ if } \widehat{\beta} > (<)\beta_R \tag{34}$$

given that R's posterior is a convex (concave) function of $\widehat{\beta}(\sigma)$ if $\widehat{\beta} > (<)\beta_R$.

Step 3 (33) and (34) together imply

$$\widetilde{\Theta}(\text{Partial},\widehat{\beta}) - \widetilde{\Theta}(\text{Full},\widehat{\beta}) > 0 \text{ if } \widehat{\beta} > \beta_R.$$

Thus, the third party would prefer full disclosure over partial disclosure whenever $\widehat{\beta} > \beta_R$, i.e., whenever $\beta_R < \widehat{\beta} < \beta_S$ or $\widehat{\beta} \geq \beta_S > \beta_R$.

Step 4 Consider $\widehat{\beta} < \beta_R < \beta_S$ with β_S being sufficiently close to 1. We have

$$\begin{aligned}
\widetilde{\Theta}(\text{Partial},\widehat{\beta}) &= E_{\widehat{\beta}}\left[\left|\widetilde{\beta}_S(\sigma,\text{Partial}) - \widetilde{\beta}_R(d,\text{Partial})\right|\right] \\
&= E_{\widehat{\beta}}\left[\widetilde{\beta}_S(\sigma,\text{Partial}) - \widetilde{\beta}_R(d,\text{Partial})\right] \\
&= E_{\widehat{\beta}}[\widetilde{\beta}_S(\sigma,\text{Partial})] - E_{\widehat{\beta}}[\widetilde{\beta}_R(d,\text{Partial})] \\
&= E_{\widehat{\beta}}\left[\widetilde{\beta}_S(\sigma,\text{Full})\right] - E_{\widehat{\beta}}\left[\widetilde{\beta}_R(d,\text{Partial})\right]
\end{aligned}$$

(i.e., we have equalities at all stages in contrast to (31)). This together with (32) and (34) implies

$$\widetilde{\Theta}(\text{Partial},\widehat{\beta}) - \widetilde{\Theta}(\text{Full},\widehat{\beta}) = E_{\widehat{\beta}}\left[\widetilde{\beta}_R(d,\text{Full})\right] - E_{\widehat{\beta}}\left[\widetilde{\beta}_R(d,\text{Partial})\right] < 0.$$

[12] See Kartik et al. (2015) for a formal definition of garbling.

Hence, in this case the third party would prefer partial disclosure over full disclosure in terms of minimizing expected actual disagreement.

Step 5 The proof for the remaining case of $\beta_S < \beta_R$ is conceptually identical, and is hence omitted. In particular, we obtain that

$$\begin{aligned}\widetilde{\Theta}(\text{Partial},\widehat{\beta}) - \widetilde{\Theta}(\text{Full},\widehat{\beta}) &> 0 \text{ if } \widehat{\beta} < \beta_R,\\ \widetilde{\Theta}(\text{Partial},\widehat{\beta}) - \widetilde{\Theta}(\text{Full},\widehat{\beta}) &< 0 \text{ if } \beta_S < \beta_R < \widehat{\beta} \text{ and } \beta_S \text{ is close to } 0.\end{aligned}$$

■

5.1 Appendix IV: Disclosure with binary signals and prior uncertainty

5.1.1 Proof of Proposition 4.a)

Step 1 Consider a putative FD equilibrium. Let $G_S(G_R)$ denote the (symmetric) cumulative distribution function of S's (R's) prior belief. Then, if the sender discloses 0-signal, the receiver with the prior β_R believes that the disagreement is

$$E_R[\Delta^{FD}(1)] = \int_{\beta_S=0}^{1} \left|\widetilde{\beta}_S(0) - \widetilde{\beta}_R(0)\right| dG_S(\beta_S).$$

In turn, the sender expects that the receiver's perceived disagreement is

$$E_S E_R[\Delta^{FD}(1)] = \int_{\beta_R=0}^{1}\int_{\beta_S=0}^{1} \left|\widetilde{\beta}_S(0) - \widetilde{\beta}_R(0)\right| dG_S(\beta_S) dG_R(\beta_R).$$

If the sender does not disclose, the expected perceived disagreement is

$$E_S E_R[\Delta^{FD}(\varnothing)] = \int_{\beta_R=0}^{1}\int_{\beta_S=0}^{1} |\beta_S - \beta_R| dG_S(\beta_S) dG_R(\beta_R).$$

In FD equilibrium we must have $E_S E_R[\Delta^{FD}(1)] - E_S E_R[\Delta^{FD}(\varnothing)] < 0$. We have

$$\begin{aligned}&E_S E_R[\Delta^{FD}(1)] - E_S E_R[\Delta^{FD}(\varnothing)]\\ =\ &\int_{\beta_R=0}^{1}\int_{\beta_S=0}^{1} \left(\left|\widetilde{\beta}_S(0) - \widetilde{\beta}_R(0)\right| - |\beta_S - \beta_R|\right) dG_S(\beta_S) dG_R(\beta_R).\end{aligned}$$

Denote $\widetilde{\beta}(\sigma, \beta)$ the posterior belief given obtained/disclosed signal σ and prior belief β. Besides, denote $\kappa(\beta_i, \beta_j) = \left|\widetilde{\beta}(1, \beta_i) - \widetilde{\beta}(1, \beta_j)\right| - |\beta_i - \beta_j|$. Then,

$$\begin{aligned}
& \int_{\beta_R=0}^{1}\int_{\beta_S=0}^{1}\left(\left|\widetilde{\beta}_S(0) - \widetilde{\beta}_R(0)\right| - |\beta_S - \beta_R|\right) dG_S(\beta_S)dG_R(\beta_R) \\
= & \int_{\beta_R=0}^{1}\int_{\beta_S=0}^{1}\kappa(\beta_S, \beta_R)dG_S(\beta_S)dG_R(\beta_R) \\
= & \int_{\beta_R=0}^{0.5}\int_{\beta_S=0}^{1}\kappa(\beta_S, \beta_R)dG_S(\beta_S)dG_R(\beta_R) \\
& + \int_{\beta_R=0}^{0.5}\int_{\beta_S=0}^{1}\kappa(\beta_S, 1-\beta_R)dG_S(\beta_S)dG_R(1-\beta_R) \\
= & \int_{\beta_R=0}^{0.5}\int_{\beta_S=0}^{1}\kappa(\beta_S, \beta_R)dG_S(\beta_S)dG_R(\beta_R) \\
& + \int_{\beta_R=0}^{0.5}\int_{\beta_S=0}^{1}\kappa(\beta_S, 1-\beta_R)dG_S(\beta_S)dG_R(\beta_R) \\
= & \int_{\beta_R=0}^{0.5}\int_{\beta_S=0}^{1}\left(\kappa(\beta_S, \beta_R) + \kappa(\beta_S, 1-\beta_R)\right) dG_S(\beta_S)dG_R(\beta_R),
\end{aligned}$$

where the third equality follows due to symmetry of G. Next, denote $\lambda(\beta_S, \beta_R) = \kappa(\beta_S, \beta_R) + \kappa(\beta_S, 1-\beta_R)$. Then, similarly,

$$\begin{aligned}
& \int_{\beta_R=0}^{0.5}\int_{\beta_S=0}^{1}\left(\kappa(\beta_S, \beta_R) + \kappa(\beta_S, 1-\beta_R)\right) dG_S(\beta_S)dG_R(\beta_R) && (35) \\
= & \int_{\beta_R=0}^{0.5}\int_{\beta_S=0}^{1}\lambda(\beta_S, \beta_R)dG_S(\beta_S)dG_R(\beta_R) \\
= & \int_{\beta_R=0}^{0.5}\left(\int_{\beta_S=0}^{0.5}\lambda(\beta_S, \beta_R)dG_S(\beta_S) + \int_{\beta_S=0}^{0.5}\lambda(1-\beta_S, \beta_R)dG_S(1-\beta_S)\right) dG_R(\beta_R) && (36) \\
= & \int_{\beta_R=0}^{0.5}\int_{\beta_S=0}^{0.5}(\lambda(\beta_S, \beta_R) + \lambda(1-\beta_S, \beta_R))dG_S(\beta_S)dG_R(\beta_R). && (37)
\end{aligned}$$

Let us now show that $\lambda(\beta_S, \beta_R) + \lambda(1-\beta_S, \beta_R) < 0$ for any $\beta_S < 0.5$ and $\beta_R < 0.5$ in which case the whole integral on the right-hand side is negative. Denote as before $x = \max\{\beta_S, \beta_R\}$ and $y = \min\{\beta_S, \beta_R\}$. Then, (noting that $1-y > 1-x > x > y$ due to

both $x < 0.5$ and $y < 0.5$)

$$\begin{aligned}
& \lambda(\beta_S, \beta_R) + \lambda(1-\beta_S, \beta_R) \\
= \; & \kappa(\beta_S, \beta_R) + \kappa(\beta_S, 1-\beta_R) + \kappa(1-\beta_S, \beta_R) + \kappa(1-\beta_S, 1-\beta_R) \\
= \; & \left(\widetilde{\beta}(1,x) - \widetilde{\beta}(1,y)\right) - (x-y) \\
& + \left(\widetilde{\beta}(1,1-x) - \widetilde{\beta}(1,y)\right) - (1-x-y) \\
& + \left(\widetilde{\beta}(1,1-y) - \widetilde{\beta}(1,x)\right) - (1-y-x) \\
& + \left(\widetilde{\beta}(1,1-y) - \widetilde{\beta}(1,1-x)\right) - (1-y-(1-x)) \\
= \; & 2(\widetilde{\beta}(1,1-y) - \widetilde{\beta}(1,y) + 2y - 1) \\
= \; & 2\left(\frac{(1-y)(1-p)}{(1-y)(1-p)+yp} - \frac{y(1-p)}{y(1-p)+(1-y)p} + 2y - 1\right) \\
= \; & -\frac{2(1-2p)^2(1-y)(1-2y)y}{(1-p+y(2p-1))(y+p(1-2y))} < 0,
\end{aligned}$$

where the inequality follows due to $y < 0.5$.

Step 2 By symmetry considerations, the same property holds for 1-signals, i.e. $E_S E_R[\Delta^{FD}(0)] - E_S E_R[\Delta^{FD}(\varnothing)] < 0$. Formally, the proof proceeds analogously redefining $\kappa(\beta_i, \beta_j) = \left|\widetilde{\beta}(0,\beta_i) - \widetilde{\beta}(0,\beta_j)\right| - |\beta_i - \beta_j|$. ■

5.1.2 Proof of Proposition 4.b)

In what follows, we assume without loss of generality that MLRP is satisfied as

$$\frac{\partial}{\partial x}\frac{g_S(x)}{g_R(x)} > 0. \tag{38}$$

Step 1. Denote the difference in disagreement under disclosure and no disclosure in a putative FD-equilibrium as

$$\begin{aligned}
\kappa_0(\beta_S, \beta_R) &= |\beta_S - \beta_R| - \left|\widetilde{\beta}(0,\beta_S) - \widetilde{\beta}(0,\beta_R)\right|, \\
\kappa_1(\beta_S, \beta_R) &= |\beta_S - \beta_R| - \left|\widetilde{\beta}(1,\beta_S) - \widetilde{\beta}(1,\beta_R)\right|.
\end{aligned}$$

In FD, we have

$$\begin{aligned}\int_{\beta_R=0}^{1}\int_{\beta_S=0}^{1}\kappa_0(\beta_S,\beta_R)g_S(\beta_S)g_R(\beta_R)d\beta_S d\beta_R &\geq 0,\\ \int_{\beta_R=0}^{1}\int_{\beta_S=0}^{1}\kappa_1(\beta_S,\beta_R)g_S(\beta_S)g_R(\beta_R)d\beta_S d\beta_R &\geq 0.\end{aligned}$$

Since the joint distribution of priors is completely symmetric with respect to either boundary (0 or 1), the effect of 0-disclosure on the expected disagreement should be equivalent to the effect of 1-disclosure, i.e.

$$\int_{\beta_R=0}^{1}\int_{\beta_S=0}^{1}\kappa_0(\beta_S,\beta_R)g_S(\beta_S)g_R(\beta_R)d\beta_S d\beta_R = \int_{\beta_R=0}^{1}\int_{\beta_S=0}^{1}\kappa_1(\beta_S,\beta_R)g_S(\beta_S)g_R(\beta_R)d\beta_S d\beta_R.$$

This implies that for $i = 0, 1$

$$\int_{\beta_R=0}^{1}\int_{\beta_S=0}^{1}\kappa_i(\beta_S,\beta_R)g_S(\beta_S)g_R(\beta_R)d\beta_S d\beta_R \geq 0 \tag{39}$$

$$\Leftrightarrow \int_{\beta_R=0}^{1}\int_{\beta_S=0}^{1}\eta(\beta_S,\beta_R)g_S(\beta_S)g_R(\beta_R)d\beta_S d\beta_R \geq 0.$$

where $\eta(\beta_S,\beta_R) = \kappa_0(\beta_S,\beta_R) + \kappa_1(\beta_S,\beta_R)$.

Step 2. We have

$$\begin{aligned}
&\int_{\beta_R=0}^{1}\int_{\beta_S=0}^{1}\eta(\beta_S,\beta_R)g_S(\beta_S)g_R(\beta_R)d\beta_S d\beta_R \\
=\ &\int_{\beta_R=0}^{0.5}\int_{\beta_S=0}^{1}\eta(\beta_S,\beta_R)g_S(\beta_S)g_R(\beta_R)d\beta_S d\beta_R \\
&+\int_{\beta_R=0}^{0.5}\int_{\beta_S=0}^{1}\eta(\beta_S,1-\beta_R)g_S(\beta_S)g_R(1-\beta_R)d\beta_S d\beta_R \\
=\ &\int_{\beta_R=0}^{0.5}\int_{\beta_S=0}^{0.5}\eta(\beta_S,\beta_R)g_S(\beta_S)g_R(\beta_R)d\beta_S d\beta_R \\
&+\int_{\beta_R=0}^{0.5}\int_{\beta_S=0}^{0.5}\eta(1-\beta_S,\beta_R)g_S(1-\beta_S)g_R(\beta_R)d\beta_S d\beta_R \\
&+\int_{\beta_R=0}^{0.5}\int_{\beta_S=0}^{0.5}\eta(\beta_S,1-\beta_R)g_S(\beta_S)g_R(1-\beta_R)d\beta_S d\beta_R \\
&+\int_{\beta_R=0}^{0.5}\int_{\beta_S=0}^{0.5}\eta(1-\beta_S,1-\beta_R)g_S(1-\beta_S)g_R(1-\beta_R)d\beta_S d\beta_R \\
=\ &\int_{\beta_R=0}^{0.5}\int_{\beta_S=0}^{0.5}\varsigma(\beta_S,\beta_R)d\beta_S d\beta_R,
\end{aligned}$$

where

$$\begin{aligned}
\varsigma(\beta_S,\beta_R) =\ &\eta(\beta_S,\beta_R)g_S(\beta_S)g_R(\beta_R)+\eta(1-\beta_S,\beta_R)g_S(1-\beta_S)g_R(\beta_R) \\
&+\eta(\beta_S,1-\beta_R)g_S(\beta_S)g_R(1-\beta_R)+\eta(1-\beta_S,1-\beta_R)g_S(1-\beta_S)g_R(1-\beta_R).
\end{aligned}$$

Hence, given Step 1, for the main claim it is sufficient to show that $\varsigma(\beta_S,\beta_R)\geq 0$ for any $\{\beta_S,\beta_R\}\in[0,0.5]^2$.

Step 3. Let us show that $\varsigma(\beta_S,\beta_R)$ is increasing in p for $p\in(1/2,1)$ and $\{\beta_S,\beta_R\}\in[0,0.5]^2$. To simplify the notation, let us denote $g_S(\beta_S)\equiv g_{S1}$, $g_R(\beta_R)\equiv g_{R1}$, $g_S(1-\beta_S)\equiv g_{S2}$, $g_R(1-\beta_R)\equiv g_{R2}$.

Consider first $0.5\geq\beta_R>\beta_S$. Substituting all expressions into $\varsigma(\beta_S,\beta_R)$ and simplifying, we obtain

$$\begin{aligned}
\varsigma(\beta_S,\beta_R) =\ &\tau_1(\beta_R+\beta_S-1)(g_{R2}g_{S1}+g_{R1}g_{S2}) \\
&+\tau_2(\beta_R-\beta_S)(g_{R1}g_{S1}+g_{R2}g_{S2}),
\end{aligned}$$

where

$$\tau_1 = -2 - \frac{(1-p)p}{(\beta_R + p - 2\beta_R p)(\beta_S + p - 2\beta_S p - 1)} - \frac{(1-p)p}{(\beta_R + p - 2\beta_R p - 1)(\beta_S + p - 2\beta_S p)},$$

$$\tau_2 = 2 - \frac{(1-p)p}{(\beta_R + p - 2\beta_R p - 1)(\beta_S + p - 2\beta_S p - 1)} - \frac{(1-p)p}{(\beta_R + p - 2\beta_R p)(\beta_S + p - 2\beta_S p)}.$$

Taking the derivative of $\varsigma(\beta_S, \beta_R)$ with respect to p and simplifying we obtain

$$\frac{\partial \varsigma(\beta_S, \beta_R)}{\partial p} = T_1 + T_2, \tag{40}$$

$$T_1 = (1-\beta_R)\beta_R \frac{(2p-1)(1-2\beta_R)}{(\beta_R + p - 2\beta_R p - 1)^2(\beta_R + p - 2\beta_R p)^2} \tag{41}$$

$$\times (g_{R2} - g_{R1})(g_{S1} - g_{S2}), \tag{42}$$

$$T_2 = (1-\beta_S)\beta_S \frac{(2p-1)(1-2\beta_S)}{(\beta_S + p - 2\beta_S p - 1)^2(\beta_S + p - 2\beta_S p)^2} \tag{43}$$

$$\times (g_{R2} + g_{R1})(g_{S1} + g_{S2}). \tag{44}$$

Consider now the case $\beta_R < \beta_S \leq 0.5$. Substituting all expressions into $\varsigma(\beta_S, \beta_R)$ and simplifying, we obtain in this case

$$\begin{aligned} \varsigma(\beta_S, \beta_R) = {} & \tau_1(\beta_R + \beta_S - 1)(g_{R2}g_{S1} + g_{R1}g_{S2}) \\ & + \tau_2(\beta_S - \beta_R)(g_{R1}g_{S1} + g_{R2}g_{S2}), \end{aligned}$$

Taking the derivative of $\varsigma(\beta_S, \beta_R)$ with respect to p and simplifying we obtain

$$\frac{\partial \varsigma(\beta_S, \beta_R)}{\partial p} = \widehat{T}_1 + \widehat{T}_2, \tag{45}$$

$$\widehat{T}_1 = (1-\beta_R)\beta_R \frac{(2p-1)(1-2\beta_R)}{(\beta_R + p - 2\beta_R p - 1)^2(\beta_R + p - 2\beta_R p)^2} \tag{46}$$

$$\times (g_{R2} + g_{R1})(g_{S1} + g_{S2}), \tag{47}$$

$$\widehat{T}_2 = (1-\beta_R)\beta_R \frac{(2p-1)(1-2\beta_S)}{(\beta_S + p - 2\beta_S p - 1)^2(\beta_S + p - 2\beta_S p)^2} \tag{48}$$

$$\times (g_{R2} - g_{R1})(g_{S1} - g_{S2}). \tag{49}$$

Recall that $\{\beta_S, \beta_R\} \in [0, 0.5]^2$ by assumption. Hence, to show that $\frac{\partial \varsigma(\beta_S, \beta_R)}{\partial p} \geq 0$ in either case we need to show that

$$(g_{R2} - g_{R1})(g_{S1} - g_{S2}) > 0.$$

This is done in the next step.

Step 4. By initial assumption, we have that for any x

$$g_R(x) = g_S(1-x).$$

In particular, this implies

$$\frac{g_R(0.5)}{g_S(0.5)} = 1.$$

Note that then the MLRP in (38) implies that for any $\beta_R < 0.5$ and $\beta_S < 0.5$

$$\begin{aligned} \frac{g_S(\beta_S)}{g_R(\beta_S)} &< \frac{g_S(0.5)}{g_R(0.5)} < \frac{g_S(1-\beta_R)}{g_R(1-\beta_R)}, \\ \frac{g_S(\beta_S)}{g_R(\beta_S)} &< 1 < \frac{g_S(1-\beta_R)}{g_R(1-\beta_R)}. \end{aligned} \tag{50}$$

Since by initial assumption $g_R(x) = g_S(1-x)$, (50) is equivalent to

$$\frac{g_S(\beta_S)}{g_S(1-\beta_S)} < 1 < \frac{g_R(\beta_R)}{g_R(1-\beta_R)}.$$

In terms of our previous notation, this is equivalent to

$$\begin{aligned} g_{S1} &< g_{S2}, \\ g_{R1} &> g_{R2.} \end{aligned}$$

Finally, this leads to

$$(g_{R2} - g_{R1})(g_{S1} - g_{S2}) > 0. \tag{51}$$

Step 5. Applying (51) to the expressions for $\frac{\partial \varsigma(\beta_S, \beta_R)}{\partial p}$ from Step 3, we obtain

$$\frac{\partial \varsigma(\beta_S, \beta_R)}{\partial p} \geq 0.$$

At the same time, it is easy to verify that $\varsigma(\beta_S, \beta_R) = 0$ for $p = 1/2$. Consequently, $\varsigma(\beta_S, \beta_R) \geq 0$ for any $p > 1/2$. Then, by Step 2 this results in

$$\int_{\beta_R=0}^{1} \int_{\beta_S=0}^{1} \eta(\beta_S, \beta_R) g_S(\beta_S) g_R(\beta_R) d\beta_S d\beta_R \geq 0.$$

By Step 1, this implies that the incentive constraints for full disclosure are satisfied.

■

5.1.3 Proof of Proposition 4.c)

Step 1. Let us show that for sufficiently high x, it holds that $\beta_S > \beta_S^{**}(\beta_R) = \frac{p(1-\beta_R)}{\beta_R + p(1-2\beta_R)}$ for any $\{\beta_S, \beta_R\} \in [x,1]^2$. Indeed, it is easy to verify that $x > \beta_S^{**}(x)$ if and only if $x > \frac{p}{p+\sqrt{p(1-p)}}$. Thus, we have that for any $\{\beta_S, \beta_R\} \in [x,1]^2$ it holds

$$\beta_S \geq x > \beta^{**}(x) \geq \beta^{**}(\beta_R),$$

where the last inequality is due to $\beta_S^{**}(x)$ decreasing in x. Hence, $\beta_S > \beta_S^{**}(\beta_R)$ for any $\{\beta_S, \beta_R\} \in [x,1]^2$.

Analogously, one can show that for any sufficiently small y (in particular, for any $y < \frac{p+\sqrt{p(1-p)}-1}{2p-1}$), it holds $\beta_S < \beta_S^{*}(\beta_R)$ for any $\{\beta_S, \beta_R\} \in [0,y]^2$.

Step 2. Let us show that if the common distribution of priors g is shifted to the right, then D1 equilibrium always exists. The incentive constraints for D1 are (see Step 1 in the proof of Proposition 4.a)

$$\int_{\beta_R=0}^{1}\int_{\beta_S=0}^{1} \widetilde{\kappa}_0(\beta_S,\beta_R) g(\beta_S) g(\beta_R) d\beta_S d\beta_R \leq 0, \tag{52}$$

$$\int_{\beta_R=0}^{1}\int_{\beta_S=0}^{1} \widetilde{\kappa}_1(\beta_S,\beta_R) g(\beta_S) g(\beta_R) d\beta_S d\beta_R \geq 0, \tag{53}$$

where

$$\begin{aligned} \widetilde{\kappa}_0(\beta_S,\beta_R) &= \Delta^{D1}(\varnothing;\beta_S,\beta_R) - \Delta(0;\beta_S,\beta_R), \\ \widetilde{\kappa}_1(\beta_S,\beta_R) &= \Delta^{D1}(\varnothing;\beta_S,\beta_R) - \Delta(1;\beta_S,\beta_R). \end{aligned}$$

At the same time, for any constellation $\{\beta_S, \beta_R\} \in [x,1]^2$ we have $\beta_S > \beta_S^{**}(\beta_R)$ by Step 1, which then implies by Proposition 1

$$\begin{aligned} \widetilde{\kappa}_0(\beta_S,\beta_R) &\leq 0, \\ \widetilde{\kappa}_1(\beta_S,\beta_R) &\geq 0. \end{aligned}$$

Consequently,

$$\int_{\beta_R=x}^{1}\int_{\beta_S=x}^{1} \kappa_0(\beta_S,\beta_R) g_S(\beta_S) g_R(\beta_R) d\beta_S d\beta_R \leq 0, \tag{54}$$

$$\int_{\beta_R=x}^{1}\int_{\beta_S=x}^{1} \kappa_1(\beta_S,\beta_R) g_S(\beta_S) g_R(\beta_R) d\beta_S d\beta_R \geq 0. \tag{55}$$

Finally, (54) and (55) result in (52) and (53) as far as g is sufficiently skewed to the right.

Step 3. The non-existence of other pure strategy equilibria (besides D1) if g is sufficiently shifted to the right follows by the analogous argument. In particular, by Step 1 and Proposition 1 for any given constellation $\{\beta_S, \beta_R\} \in [x,1]^2$ the S's incentive constraints for other equilibria (D0 and FD) are not satisfied. Consequently, they are still not satisfied once we integrate them over all possible constellations $\{\beta_S, \beta_R\} \in [x,1]^2$ like in Step 2. If the probability mass set on $\{\beta_S, \beta_R\} \notin [x,1]^2$ gets sufficiently small, the same applies to the integration over all possible constellations $\{\beta_S, \beta_R\} \in [0,1]^2$.

Step 4. Consider the case when the distribution g is sufficiently skewed to the left, i.e. to values $[0,y]^2$. As before, Step 1 implies that for any given $\{\beta_S, \beta_R\} \in [0,y]^2$ we have $\beta_S < \beta_S^*(\beta_R)$, i.e. the S's incentive constraints for D0 are satisfied, while for D1 and FD they are not satisfied. Consequently, the same holds once we integrate them over all possible priors constellations in $[0,y]^2$, and hence in $[0,1]^2$ (under sufficiently skewed distribution).

5.1.4 Proof of Proposition 4.d)

Suppose that S's prior β_S is commonly known. That of R is drawn from a symmetric distribution G over $[0,1]$. Then, by the same steps as in the proof of Proposition 4.a we obtain

$$\begin{aligned} E_S E_R[\Delta^{FD}(1)] - E_S E_R[\Delta^{FD}(\varnothing)] &= \int_{\beta_R=0}^{1} \left(\left| \widetilde{\beta}_S(0) - \widetilde{\beta}_R(0) \right| - |\beta_S - \beta_R| \right) dG_R(\beta_R) \\ &= \int_{\beta_R=0}^{0.5} \left(\kappa(\beta_S, \beta_R) + \kappa(\beta_S, 1-\beta_R) \right) dG_R(\beta_R). \quad (56) \end{aligned}$$

Consider $\beta_R < 0.5$ such that $1-\beta_R > \beta_S > \beta_R$. For such β_R it holds

$$\begin{aligned}
\kappa(\beta_S,\beta_R)+\kappa(\beta_S,1-\beta_R) &= \left(\widetilde{\beta}(1,\beta_S)-\widetilde{\beta}(1,\beta_R)\right)-(\beta_S-\beta_R) \\
&\quad +\left(\widetilde{\beta}(1,1-\beta_R)-\widetilde{\beta}(1,\beta_S)\right)-(1-\beta_R-\beta_S) \\
&= \widetilde{\beta}(1,1-\beta_R)-\widetilde{\beta}(1,\beta_R)+2\beta_R-1 \\
&= \frac{(1-\beta_R)(1-p)}{(1-\beta_R)(1-p)+\beta_R p}-\frac{\beta_R(1-p)}{\beta_R(1-p)+(1-\beta_R)p}+2\beta_R-1 \\
&= -\frac{(1-2p)^2(1-\beta_R)(1-2\beta_R)\beta_R}{(1-p+\beta_R(2p-1))(\beta_R+p(1-2\beta_R))}<0.
\end{aligned}$$

Since the probability mass of $\beta_R < 0.5$ such that the condition $1-\beta_R > \beta_S > \beta_R$ is satisfied is sufficiently large for β_S sufficiently close to 0.5, the right-hand side of (56) is negative as well. Hence, the sender would prefer to disclose 0-signal over no disclosure. The same claim for 1-signals follows by symmetry considerations. Consequently, the FD equilibrium exists. ■

5.2 Appendix V: Endogenous matching (Proof of Proposition 5)

5.3 Appendix VI: Disclosure with continuous signals (Proof of Proposition 6)

Proposition 6 follows from a set of Lemmas, which are stated and proved in what follows. Below we denote the R's perceived disagreement for the cases of disclosure and non-disclosure in SDE as, respectively

$$\begin{aligned}
\Delta(\iota) &= \left|\widetilde{\beta}_S(\iota)-\widetilde{\beta}_R(\iota)\right| \text{ for } \iota\in[\underline{s},\bar{s}], \\
\Delta(\varnothing,s_1,s_2) &= \left|E_R[\widetilde{\beta}_S|\varnothing,s_1,s_2]-\widetilde{\beta}_R(\varnothing|s_1,s_2)\right|.
\end{aligned}$$

Lemma V.A *If $\beta_S \neq \beta_R$, then $\Delta(s)$ satisfies the following:*

i) There exists $\widehat{s}$ such that $\Delta(s)$ is increasing in s for all $s<\widehat{s}$ and decreasing in s for all $s>\widehat{s}$.

ii) $\widetilde{s} < (>)\widehat{s}$ if the player with the lower prior is less (more) extreme. Instead, $\widetilde{s}=\widehat{s}$ if $\beta_S = 1-\beta_R$, i.e. if players are equally extreme.

Proof:

Step 1 i) is immediate. To show ii) we first prove that there is a unique $\widehat{s}$ such that

$$\frac{d}{ds}\left(\widetilde{\beta}_S(\widehat{s}) - \widetilde{\beta}_R(\widehat{s})\right) = 0$$

Indeed,

$$\begin{aligned}
\frac{d}{ds}\left(\widetilde{\beta}_S(s) - \widetilde{\beta}_R(s)\right) &= \frac{d}{ds}\left(\frac{\beta_S}{\beta_S + (1-\beta_S)\frac{f(s|0)}{f(s|1)}} - \frac{\beta_R}{\beta_R + (1-\beta_R)\frac{f(s|0)}{f(s|1)}}\right) \\
&= \left(\frac{\beta_R(1-\beta_R)}{\left(\beta_R + (1-\beta_R)\frac{f(s|0)}{f(s|1)}\right)^2} - \frac{\beta_S(1-\beta_S)}{\left(\beta_S + (1-\beta_S)\frac{f(s|0)}{f(s|1)}\right)^2}\right)\frac{d}{ds}\frac{f(s|0)}{f(s|1)} \qquad (57)
\end{aligned}$$

Consider the solution to

$$\beta_R(1-\beta_R)\left(\beta_S + (1-\beta_S)\frac{f(s|0)}{f(s|1)}\right)^2 = \beta_S(1-\beta_S)\left(\beta_R + (1-\beta_R)\frac{f(s|0)}{f(s|1)}\right)^2.$$

Both sides are increasing in s, but we claim that they increase at different rates. To see this, note that

$$\begin{aligned}
&\frac{d}{ds}\beta_R(1-\beta_R)\left(\beta_S + (1-\beta_S)\frac{f(s|0)}{f(s|1)}\right)^2 \\
= \;& 2\beta_R(1-\beta_R)(1-\beta_S)\left(\beta_S + (1-\beta_S)\frac{f(s|0)}{f(s|1)}\right)\frac{d}{ds}\frac{f(s|0)}{f(s|1)}, \\
&\frac{d}{ds}\beta_S(1-\beta_S)\left(\beta_R + (1-\beta_R)\frac{f(s|0)}{f(s|1)}\right)^2 \\
= \;& 2\beta_S(1-\beta_R)(1-\beta_S)\left(\beta_R + (1-\beta_R)\frac{f(s|0)}{f(s|1)}\right)\frac{d}{ds}\frac{f(s|0)}{f(s|1)}.
\end{aligned}$$

The result then follows from the fact that

$$\begin{aligned}
&2\beta_R(1-\beta_R)(1-\beta_S)\left(\beta_S + (1-\beta_S)\frac{f(s|0)}{f(s|1)}\right)\frac{d}{ds}\frac{f(s|0)}{f(s|1)} \\
\gtreqless\; & 2\beta_S(1-\beta_R)(1-\beta_S)\left(\beta_R + (1-\beta_R)\frac{f(s|0)}{f(s|1)}\right)\frac{d}{ds}\frac{f(s|0)}{f(s|1)}
\end{aligned}$$

is equivalent to

$$\beta_R\beta_S + \beta_R(1-\beta_S)\frac{f(s|0)}{f(s|1)} \gtreqless \beta_R\beta_S + \beta_S(1-\beta_R)\frac{f(s|0)}{f(s|1)}$$

which, in turn, is equivalent to $\beta_R \gtreqless \beta_S$. Hence, $\widehat{s}$ (where $\Delta(s)$ reaches its extremum) must be unique. Then, claim ii) follows from continuity and (i) together with $\Delta(\widetilde{s}) = |\beta_S - \beta_R| > 0$.

Step 2 To show (iii), define again $x = \max\{\beta_S, \beta_R\}$ and $y = \min\{\beta_S, \beta_R\}$ such that $\Delta(s) = \widetilde{\beta}(s, x) - \widetilde{\beta}(s, y)$. From (57) we then have:

$$\begin{aligned} \frac{d}{ds}\Delta(\widetilde{s}) \quad &= \quad (y(1-y) - x(1-x)) \frac{d}{ds}\frac{f(\widetilde{s}|h)}{f(\widetilde{s}|l)} \gtreqless 0 \\ &\iff \quad y(1-y) \gtreqless x(1-x), \end{aligned}$$

so that by claim ii) $\widetilde{s} < (>)\widehat{s}$ if y is less (more) extreme than x, and $\widetilde{s} = \widehat{s}$ if players are equally extreme. ■

Lemma V.B *(i) If $\beta_S = \{\beta_R, 1-\beta_R\}$, then there exists FD.*
(ii) If $\beta_S \neq \{\beta_R, 1-\beta_R\}$, then a positive measure of signals is disclosed.
Proof:

Step 1. Let us show the existence of FD for $\beta_S = \{\beta_R, 1-\beta_R\}$. If $\beta_S = \beta_R$ then trivially $\Delta(s) = 0$ for any s, so that S has always an incentive to disclose s. If $\beta_S = 1 - \beta_R$, then $\widetilde{s} = \widehat{s}$ by Lemma V.A(ii). Consequently, for any $s \in [\underline{s}, \bar{s}]$ we obtain

$$\Delta^{FD}(\varnothing) = |\beta_S - \beta_R| = \Delta(\widetilde{s}) = \Delta(\widehat{s}) \geq \Delta(s),$$

where the last inequality is by Lemma V.A (i). Hence, S has an incentive to disclose all signals in equilibrium.

Step 2. Let us show that for $\beta_S \neq \{\beta_R, 1-\beta_R\}$ there exists no equilibrium where the set of disclosed signals has 0-measure. Assume by contradiction that this is the case. Then, the perceived disagreement upon non-disclosure is $|\beta_S - \beta_R|$ (due to $\varphi < 1$). Then, since $\Delta(s)$ is single peaked at $\widehat{s}$ by Lemma V.A(i) and $\widetilde{s} \neq \widehat{s}$ by Lemma V.A(ii), we have

$$\Delta(\widehat{s}) > \Delta(\widetilde{s}) = |\beta_S - \beta_R| = \Delta(\varnothing),$$

so that S has an incentive not to disclose all signals sufficiently close to $\widehat{s}$, which is a contradiction.

Lemma V.C *If $\beta_S \neq \{\beta_R, 1-\beta_R\}$, then the unique equilibrium is SDE.*

Proof:

Step 0 Steps 1-2 introduce key equilibrium conditions. In steps 3-4, we show that there exists a unique SDE. Step 5 proves that any equilibrium is an SDE.

In what follows, we assume $\beta_S > \beta_R$. The proof for the reverse case follows the same steps and is omitted.

Step 1 Consider a putative simple disclosure equilibrium. Denote the set of signals by Ψ. Denote the (sub)set of Ψ that is being disclosed by Ψ^d and the complement by $\Psi^\varnothing$. From R's point of view, S does not disclose an observed signal with probability

$$\Pr_R(s \in \Psi^\varnothing) = \beta_R \int_{\Psi^\varnothing} f(s|1)ds + (1-\beta_R) \int_{\Psi^\varnothing} f(s|0)ds.$$

When S does not disclose, R's posterior is

$$\begin{aligned} \widetilde{\beta}_R(\varnothing) &= \frac{\varphi}{(1-\varphi)+\varphi \Pr_R(s \in \Psi^\varnothing)} \int_{\Psi^\varnothing} \left(\beta_R f(s|1) + (1-\beta_R) f(s|0)\right) \widetilde{\beta}_R(s)ds \\ &\quad + \frac{(1-\varphi)}{(1-\varphi)+\varphi \Pr_R(s \in \Psi^\varnothing)} \beta_R \end{aligned}$$

Similarly, R's belief about S's posterior in this case is

$$\begin{aligned} E_R[\widetilde{\beta}_S|\varnothing] &= \frac{\varphi}{(1-\varphi)+\varphi \Pr_R(s \in \Psi^\varnothing)} \int_{\Psi^\varnothing} \left(\beta_R f(s|1) + (1-\beta_R) f(s|0)\right) \widetilde{\beta}_S(s)ds \\ &\quad + \frac{(1-\varphi)}{(1-\varphi)+\varphi \Pr_R(s \in \Psi^\varnothing)} \beta_S. \end{aligned}$$

Step 2 Given the definition of SDE and the fact that $\Delta(s)$ is single peaked, S must be indifferent between disclosure and non-disclosure at s_1 and s_2. Hence, we require

$$\Delta(s) = \Delta(\varnothing, s_1, s_2) \text{ for } s = s_1, s_2. \tag{58}$$

Next, implicitly define $s_2^*(s_1)$ as a value of $s_2 \neq s_1$ equalizing

$$\Delta(s_1) = \Delta(s_2)$$

for given $s_1 < \widehat{s}$, which is unique by Lemma V.A (i). We additionally define $s_2^*(\widehat{s}) = \widehat{s}$. Then, the equilibrium condition (58) holds if and only if

$$\Delta(s_1) = \Delta(\varnothing, s_1, s_2^*(s_1)).$$

Define function

$$\gamma(s_1) \equiv \Delta(\varnothing, s_1, s_2^*(s_1)) - \Delta(s_1)$$

so that SDE exists if and only if

$$\gamma(s_1) = 0.$$

In the next steps, we show that such value of s_1 always exists and is unique.

Step 3. Let us show the existence of SDE, i.e. that there exists s_1 such that $\gamma(s_1) = 0$. Denote s_1' the value of s_1 such that $\Delta(s_1') = \Delta(\widetilde{s}) = \beta_S - \beta_R$. Note that

$$s_1' \neq \widehat{s} \tag{59}$$

since $\widetilde{s} \neq \widehat{s}$ due to $\beta_S \neq \{\beta_R, 1-\beta_R\}$ (by Lemma V.A(ii)). Let us prove that $\gamma(s_1') > 0$. By Step 1 we have

$$\begin{aligned}
\Delta(\varnothing, s_1', s_2^*(s_1')) &= \frac{\varphi}{(1-\varphi) + \varphi \Pr_R(s \in S^{nd})} \int_{s_1'}^{s_2^*(s_1')} (\beta_S(s) - \beta_R(s))\, \widetilde{f}(s) ds \\
&\quad + \left(1 - \frac{\varphi \Pr_R(s \in S^{nd})}{(1-\varphi) + \varphi \Pr_R(s \in S^{nd})}\right) (\beta_S - \beta_R) \\
&= \frac{\varphi}{(1-\varphi) + \varphi \Pr_R(s \in S^{nd})} \int_{s_1'}^{s_2^*(s_1')} (\beta_S(s) - \beta_R(s))\, \widetilde{f}(s) ds \\
&\quad + \left(1 - \frac{\varphi \Pr_R(s \in S^{nd})}{(1-\varphi) + \varphi \Pr_R(s \in S^{nd})}\right) \Delta(s_1') \\
&> \Delta(s_1')
\end{aligned} \tag{60}$$

where the second equality is by construction of s_1', and the strict inequality follows from the fact that $\beta_S(s) - \beta_R(s) > \Delta(s)$ for all $s \in (s_1', s_2^*(s_1'))$ by 59. This implies that $\gamma(s_1') = \Delta(\varnothing, s_1', s_2^*(s_1')) - \Delta(s_1') > 0$.

Now let us show that $\gamma(\widehat{s}) < 0$. Since $\widehat{s} = s_2^*(\widehat{s})$ by construction, it holds that $\Pr_R(s \in S^{nd} | s_1 = s_2 = \widehat{s}) = 0$ so that

$$\Delta(\varnothing, \widehat{s}, s_2^*(\widehat{s})) = \beta_S - \beta_R = \Delta(s_1') < \Delta(\widehat{s}), \tag{61}$$

where the inequality is again due to 59.

Thus, we have shown that $\gamma(s_1') > 0$ and $\gamma(\widehat{s}) < 0$. From continuity of $\gamma(s)$ it then follows that there exists at least one $\underline{s} < s_1 < \widehat{s}$ such that $\gamma(s_1) = 0$.

Step 4. Let us show the uniqueness of SDE. By Step 1 we have

$$\begin{aligned}
\Delta(\varnothing, s_1, s_2^*(s_1)) &= \frac{\varphi}{(1-\varphi)+\varphi \Pr_R(s \in S^{nd})} \int_{s_1}^{s_2^*(s_1)} (\beta_S(s) - \beta_R(s)) \widetilde{f}(s) ds \\
&\quad + \left(1 - \frac{\varphi \Pr_R(s \in S^{nd})}{(1-\varphi)+\varphi \Pr_R(s \in S^{nd})}\right) (\beta_S - \beta_R) \\
&= \frac{\varphi}{(1-\varphi)+\varphi \Pr_R(s \in S^{nd})} \left(\begin{array}{c} \int_{s_1}^{s_2^*(s_1)} (\beta_S(s) - \beta_R(s)) \widetilde{f}(s) ds \\ - \Pr_R(s \in S^{nd}) (\beta_S - \beta_R) \end{array} \right) \\
&\quad + (\beta_S - \beta_R).
\end{aligned}$$

Denote $\eta(s_1) = \frac{\varphi}{(1-\varphi)+\varphi \Pr_R(s \in S^{nd})}$ so that

$$\begin{aligned}
\eta'(s_1) &= -\left(\frac{\varphi}{(1-\varphi)+\varphi \Pr_R(s \in S^{nd})}\right)^2 \left(\frac{\partial \Pr_R(s \in S^{nd})}{\partial s_1} + \frac{\partial \Pr_R(s \in S^{nd})}{\partial s_2^*} \frac{\partial s_2^*}{\partial s_1}\right) \\
&= \eta^2 \left(\widetilde{f}(s_1) - \widetilde{f}(s_2^*) \frac{\partial s_2^*}{\partial s_1}\right).
\end{aligned}$$

Then, taking the derivative of $\Delta(\varnothing, s_1, s_2^*(s_1))$ with respect to s_1 we obtain

$$\begin{aligned}
\frac{\partial \Delta(\varnothing, s_1, s_2^*(s_1))}{\partial s_1} &= \eta'(s_1)\left(\int_{s_1}^{s_2^*}(\beta_S(s)-\beta_R(s))\widetilde{f}(s)ds - \Pr_R(s\in S^{nd})\,(\beta_S-\beta_R)\right) \\
&\quad +\eta(s_1)(-(\beta_S(s_1)-\beta_R(s_1))\widetilde{f}(s_1)+(\beta_S(s_2^*)-\beta_R(s_2^*))\widetilde{f}(s_2^*)\frac{\partial s_2^*}{\partial s_1} \\
&\quad +(\beta_S-\beta_R)\left(\widetilde{f}(s_1)-\widetilde{f}(s_2^*)\frac{\partial s_2^*}{\partial s_1}\right) \\
&= \eta(s_1)^2\left(\widetilde{f}(s_1)-\widetilde{f}(s_2^*)\frac{\partial s_2^*}{\partial s_1}\right)\left(\int_{s_1}^{s_2^*}(\beta_S(s)-\beta_R(s))\widetilde{f}(s)ds - \Pr_R(s\in S^{nd})\,(\beta_S-\beta_R)\right) \\
&\quad -\eta(s_1)((\beta_S(s_1)-\beta_R(s_1))\left(\widetilde{f}(s_1)-\widetilde{f}(s_2^*)\frac{\partial s_2^*}{\partial s_1}\right) \\
&\quad +(\beta_S-\beta_R)\left(\widetilde{f}(s_1)-\widetilde{f}(s_2^*)\frac{\partial s_2^*}{\partial s_1}\right) \\
&= \eta(s_1)\left(\widetilde{f}(s_1)-\widetilde{f}(s_2^*)\frac{\partial s_2^*}{\partial s_1}\right)\begin{pmatrix}\eta(s_1)\int_{s_1}^{s_2^*}(\beta_S(s)-\beta_R(s))\widetilde{f}(s)ds \\ +(1-\eta(s_1)\Pr_R(s\in S^{nd}))\,(\beta_S-\beta_R) \\ -(\beta_S(s_1)-\beta_R(s_1))\end{pmatrix} \\
&= \eta(s_1)\left(\widetilde{f}(s_1)-\widetilde{f}(s_2^*)\frac{\partial s_2^*}{\partial s_1}\right)(\Delta(\varnothing,s_1,s_2^*(s_1))-\Delta(s_1)) \\
&= \eta(s_1)\left(\widetilde{f}(s_1)-\widetilde{f}(s_2^*)\frac{\partial s_2^*}{\partial s_1}\right)\gamma(s_1).
\end{aligned}$$

Thus,

$$\frac{\partial \gamma(s_1)}{\partial s_1} = \eta(s_1)\left(\widetilde{f}(s_1)-\widetilde{f}(s_2)\frac{\partial s_2^*}{\partial s_1}\right)\gamma(s_1)-\Delta'(s_1). \tag{62}$$

Note that $\Delta'(s_1)>0$ and $\frac{\partial s_2^*}{\partial s_1}<0$ for any $s_1<\widehat{s}$ by Lemma V.A (i). Then, (62) implies that for any $s_1<\widehat{s}$ such that $\gamma(s_1)\le 0$ it holds $\gamma'(s_1)<0$. Consequently, if $\gamma(s')=0$ for some $s'<\widehat{s}$, it is strictly decreasing for all $s_1\in[s',\widehat{s})$. Hence, the equilibrium condition $\gamma(s_1)=0$ can be satisfied for at most one value of $s_1<\widehat{s}$.

Step 5 We prove by contradiction that any equilibrium is a simple disclosure equilibrium. Assume thus an equilibrium which is not an SDE. By Lemma V.B, the set of non-disclosed signals has a positive measure. Upon non-disclosure, let the perceived disagreement be denoted by $C>0$. Conditional on obtaining a signal, S wants to disclose if and only if the resulting disagreement $\Delta(s)$ is smaller than C. Recall now that $\Delta(s)$ is

single peaked at $\widehat{s}$ by Lemma V.A(i). Hence, given that a positive measure of signals is not disclosed, we must have $C < \Delta(\widehat{s})$. Then, there are s_1, s_2 satisfying $\underline{s} < s_1 < s_2 < \overline{s}$ such that the actual disagreement is strictly higher than C after disclosing $s \in (s_1, s_2)$ and strictly lower than C after disclosing $s < s_1$ and $s > s_2$. In other words, this implies that for any putative equilibrium, there are s_1, s_2 satisfying $\underline{s} < s_1 < s_2 < \overline{s}$ such that S would strictly prefer not to disclose for $\sigma \in (s_1, s_2)$ and strictly prefer to disclose if $\sigma < s_1$ and $\sigma > s_2$. A putative equilibrium which is not an SDE thus gives rise to strict deviation incentives for S. ■

Lemma V.D

a) Assume that $\beta_S > \beta_R$. If $\beta_R < 1 - \beta_S$, i.e. R is more extreme than S, then any equilibrium features $s_1 < s_2 < \widetilde{s}$, i.e. all signals congruent with R′s prior bias are disclosed. If $\beta_R > 1 - \beta_S$, i.e. R is less extreme than S, then any equilibrium features $\widetilde{s} < s_1 < s_2$, i.e. all signals congruent with S′s prior bias are disclosed.

b) Assume that $\beta_S < \beta_R$. If $\beta_R > 1 - \beta_S$, i.e. R is more extreme than S, then any equilibrium features $\widetilde{s} < s_1 < s_2$, i.e. all signals congruent with R′s prior bias are disclosed. If $\beta_R < 1 - \beta_S$, i.e. R is less extreme than S, then any equilibrium features $s_1 < s_2 < \widetilde{s}$, i.e. all signals congruent with S′s prior bias are disclosed.

Proof:

We use the definitions of $\gamma(s_1)$, s_1' and $s_2^*(s_1)$ used in the proof of Lemma V.C (see Steps 2 and 3 there). By (61), $\gamma(\widehat{s}) < 0$. At the same time, by (60) we have that $\gamma(s_1') > 0$. Consequently, by the uniqueness of the SDE

$$s_1' < s_1. \tag{63}$$

Given that $\Delta(s)$ is single-peaked and the definition of s_2^*, this further implies

$$s_2 = s_2^*(s_1) < s_2^*(s_1'). \tag{64}$$

Now note that by construction $s_1' = \widetilde{s}$ if $\widetilde{s} < \widehat{s}$, and $s_2^*(s_1') = \widetilde{s}$ if $\widetilde{s} > \widehat{s}$. Then, the claims a) and b) follow by Lemma V.A(ii) together with (63) and (64). ■

Lemma V.E *a) If φ increases and $\beta_R \neq 1 - \beta_S$, the equilibrium becomes more Blackwell informative.*

Proof:

Step 0 We focus on the case of $\beta_S > \beta_R$. The proof of the reverse case is identical.

Step 2 Fix s_1 and s_2 at their unique equilibrium values s_1^φ and s_2^φ for given φ. Note that by Step 1 of Lemma V.C

$$\begin{aligned}
&\Delta(\varnothing, s_1^\varphi, s_2^\varphi) = E_R[\widetilde{\beta}_S|\varnothing, s_1^\varphi, s_2^\varphi] - \widetilde{\beta}_R(\varnothing, s_1^\varphi, s_2^\varphi) \\
= \quad & \frac{\varphi \int_{s_1^\varphi}^{s_2^\varphi} (\beta_R f(s|1) + (1-\beta_R) f(s|0)) \left(\widetilde{\beta}_S(s) - \widetilde{\beta}_R(s)\right) ds + (1-\varphi)(\beta_S - \beta_R)}{(1-\varphi) + \varphi \Pr_R(s \in \Psi^\varnothing)},
\end{aligned}$$

and note that the latter expression is trivially always positive given the assumption that $\beta_S > \beta_R$. Letting

$$A = \int_{s_1^\varphi}^{s_2^\varphi} (\beta_R f(s|1) + (1-\beta_R) f(s|0)) \left(\widetilde{\beta}_S(s) - \widetilde{\beta}_R(s)\right) ds$$

and $\delta = (\beta_S - \beta_R)$, it follows that :

$$\frac{\partial \Delta(\varnothing, s_1^\varphi, s_2^\varphi)}{\partial \varphi} = \frac{\begin{array}{c}(A-\delta)\left[(1-\varphi) + \varphi \Pr_R(s \in \Psi^\varnothing)\right] \\ -\left[\varphi A + (1-\varphi)\delta\right]\left[-1 + \Pr_R(s \in \Psi^\varnothing)\right]\end{array}}{\left[(1-\varphi) + \varphi \Pr_R(s \in \Psi^\varnothing)\right]^2}.$$

The above expression simplifies as follows:

$$\begin{aligned}
& \frac{\begin{array}{c}(A-\delta)\left[(1-\varphi) + \varphi \Pr_R(s \in \Psi^\varnothing)\right] \\ -\left[\varphi A + (1-\varphi)\delta\right]\left[-1 + \Pr_R(s \in \Psi^\varnothing)\right]\end{array}}{\left[(1-\varphi) + \varphi \Pr_R(s \in \Psi^\varnothing)\right]^2} \\
= \quad & \frac{\begin{array}{c}A(1-\varphi) + A\varphi \Pr_R(s \in \Psi^\varnothing) - \delta(1-\varphi) - \delta\varphi \Pr_R(s \in \Psi^\varnothing) \\ +\varphi A - \varphi A \Pr_R(s \in \Psi^\varnothing) + (1-\varphi)\delta - (1-\varphi)\delta \Pr_R(s \in \Psi^\varnothing)\end{array}}{\left[(1-\varphi) + \varphi \Pr_R(s \in \Psi^\varnothing)\right]^2} \\
= \quad & \frac{A - \Pr_R(s \in \Psi^\varnothing)\delta}{\left[(1-\varphi) + \varphi \Pr_R(s \in \Psi^\varnothing)\right]^2} \\
= \quad & \frac{\int_{s_1^\varphi}^{s_2^\varphi} (\beta_R f(s|1) + (1-\beta_R) f(s|0)) \left(\widetilde{\beta}_S(s) - \widetilde{\beta}_R(s)\right) ds - \Pr_R(s \in \Psi^\varnothing)(\beta_S - \beta_R)}{\frac{1}{\Pr_R(s \in \Psi^\varnothing)}\left[(1-\varphi) + \varphi \Pr_R(s \in \Psi^\varnothing)\right]^2} > 0
\end{aligned}$$

To see that the last inequality holds, note that

$$\begin{aligned} & \frac{1}{\Pr_R(s \in \Psi^{\varnothing})} \int_{s_1^{\varphi}}^{s_2^{\varphi}} \left(\beta_R f(s|1) + (1-\beta_R) f(s|0)\right) \left(\widetilde{\beta}_S(s) - \widetilde{\beta}_R(s)\right) ds \\ = \; & E[\Delta(s) | s_1^{\varphi} \leq s \leq s_2^{\varphi}] \\ > \; & \Delta(\widetilde{s}) = \beta_S - \beta_R, \end{aligned}$$

where the inequality is due to S having a strict incentive to disclose $\widetilde{s}$ in equilibrium by Lemma V.D.

Step 3 Note that given our previous step, assuming $\varphi' > \varphi$ it holds true that

$$\Delta(\varnothing, s_1^{\varphi}, s_2^{\varphi}, \varphi') > \Delta(\varnothing, s_1^{\varphi}, s_2^{\varphi}, \varphi).$$

Using the notation from Step 3 of the proof of Lemma V.C, we obtain

$$\gamma(s_1^{\varphi}, \varphi') = \Delta(\varnothing, s_1^{\varphi}, s_2^{\varphi}, \varphi') - \Delta(s_1^{\varphi}) > \Delta(\varnothing, s_1^{\varphi}, s_2^{\varphi}, \varphi) - \Delta(s_1^{\varphi}) = 0, \tag{65}$$

where the last equality is by the equilibrium condition. Given that also $\gamma(\widehat{s}, \varphi') < 0$ by (61), $\gamma(s_1^{\varphi}, \varphi') > 0$ yields that for the unique equilibrium level of $s_1^{\varphi'}$ rendering $\gamma(s_1^{\varphi'}, \varphi') = 0$ (since the unique equilibrium is SDE by Lemma V.B) it holds $s_1^{\varphi'} > s_1^{\varphi}$. Given that $\Delta(s)$ is single peaked, it follows that $s_2^{\varphi'} < s_2^{\varphi}$ so that $\left(s_1^{\varphi'}, s_2^{\varphi'}\right) \subset \left(s_1^{\varphi}, s_2^{\varphi}\right)$. This implies that $\left\{s_1^{\varphi'}, s_2^{\varphi'}\right\}$ is more Blackwell informative than $\left\{s_1^{\varphi}, s_2^{\varphi}\right\}$. ■

5.4 Appendix VII: Hidden cost of PC with continuous signals

5.4.1 Proof of Proposition 8

Step 0 We prove Point 1 in what follows. By assumption, it holds true that $\widetilde{s} < s_1 < s_2$. By Lemmas V.B and V.C it follows that $\beta_S \neq 1 - \beta_R$ and $\beta_R > 1 - \beta_S$. We focus on proving that S would strictly prefer to commit to full disclosure if $\beta_S > \beta_R$ so that $\beta_R \in (1 - \beta_S, \beta_S)$. The proof that S instead prefers equilibrium disclosure given $\beta_S < \beta_R$ and $\widetilde{s} < s_1 < s_2$ is briefly outlined in our final step. The proof of Point 2 is conceptually identical to that of Point 1 and thus entirely omitted.

Step 1 From S's perspective, the ex ante perceived disagreement in the SDE featuring thresholds $\{s_1, s_2\}$ is given by:

$$\begin{aligned}
&(1-\varphi)\left[E_R[\widetilde{\beta}_S|\varnothing] - \widetilde{\beta}_R(\varnothing)\right] \\
&+\varphi\int_{s_1}^{s_2}\left(\beta_S f(s|1) + (1-\beta_S) f(s|0)\right)\left[E_R[\widetilde{\beta}_S|\varnothing] - \widetilde{\beta}_R(\varnothing)\right] ds \\
&+\varphi\int_{s\notin\Psi^{\varnothing}}\left(\beta_S f(s|1) + (1-\beta_S) f(s|0)\right)\left[\widetilde{\beta}_S(s) - \widetilde{\beta}_R(s)\right] ds.
\end{aligned}$$

Recall also that we know from Step 1 in the proof of Lemma V.C that

$$\begin{aligned}
& E_R[\widetilde{\beta}_S|\varnothing] - \widetilde{\beta}_R(\varnothing) \\
= \; & \frac{\varphi}{(1-\varphi)+\varphi\Pr_R(s\in\Psi^{\varnothing})}\int_{s_1}^{s_2}\left(\beta_R f(s|1) + (1-\beta_R) f(s|0)\right)\left(\widetilde{\beta}_S(s) - \widetilde{\beta}_R(s)\right) ds \\
& +\frac{(1-\varphi)}{(1-\varphi)+\varphi\Pr_R(s\in\Psi^{\varnothing})}\left(\beta_S - \beta_R\right).
\end{aligned}$$

Step 2 We here consider a putative full disclosure equilibrium. From S's perspective, the ex ante perceived disagreement in an equilibrium with full disclosure is simply

$$\begin{aligned}
&\varphi\int_{s_1}^{s_2}\left(\beta_S f(s|1) + (1-\beta_S) f(s|0)\right)\left[\widetilde{\beta}_S(s) - \widetilde{\beta}_R(s)\right] ds \\
&+\varphi\int_{s\notin\Psi^{\varnothing}}\left(\beta_S f(s|1) + (1-\beta_S) f(s|0)\right)\left[\widetilde{\beta}_S(s) - \widetilde{\beta}_R(s)\right] ds \\
&+(1-\varphi)\left[\beta_S - \beta_R\right].
\end{aligned}$$

Step 3 We introduce two expressions which we shall call Θ(Partial) and Θ(Full). These describe the expected perceived disagreement in S's eyes under each of the two disclosure rules, when restricting ourselves to those events where either $s \in [s_1, s_2]$ or S holds no signal (as otherwise the perceived disagreement is identical under the two regimes). We have:

$$\begin{aligned}
\Theta(\text{Partial}) &= \left[\varphi\Pr_S(s\in\Psi^{\varnothing}) + (1-\varphi)\right]\left[E_R[\widetilde{\beta}_S|\varnothing] - \widetilde{\beta}_R(\varnothing)\right] \\
&= \left[\varphi\Pr_S(s\in\Psi^{\varnothing}) + (1-\varphi)\right] \\
&\quad\times\left[\begin{array}{c}\frac{\varphi}{(1-\varphi)+\varphi\Pr_R(s\in\Psi^{\varnothing})}\int_{s_1}^{s_2}\left(\beta_R f(s|1) + (1-\beta_R) f(s|0)\right)\left(\widetilde{\beta}_S(s) - \widetilde{\beta}_R(s)\right) ds \\ +\frac{(1-\varphi)}{(1-\varphi)+\varphi\Pr_R(s\in\Psi^{\varnothing})}\left(\beta_S - \beta_R\right)\end{array}\right]
\end{aligned}$$

and

$$\Theta(\text{Full}) = \varphi \int_{s_1}^{s_2} \left[\beta_S f(s|1) + (1-\beta_S) f(s|0)\right] \left[\widetilde{\beta}_S(s) - \widetilde{\beta}_R(s)\right] ds + (1-\varphi)(\beta_S - \beta_R).$$

Our objective is to identify conditions under which $\Theta(\text{Partial}) > \Theta(\text{Full})$, i.e.

$$\left[\varphi \Pr_S(s \in \Psi^{\varnothing}) + (1-\varphi)\right] \left[E_R[\widetilde{\beta}_S|\varnothing] - \widetilde{\beta}_R(\varnothing)\right] \tag{66}$$

$$> \quad \varphi \int_{s_1}^{s_2} \left[\beta_S f(s|1) + (1-\beta_S) f(s|0)\right] \left[\widetilde{\beta}_S(s) - \widetilde{\beta}_R(s)\right] ds + (1-\varphi)(\beta_S - \beta_R). \tag{67}$$

Step 4 Define $\Pr_{\widehat{\beta}_R}(s \in \Psi^{\varnothing})$ as the ex ante probability that $s \in [s_1, s_2]$, given the prior $\widehat{\beta}_R$. I.e. define:

$$\Pr_{\widehat{\beta}_R}(s \in \Psi^{\varnothing}) = \int_{s_1}^{s_2} (\widehat{\beta}_R f(s|1) + \left(1 - \widehat{\beta}_R\right) f(s|0)) ds.$$

We define $\Delta\left(\beta_S, \beta_R, \widehat{\beta}_R\right)$ as a slightly modified version of $E_R[\widetilde{\beta}_S|\varnothing] - \widetilde{\beta}_R(\varnothing)$, with the only difference that the expected distribution of signals is calculated based on the prior $\widehat{\beta}_R$. We let

$$\begin{aligned} & \Delta\left(\beta_S, \beta_R, \widehat{\beta}_R\right) \\ = \quad & \frac{\varphi}{(1-\varphi) + \varphi \Pr_{\widehat{\beta}_R}(s \in \Psi^{\varnothing})} \int_{s_1}^{s_2} (\widehat{\beta}_R f(s|1) + \left(1 - \widehat{\beta}_R\right) f(s|0)) \left(\widetilde{\beta}_S(s) - \widetilde{\beta}_R(s)\right) ds \\ & + \frac{(1-\varphi)}{(1-\varphi) + \varphi \Pr_{\widehat{\beta}_R}(s \in \Psi^{\varnothing})} (\beta_S - \beta_R). \end{aligned}$$

Let us finally define

$$\widehat{\Theta}(\text{Partial}, \widehat{\beta}_R) = \left[\varphi \Pr_S(s \in \Psi^{\varnothing}) + (1-\varphi)\right] \left[\Delta\left(\beta_S, \beta_R, \widehat{\beta}_R\right)\right]$$

and note that $\widehat{\Theta}(\text{Partial}, \beta_R) = \Theta(\text{Partial})$.

In what follows, we shall consider the value of the above function for $\widehat{\beta}_R = \beta_S$ and for $\widehat{\beta}_R \in (1-\beta_S, \beta_S)$. We show in step 5 that $\widehat{\Theta}(\text{Partial}, \beta_S) = \Theta(\text{Full})$. We show in step 6 that for any $\widehat{\beta}_R \in (1-\beta_S, \beta_S)$ $\widehat{\Theta}(\text{Partial}, \widehat{\beta}_R) > \Theta(\text{Full})$. Given that by assumption $\beta_R \in (1-\beta_S, \beta_S)$, this implies that in particular $\widehat{\Theta}(\text{Partial}, \beta_R) = \Theta(\text{Partial}) > \Theta(\text{Full})$.

Step 5 Note that when setting $\widehat{\beta}_R = \beta_S$, we have:

$$\begin{aligned}
& \widehat{\Theta}(\text{Partial}, \beta_S) && (68)\\
= \; & \left[\varphi \Pr_S(s \in \Psi^{\varnothing}) + (1-\varphi)\right] \left[\Delta\left(\beta_S, \beta_R, \beta_S\right)\right] && (69)\\
= \; & \left[\varphi \Pr_S(s \in \Psi^{\varnothing}) + (1-\varphi)\right] && (70)\\
& \times \left[\begin{array}{c} \frac{\varphi}{(1-\varphi)+\varphi \Pr_S(s\in\Psi^{\varnothing})} \int_{s_1}^{s_2} \left(\beta_S f(s|1) + (1-\beta_S) f(s|0)\right) \left(\widetilde{\beta}_S(s) - \widetilde{\beta}_R(s)\right) ds \\ + \frac{(1-\varphi)}{(1-\varphi)+\varphi \Pr_S(s\in\Psi^{\varnothing})} \left(\beta_S - \beta_R\right) \end{array}\right] && (71)\\
= \; & \varphi \int_{s_1}^{s_2} \left[\beta_S f(s|1) + (1-\beta_S) f(s|0)\right] \left[\widetilde{\beta}_S(s) - \widetilde{\beta}_R(s)\right] ds + (1-\varphi)\left(\beta_S - \beta_R\right) && (72)\\
= \; & \Theta(\text{Full}). && (73)
\end{aligned}$$

Step 6 Here, we show that $\Delta\left(\beta_S, \beta_R, \widehat{\beta}_R\right)$ increases (resp. decreases) as $\widehat{\beta}_R$ decreases (resp. increases), for $\widehat{\beta}_R \leq \beta_S$. Note that we can rewrite $\Delta\left(\beta_S, \beta_R, \widehat{\beta}_R\right)$ as follows:

$$\begin{aligned}
& \Delta\left(\beta_S, \beta_R, \widehat{\beta}_R\right) \\
= \; & \left[\begin{array}{c} \frac{\varphi \Pr_{\widehat{\beta}_R}(s\in\Psi^{\varnothing})}{(1-\varphi)+\varphi \Pr_{\widehat{\beta}_R}(s\in\Psi^{\varnothing})} \int_{s_1}^{s_2} \frac{\left(\widehat{\beta}_R f(s|1) + \left(1-\widehat{\beta}_R\right) f(s|0)\right)}{\Pr_{\widehat{\beta}_R}(s\in\Psi^{\varnothing})} \left(\widetilde{\beta}_S(s) - \widetilde{\beta}_R(s)\right) ds \\ + \frac{(1-\varphi)}{(1-\varphi)+\varphi \Pr_{\widehat{\beta}_R}(s\in\Psi^{\varnothing})} \left(\beta_S - \beta_R\right) \end{array}\right].
\end{aligned}$$

From the above expression, note that $\Delta\left(\beta_S, \beta_R, \widehat{\beta}_R\right)$ is thus a weighted average of the expressions

$$\begin{aligned}
& E_{\widehat{\beta}_R}\left[\widetilde{\beta}_S(s) - \widetilde{\beta}_R(s) \,|s \in [s_1, s_2]\right] \\
= \; & \int_{s_1}^{s_2} \frac{\left(\widehat{\beta}_R f(s|1) + \left(1-\widehat{\beta}_R\right) f(s|0)\right)}{\Pr_{\widehat{\beta}_R}(s \in \Psi^{\varnothing})} \left(\widetilde{\beta}_S(s) - \widetilde{\beta}_R(s)\right) ds
\end{aligned}$$

and $(\beta_S - \beta_R)$. The first expression is weighted by $\frac{\varphi \Pr_{\widehat{\beta}_R}(s\in\Psi^{\varnothing})}{(1-\varphi)+\varphi \Pr_{\widehat{\beta}_R}(s\in\Psi^{\varnothing})}$ and the second is weighted by $\frac{(1-\varphi)}{(1-\varphi)+\varphi \Pr_{\widehat{\beta}_R}(s\in\Psi^{\varnothing})}$. In other words, $\Delta\left(\beta_S, \beta_R, \widehat{\beta}_R\right)$ can be written as:

$$\Delta\left(\beta_S, \beta_R, \widehat{\beta}_R\right) = p(\widehat{\beta}_R) A(\widehat{\beta}_R) + (1 - p(\widehat{\beta}_R))(\beta_S - \beta_R),$$

where we let

$$p(\widehat{\beta}_R) = \frac{\varphi \Pr_{\widehat{\beta}_R}(s \in \Psi^{\varnothing})}{(1-\varphi) + \varphi \Pr_{\widehat{\beta}_R}(s \in \Psi^{\varnothing})}$$

and we let

$$A(\widehat{\beta}_R) = E_{\widehat{\beta}_R}\left[\widetilde{\beta}_S(s) - \widetilde{\beta}_R(s)\,|s \in [s_1, s_2]\right].$$

The derivative of $\Delta\left(\beta_S, \beta_R, \widehat{\beta}_R\right)$ w.r.t. $\widehat{\beta}_R$ is thus given by

$$\begin{aligned}\frac{\partial \Delta\left(\beta_S, \beta_R, \widehat{\beta}_R\right)}{\partial \widehat{\beta}_R} &= \frac{\partial p(\widehat{\beta}_R)}{\partial \widehat{\beta}_R} A(\widehat{\beta}_R) + p(\widehat{\beta}_R)\frac{\partial A(\widehat{\beta}_R)}{\partial \widehat{\beta}_R} - \frac{\partial p(\widehat{\beta}_R)}{\partial \widehat{\beta}_R}(\beta_S - \beta_R) \\ &= p(\widehat{\beta}_R)\frac{\partial A(\widehat{\beta}_R)}{\partial \widehat{\beta}_R} + \frac{\partial p(\widehat{\beta}_R)}{\partial \widehat{\beta}_R}\left[A(\widehat{\beta}_R) - (\beta_S - \beta_R)\right].\end{aligned}$$

In order to prove that $\frac{\partial \Delta\left(\beta_S, \beta_R, \widehat{\beta}_R\right)}{\partial \widehat{\beta}_R} < 0$, it thus suffices to show that $\frac{\partial A(\widehat{\beta}_R)}{\partial \widehat{\beta}_R} < 0$,

$$\left[A(\widehat{\beta}_R) - (\beta_S - \beta_R)\right] > 0$$

and $\frac{\partial p(\widehat{\beta}_R)}{\partial \widehat{\beta}_R} < 0$. We show in what follows that these properties are indeed satisfied for $\widehat{\beta}_R \in (1 - \beta_S, \beta_S]$.

Note first that $\frac{\partial \Pr_{\widehat{\beta}_R}(s \in \Psi^{\varnothing})}{\partial \widehat{\beta}_R} = \int_{s_1}^{s_2}(f(s|1) - f(s|0))ds$, which is strictly negative given that we know that $f(s|0) > f(s|1)$ for any $s \in [s_1, s_2]$, recalling that $\widetilde{s} < s_1 < s_2$ by assumption. It follows immediately that $\frac{(1-\varphi)}{(1-\varphi)+\varphi \Pr_{\widehat{\beta}_R}(s \in \Psi^{\varnothing})} = 1 - p(\widehat{\beta}_R)$ increases in $\widehat{\beta}_R$ and that $\frac{\varphi \Pr_{\widehat{\beta}_R}(s \in \Psi^{\varnothing})}{(1-\varphi)+\varphi \Pr_{\widehat{\beta}_R}(s \in \Psi^{\varnothing})} = p(\widehat{\beta}_R)$ decreases in $\widehat{\beta}_R$. Second, to show that $A(\widehat{\beta}_R) - (\beta_S - \beta_R) > 0$ note that by the fact that $\widetilde{s} < s_1$, $\Delta(s_1) = \Delta(s_2)$ in SDE and the hump shape of $\Delta(s)$ we obtain

$$\beta_S - \beta_R = \Delta(\widetilde{s}) < \Delta(s_1) < \Delta(s) | s \in (s_1, s_2).$$

Third, we now show that $A(\widehat{\beta}_R) = E_{\widehat{\beta}_R}\left[\left(\widetilde{\beta}_S(s) - \widetilde{\beta}_R(s)\right)|s \in [s_1, s_2]\right]$ decreases as $\widehat{\beta}_R$

increases. Note that:

$$
\begin{aligned}
&\frac{\partial\left[\int_{s_1}^{s_2}\frac{\left(\widehat{\beta}_R f(s|1)+\left(1-\widehat{\beta}_R\right)f(s|0)\right)}{\Pr_{\widehat{\beta}_R}(s\in\Psi^{\varnothing})}\left(\widetilde{\beta}_S(s)-\widetilde{\beta}_R(s)\right)ds\right]}{\partial\widehat{\beta}_R}\\
&= \int_{s_1}^{s_2}\frac{\begin{pmatrix}\left(f(s|1)-f(s|0)\right)\left[\int_{s_1}^{s_2}\widehat{\beta}_R f(s|1)+\left(1-\widehat{\beta}_R\right)f(s|0)ds\right]\\ -\left[\widehat{\beta}_R f(s|1)+\left(1-\widehat{\beta}_R\right)f(s|0)\right]\left[\int_{s_1}^{s_2}(f(s|1)-f(s|0))ds\right]\end{pmatrix}}{\left[\Pr_{\widehat{\beta}_R}(s\in\Psi^{\varnothing})\right]^2}\left(\widetilde{\beta}_S(s)-\widetilde{\beta}_R(s)\right)ds\\
&= \frac{\begin{pmatrix}\left[\int_{s_1}^{s_2}\widehat{\beta}_R f(s|1)+\left(1-\widehat{\beta}_R\right)f(s|0)ds\right]\left[\int_{s_1}^{s_2}(f(s|1)-f(s|0))\left(\widetilde{\beta}_S(s)-\widetilde{\beta}_R(s)\right)ds\right]\\ -\left[\int_{s_1}^{s_2}(f(s|1)-f(s|0))ds\right]\left[\int_{s_1}^{s_2}\left(\widehat{\beta}_R f(s|1)+\left(1-\widehat{\beta}_R\right)f(s|0)\right)\left(\widetilde{\beta}_S(s)-\widetilde{\beta}_R(s)\right)ds\right]\end{pmatrix}}{\left[\Pr_{\widehat{\beta}_R}(s\in\Psi^{\varnothing})\right]^2}\\
&= \frac{\begin{pmatrix}-\left[\int_{s_1}^{s_2}(f(s|1)-f(s|0))ds\right]\left[\int_{s_1}^{s_2}\left(\widehat{\beta}_R f(s|1)+\left(1-\widehat{\beta}_R\right)f(s|0)\right)\left(\widetilde{\beta}_S(s)-\widetilde{\beta}_R(s)\right)ds\right]\\ +\left[\int_{s_1}^{s_2}\widehat{\beta}_R f(s|1)+\left(1-\widehat{\beta}_R\right)f(s|0)ds\right]\left[\int_{s_1}^{s_2}(f(s|1)-f(s|0))\left(\widetilde{\beta}_S(s)-\widetilde{\beta}_R(s)\right)ds\right]\end{pmatrix}}{\left[\Pr_{\widehat{\beta}_R}(s\in\Psi^{\varnothing})\right]^2}\\
&< \frac{\begin{pmatrix}-\left[\int_{s_1}^{s_2}(f(s|1)-f(s|0))ds\right]\left[\int_{s_1}^{s_2}\left(\widehat{\beta}_R f(s|1)+\left(1-\widehat{\beta}_R\right)f(s|0)\right)\left(\widetilde{\beta}_S(s)-\widetilde{\beta}_R(s)\right)ds\right]\\ +\left[\int_{s_1}^{s_2}\widehat{\beta}_R f(s|1)+\left(1-\widehat{\beta}_R\right)f(s|0)ds\right]\left[\int_{s_1}^{s_2}(f(s|1)-f(s|0))\right]\left[\int_{s_1}^{s_2}\left(\widetilde{\beta}_S(s)-\widetilde{\beta}_R(s)\right)ds\right]\end{pmatrix}}{\left[\Pr_{\widehat{\beta}_R}(s\in\Psi^{\varnothing})\right]^2}\\
&= \frac{-\left[\int_{s_1}^{s_2}(f(s|1)-f(s|0))ds\right]\begin{pmatrix}\left[\int_{s_1}^{s_2}\left(\widehat{\beta}_R f(s|1)+\left(1-\widehat{\beta}_R\right)f(s|0)\right)\left(\widetilde{\beta}_S(s)-\widetilde{\beta}_R(s)\right)ds\right]\\ -\left[\int_{s_1}^{s_2}\widehat{\beta}_R f(s|1)+\left(1-\widehat{\beta}_R\right)f(s|0)ds\right]\left[\int_{s_1}^{s_2}\left(\widetilde{\beta}_S(s)-\widetilde{\beta}_R(s)\right)ds\right]\end{pmatrix}}{\left[\Pr_{\widehat{\beta}_R}(s\in\Psi^{\varnothing})\right]^2}\\
&< 0.
\end{aligned}
$$

Above, the first equality follows from the application of Leibniz' rule. The first and the second inequality follow from applying Hölder's inequality.

Thus, we have shown that $\frac{\partial\Delta\left(\beta_S,\beta_R,\widehat{\beta}_R\right)}{\partial\widehat{\beta}_R}<0$. This implies that

$$\frac{\partial\widehat{\Theta}(\text{Partial},\widehat{\beta}_R)}{\partial\widehat{\beta}_R}<0.$$

In sum, we obtain that for $\beta_S > \beta_R$ and $\widetilde{s} < s_1 < s_2$ it holds

$$\Theta(\text{Partial}) = \widehat{\Theta}(\text{Partial},\beta_R) > \widehat{\Theta}(\text{Partial},\beta_S) = \Theta(\text{Full}). \tag{74}$$

Here, the inequality follows from the previous inequality, while the second equality is by Step 5.

Step 7 Suppose now instead that $\beta_S < \beta_R$ and $\widetilde{s} < s_1 < s_2$. Note that combining the assumptions $\beta_S < \beta_R$ and $\widetilde{s} < s_1 < s_2$ implies that $\beta_R \in (\beta_S, 1)$ by Lemma V.D. The argument follows the same logic as above. It still holds true $\widehat{\Theta}(\text{Partial},\beta_S) = \Theta(\text{Full})$ and that $\widehat{\Theta}(\text{Partial},\beta_R) = \Theta(\text{Partial})$. It also still holds true that $\Theta(\text{Partial},\widehat{\beta}_R)$ is decreasing in $\widehat{\beta}_R$. It follows that

$$\Theta(\text{Partial}) = \widehat{\Theta}(\text{Partial},\beta_R) < \widehat{\Theta}(\text{Partial},\beta_S) = \Theta(\text{Full}).$$

5.4.2 Proof of Proposition 9

The argument here is exactly identical to the proof of the counterpart of this result for the case of binary signals (Proposition 3).

5.5 Appendix VIII: Instrumental disagreement aversion.

Proof of Proposition ??.

Step 1. Consider the disclosure choice of the agent in a putative FD-equilibrium if she holds a signal $\sigma \in \{0, 1\}$. Let:

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

where $\Pi^{FD}(\varnothing)$ is the value of Π after no disclosure in a putative FD-equilibrium. Recall that S prefers a higher $\Pi(d)$, since he maximized the likelihood of being hired. Hence, in an FD-equilibrium, S has no strict incentive to deviate when holding a σ-signal if and

only if $\Delta_\sigma(\beta_S, \beta_R) \geq 0$ for $\sigma \in \{0, 1\}$. We have

$$\begin{aligned}
\Delta_1(\beta_S, \beta_R) &= \Pi(1) - \Pi^{FD}(\varnothing) \\
&= \widetilde{\beta}_R(0)\widetilde{\beta}_S(0) + (1 - \widetilde{\beta}_R(0))(1 - \widetilde{\beta}_S(0)) \\
&\quad -\beta_R\beta_S - (1 - \beta_R)(1 - \beta_S) \\
&= \frac{(\beta_S + \beta_R - 2\beta_R\beta_S)(2p - 1)}{(1 - p + \beta_R(2p - 1))(1 - p + \beta_S(2p - 1))} \\
&\quad \times((\beta_R + \beta_S - 1)(1 - p) + \beta_R\beta_S(2p - 1)).
\end{aligned}$$

It is easy to show that the fraction on the right-hand side is always positive. Hence,

$$\begin{aligned}
\Delta_1(\beta_S, \beta_R) &\geq 0 \\
&\Leftrightarrow (\beta_R + \beta_S - 1)(1 - p) + \beta_R\beta_S(2p - 1) \geq 0 \\
&\Leftrightarrow \beta_S \geq \frac{(1 - p)(1 - \beta_R)}{1 - p + \beta_R(2p - 1)}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\Delta_0(\beta_S, \beta_R) &= \Pi(0) - \Pi^{FD}(\varnothing) \\
&= \widetilde{\beta}_R(1)\widetilde{\beta}_S(1) + (1 - \widetilde{\beta}_R(1))(1 - \widetilde{\beta}_S(1)) \\
&\quad -\beta_R\beta_S - (1 - \beta_R)(1 - \beta_S) \\
&= \frac{(\beta_S + \beta_R - 2\beta_R\beta_S)(2p - 1)}{(p - \beta_R(2p - 1))(p - \beta_S(2p - 1))} \\
&\quad \times(p(1 - \beta_R - \beta_S) + \beta_R\beta_S(2p - 1)).
\end{aligned}$$

Hence,

$$\begin{aligned}
\Delta_0(\beta_S, \beta_R) &\geq 0 \\
&\Leftrightarrow (p(1 - \beta_R - \beta_S) + \beta_R\beta_S(2p - 1)) \geq 0 \\
&\Leftrightarrow \beta_S \leq \frac{p(1 - \beta_R)}{\beta_R + p(1 - 2\beta_R)}.
\end{aligned}$$

Thus, $\Delta_0(\beta_S, \beta_R)$ and $\Delta_1(p, \beta_S, \beta_R)$ 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.

Step 2. Consider conditions when S find it optimal to conceal 1-signals and disclose 0-signals at the first stage. For this we need to have

$$\begin{aligned}\Pi(0) &\leq \Pi^{D0}(\varnothing),\\ \Pi(1) &\geq \Pi^{D0}(\varnothing),\end{aligned}$$

where $\Pi^{D0}(\varnothing)$ is the value of Π after no disclosure in a putative equilibrium where only 0-signals are disclosed. Note that

$$\Pi^{D0}(\varnothing) = \Pr[\sigma = 1|\varnothing]\Pi(0) + (1 - \Pr[\sigma = 1|\varnothing])\Pi^{FD}(\varnothing).$$

Consequently,

$$\Pi(0) \leq \Pi^{D0}(\varnothing) \text{ iff } \Pi(0) \leq \Pi^{FD}(\varnothing).$$

Hence, by Step 1,

$$\Pi(0) \leq \Pi^{D0}(\varnothing) \text{ iff } \beta_S \geq \frac{p(1-\beta_R)}{\beta_R + p(1-2\beta_R)}.$$

Next, $\Pi(0) \leq \Pi^{D0}(\varnothing)$ immediately implies the remaining condition $\Pi(1) > \Pi^{D0}(\varnothing)$. Indeed,

$$\frac{p(1-\beta_R)}{\beta_R + p(1-2\beta_R)} > \frac{(1-p)(1-\beta_R)}{1-p+\beta_R(2p-1)}.$$

Consequently,

$$\begin{aligned}\beta_S &\geq \frac{p(1-\beta_R)}{\beta_R + p(1-2\beta_R)}\\ &\Rightarrow \beta_S > \frac{(1-p)(1-\beta_R)}{1-p+\beta_R(2p-1)}\\ &\Leftrightarrow \Pi(1) > \Pi^{FD}(\varnothing),\end{aligned}$$

where the last step is by Step 1. Thus, D0 exists if and only if $\beta_S \geq \frac{p(1-\beta_R)}{\beta_R+p(1-2\beta_R)}$, which is equivalent to the corresponding condition in Proposition 1.

Step 3. The proof for D1-equilibrium proceeds analogously to Step 2.

5.6 Appendix IX: Joint observation of public signals

5.6.1 Proof of Proposition 15

Step 1 This proves Point 1 of the proposition. Assume without loss of generality that $x \geq y$. Given the definitions in section 2.4.1 we have

$$\begin{aligned}\Lambda^x(x,y,p) &= P(\sigma=1\,|x)D_0(x,y,p) + P(\sigma=0\,|x)D_1(x,y,p)\\ &= (xp+(1-x)(1-p))\left(\frac{xp}{xp+(1-x)(1-p)} - \frac{yp}{yp+(1-y)(1-p)}\right)\\ &\quad + (1-(xp+(1-x)(1-p)))\left(\frac{x(1-p)}{x(1-p)+(1-x)p} - \frac{y(1-p)}{y(1-p)+(1-y)p}\right).\end{aligned}$$

At the same time,

$$\Lambda^x(x,y,\frac{1}{2}) = x-y.$$

The difference $V^x(x,y,p) = \Lambda^x(x,y,\frac{1}{2}) - \Lambda^x(x,y,p)$ further simplifies to

$$V^x(x,y,p) = \frac{(1-2p)^2(x-y)(1-y)y}{(y+p-2py)(2py+1-(p+y))}. \tag{75}$$

It can be trivially shown that this expression is always positive no matter the values of x, y and p, where V^x equals to 0 if and only if $y \in \{0, x, 1\}$, which proves Point 1 of the proposition.

Step 2 Let us show that the derivative of $V^x(x,y,p)$ with respect to y is convex in y if $x > y$ and concave in y if $y > x$. Consider $x > y$. Taking the third derivative of $V^x(x,y,p)$ and simplifying we obtain

$$\frac{\partial^3 V^x(x,y,p)}{\partial y^3} = \frac{6(1-2p)^2(1-p)p}{(1-y-p(1-2y))^4(y+p(1-2y))^4}M, \tag{76}$$

where

$$\begin{aligned}M &= y^4(1-2p)^4 - 4y^3x(1-2p)^4 + p - 4p^2 + 6p^3 - 3p^4 + 6y^2(x(1-2p)^4\\ &\quad +(1-2p)^2(1-p)p) + x(1-2p)^2(1-2p+2p^2)\\ &\quad -4y(1-2p)^2(x+p-3xp-p^2+3xp^2).\end{aligned}$$

Let us show that $M > 0$. Note that M is linear in x. Hence, to prove that M as a function of x is positive on $(y,1)$ it is sufficient to show that it is positive at the boundaries of this interval. We have that at $x = y$

$$\begin{aligned} M_{|x=y} &= 6y^3(1-2p)^4 - 3y^4(1-2p)^4 + p - 4p^2 + 6p^3 - 3p^4 \\ &\quad + y(1-2p)^2(1-6p+6p^2) - 2y^2(1-2p)^2(2-9p+9p^2). \end{aligned}$$

One can verify that this function of y has no roots on [0,1]. Besides at $y = 0$ this expression turns to $p(1-4p) + 6p^3(1-0.5p) > 0$. Hence,

$$M_{|x=y} > 0. \tag{77}$$

Next,

$$\begin{aligned} M_{|x=1} &= 1 - 4y^3(1-2p)^4 + y^4(1-2p)^4 - 5p + 10p^2 - 10p^3 + 5p^4 \\ &\quad - 4y(1-2p)^2(1-2p+2p^2) + 6y^2(1-2p)^2(1-3p+3p^2). \end{aligned}$$

One can verify that this function of y has no roots on [0,1]. Besides at $y = 0$ this expression turns to $1 - 5p(1-2p) - 10p^3(1-0.5p) > 0$. Hence,

$$M_{|x=1} > 0.$$

This together with (77) and the fact that M is linear in x implies that $M > 0$. Consequently, by (76)

$$\frac{\partial^3 V^x(x,y,p)}{\partial y^3} > 0,$$

i.e., the derivative of V^x with respect to y is convex in y if $x > y$.

The claim that the derivative of V^x with respect to y is concave in y if $y > x$ follows analogously.

Step 3 Now we can prove Point 2 of the proposition. From Step 1 and the continuity of $V^x(x,y,p)$ in y it follows that

$$\begin{aligned} \frac{\partial V^x(x,y,p)}{\partial y}_{|y=0} &> 0, \\ \frac{\partial V^x(x,y,p)}{\partial y}_{|y\to x^-} &< 0, \end{aligned}$$

Since further $\frac{\partial V^x(x,y,p)}{\partial y}$ is convex in y by Step 2, it follows that it has a unique root on $(0,x)$. This implies that $V^x(x,y,p)$ is single-peaked in y for $y \in [0,x]$. The claim for $x < y$ follows analogously given that

$$\begin{aligned} \frac{\partial V^x(x,y,p)}{\partial y}_{|y=x^+} &> 0, \\ \frac{\partial V^x(x,y,p)}{\partial y}_{|y=1^-} &< 0 \end{aligned}$$

by Step 1, and $\frac{\partial V^x(x,y,p)}{\partial y}$ is concave in y for $x < y$ by Step 2.

Step 4 Now we can prove Point 3 of the proposition. Let us show that for $x < 1/2$ the maximum of $V^x(x,y,p)$ is reached for $y > 1/2$ (the reverse argument then immediately follows by symmetry considerations). First, note that for $x = 1/2$ we should have that the left and the right peaks (see Step 3) yield the same value of $V^x(x,y,p)$ by symmetry considerations. Next, we have that

$$y > x : \frac{\partial V^x(x,y,p)}{\partial x} = \frac{(1-y)y(1-2p)^2}{(y-1+p-2yp)(y+p-2yp)} < 0, \tag{78}$$

$$y < x : \frac{\partial V^x(x,y,p)}{\partial x} = -\frac{(1-y)y(1-2p)^2}{(y-1+p-2yp)(y+p-2yp)} > 0, \tag{79}$$

This implies that as x decreases, $\max_y V^x(x,y,p|y>x)$ is increasing and $\max_y V^x(x,y,p|y<x)$ is decreasing. Hence, overall $\max_y V^x(x,y,p)$ is reached at $\widehat{y} > x$. To show that $\widehat{y} > 1/2$ we use the fact that

$$\begin{aligned} & \frac{\partial V^x(x,y,p)}{\partial y}_{|y=1/2} \\ = & \frac{4(1-2p)^2(-\frac{3}{16}(1-2p)^2 - x(1-p)p + (1+x)(1-p)p + \frac{1}{4}(1-7(1-p)p))}{(1-p+\frac{1}{2}(2p-1))^2} \\ > & 0. \end{aligned}$$

Hence, the right peak (maximizing $V^x(x,y,p)$) is reached to the right of $y = 1/2$. ■

5.6.2 Proof of Proposition 16

We further denote

$$\mu(x,y) = \min\{V^x(x,y), V^y(x,y)\}.$$

Given that both players should agree to participate, the probability of signal acquisition is maximized if and only if $\mu(x,y)$ is maximized.

Step 1 Note that $\mu(x,y)$ should reach its maximum at some values $\{x^*,y^*\}$ where $x^*,y^* \neq \{0,1\}$ since $\min\{V^x(x,y),V^y(x,y)\} = 0$ if either x or y are at the boundaries while there exists some $\{x,y\}$ where $\mu(x,y) > 0$ (by Proposition 10.1).

Step 2 By (78) and (79) we have that $V^x(x,y)$ is linearly increasing (decreasing) in x for $x > y$ ($x < y$). Analogously, $V^y(x,y)$ is linearly increasing (decreasing) in y for $y > x$ ($y < x$).

Step 3 Let us show that $\mu(x,y)$ must reach its maximum at some $\{x^*,y^*\}$ where $V^x(x^*,y^*) = V^y(x^*,y^*)$. Assume by contradiction that this is not the case so that for instance, $V^x(x^*,y^*) < V^y(x^*,y^*)$. Then, by Steps 1 and 2 one can slightly change x to raise the value of $V^x(x,y)$ so that $\mu(x,y) = V^x(x,y) < V^y(x,y)$ continues to hold. In other words, one can raise $\mu(x,y)$ at least by a slight perturbation of x which proves that $\{x^*,y^*\}$ is not the optimum. The symmetric argument excludes $V^x(x^*,y^*) > V^y(x^*,y^*)$.

Step 4 We have shown that $V^x(x^*,y^*) = V^y(x^*,y^*)$. Given (75) (and the symmetrical expression for $V^y(x,y,p)$), this condition holds if and only if either $x^* = y^*$ or $\omega^0(x^*,y^*) = \omega^1(x^*,y^*)$ (in which case a difference in the probability weights on $\omega^0(x^*,y^*)$ and $\omega^1(x^*,y^*)$ does not matter). One can in turn verify that the latter condition is true if and only if either $x^* = y^*$ or $x^* = 1 - y^*$. In the first case, we have $\mu(x,x) = 0$ by Proposition 11.1 so it cannot be optimal. Hence, at the optimum it must hold $x^* = 1 - y^*$.

Step 5 Let us finally show that there is unique $x^* \in (0,1/2)$ where $\mu(x^*,1-x^*)$ is maximized (in which case $\mu(x,y)$ is also maximized by Step 4). Let us show that $\mu(x,1-x)$ is concave. Note that by symmetry considerations $V^x(x,1-x) = V^{1-x}(x,1-x)$. Hence,

$$\begin{aligned} & \frac{\partial^2 \mu(x,1-x)}{\partial x^2} \\ = & \frac{\partial^2 V^x(x,1-x)}{\partial x^2} \\ = & 2(1-p)p(2p-1)\left(\frac{1}{(p(1-2x)+x-1)^3} + \frac{1}{(p(1-2x)+x)^3}\right) < 0. \end{aligned}$$

Given that $\mu(x,1-x)$ is concave in x and is equal to 0 at $x = 0$ and $x = 1/2$ by Proposition 10.1, we obtain that there is unique $x^* \in (0,1/2)$ maximizing $\mu(x,1-x)$.