Disliking to disagree

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

April 15, 2018

Abstract

We study disclosure by a disagreement-averse sender facing a receiver with different prior beliefs. With a binary state, full disclosure is feasible only if the receiver's prior is close enough to one minus the sender's prior. If full disclosure is infeasible, only information congruent with the prior bias of the most extreme player is fully disclosed. The active avoidance of perceived disagreement can paradoxically lead to larger (perceived or actual) disagreement in beliefs from an ex ante perspective. Disagreement aversion arises endogenously within simple games of compromise decision making and delegation. Finally, moderate prior heterogeneity encourages public information acquisition in committees featuring disagreement averse players.

Keywords: strategic disclosure, psychological games, disagreement aversion

JEL classification: D81, D83, D91

1 Introduction

Political behavior is heavily influenced by the information available to citizens, which is a key reason why political forces (governments, lobbies, political parties) often devote significant resources to influencing centralized information flows (propaganda, media censorship, campaigns, etc). But citizens also obtain information in a decentralized fashion

*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.

from talking to each other. The latter channel has arguably gained in relative importance in the digital era (internet, social media).

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 not equally likely to talk to everyone. A typical instance of this is the following recommendation from a 19th century gentleman's manual:[1] "*Do not discuss politics or religion in general company. (...) To discuss those topics is to arouse feeling without any good result*." The rule is still relevant today: A 2016 poll by the online employment website *CareerBuilder* finds that 42 percent of respondents avoid talking politics at the office while 44 percent talk about it but interrupt the conversation if it becomes heated[2]. Social-psychologists have developed a wide array of concepts to describe and theorize informational biases in social networks: *taboos, Overton windows, opinion corridors, political correctness, conversational minefields, echo chambers, confirmation bias, collective ignorance, information avoidance*.

Two aspects appear to play an important role in generating these biases, namely people's tendency to avoid open conflict (of opinions) and the fact that people hold different prior beliefs, which lead them to react differently to the same information. This paper proposes to study the consequences of these two stylized facts for key instances of (bayesian) learning among rational agents.

A large body of experimental and empirical evidence documents that in stating their opinions, individuals tend to conform to what they believe others think. In the series of experiments conducted by Asch and related in the seminal paper *Opinions and social pressure* Asch (1955), subjects wrongly evaluated the length of a line after being exposed to other participants' (artificially induced) wrong assessment. Bursztyn et al. (2017a) found that subjects were more likely to reveal immigration-critical views two weeks after Donald Trump's victory than two weeks before it. Prentice and Miller (1993) found that students radically underestimated how many others were uncomfortable with campus alcohol practices because very few dared express dissent. Disagreement aversion has a

[1] "Hills Manual of Social and Business Forms", 1879.

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

variety of potential causes. The aversion may be intrinsic (i.e. a cultural trait) or instrumental (i.e. driven by the anticipation of adverse consequences). One might dislike others to think that one is wrong or fear being disliked by people who disagree with one's views. Disagreement might make others more less likely to cooperate in future tasks. Political cultures in north-western Europe put a special emphasis on reaching consensus in negotiations (e.g. the so-called Polder model of consensual politics in the Netherlands, labour market negotiations in Scandinavian countries). We refer to Golman et al. (2016) for an in-depth discussion of central causes and consequences of what the authors term a preference for *belief consonance*, structured around a distinction between so-called *group-identity* and *protected beliefs* approaches[3].

A second stylized fact is that individuals have heterogeneous priors, i.e. have different distributional beliefs about the environment (the state of the world) and as a consequence do not interpret evidence in the exact same way. Examples include climate change (whether it occurs and how to best address it), animal welfare, immigration, the economic effects of trade, religion. A key underlying driver of the phenomenon is that people have different personal histories (different experiences, socialization, frames).

We examine the implications of the two above stylized facts for key instances of bilateral bayesian learning, namely strategic information disclosure and collective information acquisition. A main source of tension is that while any informative experiment on average reduces disagreement, specific signal realizations can increase disagreement. Key questions addressed are as follows. For which types of information is learning impeded or slowed down by the disagreement aversion? What matches of individuals give rise to the most productive learning? Can a policy of preempting disagreement be counterproductive from an ex ante perspective?

The main section of our paper examines a simple game of disclosure by a party (S) who is averse to perceived disagreement on the part of the uninformed party (R). The state is binary (0 or 1) and S and R have different publicly observed prior beliefs α_S and α_R about the state being 0. A binary informative signal $\sigma \in \{0, 1\}$ of commonly known

[3]See also Golman et al. (2017) for a related discussion of motives for information avoidance.

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 a given signal precision p and a given receiver prior α_R, there is a most favorable sender prior $\alpha_S^* = 1 - \alpha_R$ such that full disclosure is feasible only if S's prior is close enough to α_S^*. Fourth, better information quality is always helpful: The higher the quality of information, the larger the set of values of α_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, which constitutes a potential starting point for a theory of echo-chambers. The second part of our main section takes an ex ante perspective on equilibrium and disagreement. We show that the drive to avoid perceived disagreement can backfire from an ex ante perspective, thereby revealing a hidden cost of political correctness. In the eyes of S, (ex ante) expected perceived disagreement can be higher in equilibrium than it would be under full disclosure, so that S would prefer to commit to full disclosure. Second, in the eyes of a third party with a prior potentially different from S's and R's, the (ex ante) expected *actual* disagreement can be higher in equilibrium than it would be under full disclosure. The final part of our main section establishes that uncertainty about priors can help disclosure.

The subsequent *extensions* section of the paper examines how our findings extend in a variety of fundamental directions. The first subsection tests the technical robustness of our results to a more general information structure. We show that the main qualitative features of our characterization survive given continuous signals satisfying the MLRP property. The second subsection examines potential micro-foundations of disagreement aversion, by embedding the disclosure stage within two-stage games in which the second stage involves either collective decision making or a delegation decision. The last subsection considers a game of costly collective acquisition of public signals by parties exhibiting disagreement aversion. The game represents conversations as a collective endeavor whose (informational) outcome is inherently unpredictable. Though the game is strategically very different from our disclosure game, it addresses the same underlying problem of learning in (prior beliefs-wise) heterogeneous groups. We find that moderate

prior misalignment optimally induces information acquisition (i.e. stimulates conversation), in a way that echoes our main findings.

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. It 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.[4]. Zanardo (2017) characterizes the set of belief disagreement functions that satisfy a set of desirable axioms. Kartik and Zanardo (2016) identify necessary and sufficient conditions under which public information reduces disagreement. Kartik et al. (2015) consider agents with different priors and show that two agents with different priors each believe that a (Blackwell) more informative public experiment will, in expectation, bring the other's posterior closer to his own prior.

An extensive body of research dating back to (Crawford and Sobel, 1982; Milgrom, 1981) studies strategic information transmission between an informed sender (S) and a receiver (R) (see Sobel, 2013, for a review). These models typically involve a difference in 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, which features privately known priors, only experts with extreme priors disclose and information has on average a moderating effect on R. In the second, the authors identify cases where competition promotes revelation. Che and Kartik (2009) examines the effect of prior 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 the above papers, S simply wants R's first order beliefs to be close to some state

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

dependent or independent bliss-point. In our paper, S also 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. As a consequence, S might for example conceal a signal that brings R's first order beliefs closer to hers, if this minimizes perceived disagreement .

A strand of the literature on strategic information transmission features an endogenous preference for belief conformity arising from reputational concerns. In (Ottaviani and Sørensen, 2006a,b; Gentzkow and Shapiro, 2006), S wishes to signal a high quality of her information to R, who ultimately observes the actual state. This leads S to bias her message towards R's prior belief, which hampers informativeness. [5] Similarly, in our setup if S's prior is more 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 R to omit signals that confirm R's prior if S's prior is more extreme than R's.

Finally, our paper 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). 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). A related model of Ely et al. (2015) considers the behavior of a principal who wishes the beliefs of an agent to follow a specific time-path exhibiting suspense or surprises.

We proceed as follows. Section 2 considers the baseline model, Section 3 provides

[5] Sobel (1985) and Morris (2001) study related sender-receiver games with an endogenous reputational concern of the sender for being perceived as unbiased, which also leads to distorted informativeness of communication. Prendergast (1993) in a principal-agent setting examines the agent's incentive to match the (noisy) information of the principal in his report. Bursztyn et al. (2017b) consider a setting where a sender has to communicate his type to a receiver and has 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. Kajackaite and Gneezy (2015); Khalmetski and Sliwka (2017) consider cheap talk with a sender who is interested in minimizing the probability assigned by R to her having lied. They find that this implies an incentive to conform to what R expects to hear.

modeling extensions, and Section 4 concludes. All proofs, unless explicitly stated otherwise, are relegated to Technical Appendixes $I - VI$.

2 Main analysis

2.1 The disclosure game

There are two agents S and R and a state $\omega \in \{0,1\}$. Player $i \in S,R$ assigns prior probability α_i to $\omega = 0$. Priors are common knowledge. S holds with probability $\varphi \in (0,1)$ a privately observed signal $\sigma \in \{0,1\}$. The signal is identical to the state with probability p, i.e. $P(\sigma = \omega) = p \ \forall \omega$. Player S can disclose the signal to R or not. Denote S's disclosure by d, where $d \in \{0,1,\emptyset\}$. R simply observes S's signal if disclosed and subsequently updates beliefs. S is averse to perceived disagreement on the part of R, i.e. wants to minimize R's ex post perception of disagreement. Let $\widetilde{\alpha}_i$ denote i's posterior given information. Denote by $\widetilde{\alpha}_R(d)$ the posterior probability assigned by R to state 0 given that S discloses d. Denote by $E_R[\widetilde{\alpha}_S \,|d]$ the expected value of S's posterior given d, in the eyes of R. Clearly, both $\widetilde{\alpha}_R(d)$ and $E_R[\widetilde{\alpha}_S \,|d]$ are functions of R's belief about S's disclosure strategy. S's utility function is given as follows:

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

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. Note that we could have assumed instead that S minimizes $E_R\left[\left|\widetilde{\alpha}_S - \widetilde{\alpha}_R\right| \middle| d\right]$. The idea of our assumption is that S only cares about not being perceived as clearly biased in one direction relative to R.

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.

Three informative pure disclosure strategies are respectively full disclosure (called FD), disclosure of only 0-signals or only 1-signals (called D0 or D1). An equilibrium featuring disclosure strategy $X \in \{FD, D0, D1\}$ is called an X-equilibrium. An equilibrium featuring an informative disclosure strategy is called informative. If $\alpha_i > (<)\frac{1}{2}$, we say that i's prior is *biased towards* state $0(1)$. If $\alpha_i > \frac{1}{2}$, a 0-signal is *congruent with* i's prior bias and a 1-signal *contradicts* it (vice versa if $\alpha_i < \frac{1}{2}$). If α_i is strictly closer to either 0 or 1 than α_j, then i is said to hold a *stronger* or *more extreme* prior than j.

2.2 Equilibrium characterization

Define the following functions:

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

The above functions have the following properties, which are formally established in the proof of our next Proposition. First, given $\alpha_R \in (0,1)$ and $p \in \left(\frac{1}{2}, 1\right)$, it always holds true that $0 < \alpha_S^*(\alpha_R, p) < \alpha_S^{**}(\alpha_R, p) < 1$. Second, α_S^* and α_S^{**} are continuous in p. Third, α_S^* is decreasing in p and α_S^{**} is increasing in p. Fourth, $\alpha_S^*(\alpha_R, \frac{1}{2}) = \alpha_S^{**}(\alpha_R, \frac{1}{2}) = 1 - \alpha_R$. Finally, $\alpha_S^*(\alpha_R, 1) = 0$ and $\alpha_S^{**}(\alpha_R, 1) = 1$. The following Proposition provides a characterization of the set of informative equilibria.

Proposition 1 *1. If* $\alpha_S \in \{\alpha_R, 1-\alpha_R\}$ *or if* $\alpha_S \notin \{\alpha_R, 1-\alpha_R\}$ *and* $\alpha_S = \left\{\alpha_S^*(\alpha_R, p), \alpha_S^{**}(\alpha_R, p)\right\}$, *the FD equilibrium exists.*

2. Given $\alpha_S \notin \{\alpha_R, 1-\alpha_R\}$:

a) If $\alpha_S \in (0, \alpha_S^*(\alpha_R, p))$ *then the unique equilibrium is D1,*

b) If $\alpha_S \in \left(\alpha_S^*(\alpha_R, p), \alpha_S^{**}(\alpha_R, p)\right)$ *then the unique equilibrium is FD,*

c) If $\alpha_S \in (\alpha_S^{**}(\alpha_R, p), 1)$ *then the unique equilibrium is D0.*

Figure 1 below provides an illustration of our characterization for $\alpha_R = .3$. The dashed curves correspond to $\alpha_S^*(.3, p)$ and $\alpha_S^{**}(.3, p)$. Between the two thick curves, the FD equilibrium equilibrium exists. Instead, above (below) of the upward (downward) sloping

thick curve, only the D0 (D1) equilibrium exists. Finally, for $\alpha_S = \alpha_R$, an FD equilibrium exists 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.

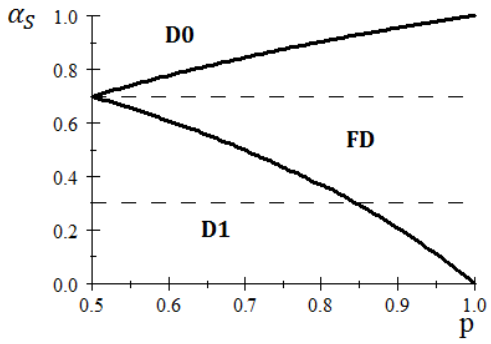

Figure 1: Equilibrium characterization.

Our characterization exhibits five key properties. First, note that except under knife-edge conditions, our statement guarantees a unique equilibrium. Uniqueness is attractive for comparative statics purposes. The second key feature is that full disclosure is not always feasible. This is surprising at first, as information on average reduces disagreement in beliefs by moving everyone's beliefs towards the truth. The answer lies in the fact that updating has two dimensions: The direction of updating of the prior and the size of the implied shift in beliefs. 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 heavily discounts a signal contradicting her prior bias (she considers it wrong with high probability). The difference in updating-intensities can be large enough to make posteriors more different than priors, for a given signal.

The third key feature is that more prior misalignment (if not too extreme) can generate more disclosure, the (disclosure-) optimal sender prior being $1 - \alpha_R$. The optimal sender's prior can thus be very different from R's prior but is always exactly as strong as R's. R and R's optimal sparring partner are thus both different and similar, depending on the dimension considered. Note that two aspects vary as a function of prior misalignment. The first is the attractiveness of the status quo. The second is how differently players

update on the basis of the same signal. As the difference in priors varies, the relative strength of these two effects changes. Let p be intermediate, and consider the following three cases, all featuring a putative FD equilibrium. Given small prior disagreement, the status quo is very attractive and different priors imply slightly different updating. As a consequence, one of the two signals increases disagreement. Given moderate prior disagreement, the status quo is now quite unattractive but priors are still close enough that updating is fairly similar. It follows that both signals decrease disagreement. Finally, given large prior disagreement, the status quo is very unattractive but very different priors imply very different updating, so that one of the two signals increases disagreement. To understand that the optimal sender prior is specifically $1 - \alpha_R$, consider the following argument. For any α_S, α_R, p, at least one type of signal (either 0 or 1) leads to a decrease in perceived (and actual) disagreement. This must be true, as we know from the work of Kartik et al. that a hypothetical full disclosure equilibrium reduces perceived disagreement in expectation. Next, note that if $\alpha_S = 1 - \alpha_R$, the effect of a 1-signal disclosure on disagreement is equivalent to the effect of a 0-disclosure. I.e.,

$$-\left|E_R[\widetilde{\alpha}_S \,|0] - \widetilde{\alpha}_R(0)\right| = -\left|E_R[\widetilde{\alpha}_S \,|1] - \widetilde{\alpha}_R(1)\right|$$

This is intuitive since priors are completely symmetric around $\frac{1}{2}$. This implies that if a disclosure of one type of signal leads to a reduction of disagreement relative to the status quo (which must be true by our first observation), then the disclosure of the other type of signal should achieve this as well. Hence, full disclosure is achievable under any p for $\alpha_S = 1 - \alpha_R$.

The fourth key feature is that sufficiently conclusive information allows for full disclosure. The intuition follows from considering the limit case of $p = 1$, in which any signal trivially reduces disagreement. It follows that for p sufficiently high, this is also true. 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.

The fifth key feature is that 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 strongest. The signal generating the largest disagreement is indeed the one that contradicts the prior

bias of the most extreme player. For an intuition, 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 (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 moderate closer to (away from) the extremist's prior.

Within a simple random matching setup, the above fifth key feature naturally leads to the implication that the more R's prior is biased towards the wrong state, the less likely she is to be exposed to the truth. Assume for example that R, whose prior α_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 (i.e. one that is congruent with the true state), the more biased she is towards the wrong state. Assume for example that $\omega = 1$. Then by Proposition 1 the ex-ante probability of R being exposed to a 1-signal (denoting by σ the signal obtained by S) is

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

which is strictly decreasing in α_R. So the more R is biased towards 0, the less likely she is to observe a 1-signal.

Within a dynamic version of the above random matching scenario where R repeatedly encounters senders over many periods, partial disclosure thus leads to very slow learning (i.e. inertia in beliefs) if the state is not congruent with R's extreme prior bias. Note that R is however not naive. In a D0 or D1 equilibrium, no disclosure is interpreted as S potentially holding information R's bias and thus does give rise to a shift in beliefs. A second observation is that the standard mechanism behind confirmation bias is arguably reversed here. Confirmation bias is often assumed to be driven by selective information search. Here, it is instead caused by the information suppliers.

2.3 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 perspective in terms of perceived disagreement and then from a third party perspective in terms of actual disagreement. Note that the ex ante evaluation of expected disagreement in equilibrium requires specifying two aspects: The prior used to weight different possible signal realizations and the applied measure of ex post disagreement, either perceived or actual.

First, from S's perspective, can the (ex ante) expected perceived disagreement be higher in (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. D0 is the unique equilibrium. Ex ante, S would strictly prefer to commit to full disclosure if $\alpha_S > \alpha_R$. If $\alpha_S < \alpha_R$, she instead ex ante strictly prefers the D0 equilibrium.*

2. Let parameters be s.t. D1 is the unique equilibrium. Ex ante, S would strictly prefer to commit to full disclosure if $\alpha_S < \alpha_R$. If $\alpha_S > \alpha_R$, she instead ex ante strictly prefers the D1 equilibrium.

S would thus ex ante prefer to commit to full disclosure if she is the most extreme player. The intuition is as follows. In a partial disclosure equilibrium (e.g. D1), the omission of 0-signals has two countervailing effects. The upside is that S benefits from hiding a 0-signal once she holds it. The downside is that when S holds no signal, R interprets silence as a possible concealment of a 0-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. Recall that in the latter case, S omits signals contradicting her bias in a partial disclosure equilibrium. The key idea is that 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 in a partial disclosure equilibrium. Instead, on the equilibrium path of the full disclosure equilibrium, R's prior does not affect her ex post perception of S's posterior.

A second key question is whether from the perspective of a third party endowed with a prior $\bar{\alpha}$, the (ex ante) expected *actual* disagreement can be higher in equilibrium than it would be under FD. I.e., would such a third party want to impose full disclosure if aiming at minimizing 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 $|\widetilde{\alpha}_S(\sigma) - \widetilde{\alpha}_R(d)|$ (where $\widetilde{\alpha}_R(d)$ is a function of R's belief about S's disclosure strategy). In what follows, if either $\alpha_i = \widehat{\alpha}$ and $\alpha_j \neq \widehat{\alpha}$ or instead $\alpha_i < \widehat{\alpha} < \alpha_j$, we say that S and R's priors are weakly on different sides of $\widehat{\alpha}$. Otherwise, they are said to be on the same side of $\widehat{\alpha}$.

Proposition 3 *Let parameters be s.t. there exists no FD equilibrium. In the eyes of a third party with prior $\widehat{\alpha} \neq \alpha_R$, the expected actual disagreement:*

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

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

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

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

a) S and R's priors are on the same side of $\widehat{\alpha}$,

b) S's prior is further away from $\widehat{\alpha}$ than R's prior and it is sufficiently extreme.

Part 1 finds that equilibrium concealment can indeed be counterproductive while part 2 instead identifies conditions under which it is helpful. A general intuition behind our results is that the third party (TP) expects new information to lead S and R*'s* beliefs to converge to her prior. The disclosure strategy 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 and R's priors are on different sides of $\widehat{\alpha}$. Here, given that S and R's beliefs move closer to $\widehat{\alpha}$, they must also be moving closer to each other. Hence TP would prefer that they learn as fast as possible and would thus prefer FD over partial revelation. The second case is that α_S and α_R on the same side of $\widehat{\alpha}$, but R is more extreme. An

instance of this is the case of $\frac{1}{2} \leq \widehat{\alpha} < \alpha_S < \alpha_R$. Again TP expects S and R to converge to her prior $\widehat{\alpha}$, 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 α_S and α_R on the same side of $\widehat{\alpha}$, but R is more extreme. An instance of this is the case of $\frac{1}{2} \leq \widehat{\alpha} < \alpha_R < \alpha_S$. Here, both players' belief decreases. But decreasing R's belief moves it away from S's. So TP would prefer to slow down R's learning and thus prefers partial disclosure.

2.4 Strangers' talk

Conversations often take place between parties who do not know each others' priors. Is such uncertainty beneficial for disclosure? The following Proposition provides an answer to the question for two cases.

Proposition 4 *a) If both priors are private knowledge and drawn from the same symmetric distribution on* $[0, 1]$*, there exists a full disclosure equilibrium.*

b) If S's prior is commonly known and sufficiently close to $\frac{1}{2}$ *while that of R is drawn from a symmetric distribution over* $[0, 1]$*, there exists a full disclosure equilibrium.*

Point a) shows that two-sided uncertainty about priors is beneficial to disclosure if the prior distribution is symmetric. This provides an argument for not encouraging revelation of information about respective biases. Point b) shows that two-sided uncertainty is not strictly necessary to ensure full revelation. The latter is compatible with S's prior being known, if S is approximately unbiased and R's prior is symmetrically distributed.

While uncertainty about prior bias is 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. We here study an extended version of our disclosure game featuring communication about priors and explicit preferences of R. Assume that S's prior is known and equal to $\frac{1}{2}$ while R's prior is privately observed. Assume that S minimizes perceived disagreement whereas R wants to learn the state, her utility function being

given by $-(a-\omega)^2$, where $a \in [0,1]$ is the action chosen by R after S's disclosure. 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. Note that given α_S and p, there are three sets of values of α_R giving rise to respectively the D0, D1 and FD equilibrium. We call *essentially truthful* an equilibrium in which R truthfully reveals to which of these three set α_R belongs. Our next Proposition presents a negative result concerning such equilibria.

Proposition 5 *Consider the extended disclosure game. Suppose that the set of possible priors of R contains two priors α_R and α_R' that imply different equilibria in the one-shot disclosure game. There exists no equilibrium featuring essentially truthful communication by R.*

Note first that 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 trigger respectively D0 and D1. The key is that R's preference over partially revealing experiments reverses her prior bias as follows: Given $\alpha_R < \frac{1}{2}$, R strictly prefers D0-communication over D1-communication (and vice versa given $\alpha_R > \frac{1}{2}$). Note furthermore that given $\alpha_S = \frac{1}{2}$, 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 lie about her bias.

3 Extensions

The first two subsections of this section examine the robustness and micro-foundability of the setup examined in the main section. The third subsection considers a game of costly collective acquisition of public signals by parties exhibiting disagreement aversion.

3.1 Continuous signals

We now show that our characterization carries over qualitatively to the case of a signal structure with continuous signals satisfying the marginal likelihood ratio property (MLRP). Assume that S's signal is drawn from an interval $[\underline{s}, \overline{s}]$. Given state $\omega \in \{0, 1\}$, the signal s received by 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 extreme signals $\underline{s}$ and $\overline{s}$ are perfectly revealing, i.e., $\frac{f(s|h)}{f(s|l)} = 0$ for $s = \underline{s}$ and $= \infty$ for $s = \overline{s}$. Upon learning s, the updated belief of i is

$$\widetilde{\alpha}_i(s) = \frac{\alpha_i f(s|l)}{\alpha_i f(s|l) + (1 - \alpha_i) f(s|h)} = \frac{\alpha_i}{\alpha_i + (1 - \alpha_i)\frac{f(s|h)}{f(s|l)}},$$

which is decreasing in s. Note that there exists a threshold signal $\tilde{s} \in (\underline{s}, \overline{s})$ such that whatever $\alpha_i \in (0, 1)$, it holds true that $\widetilde{\alpha}_i(s) \lesseqgtr \alpha_i$ for $s \gtreqless \tilde{s}$. Signal $\tilde{s}$ satisfies $f(\tilde{s}|h) = f(\tilde{s}|l)$ 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 $\alpha_j < (>)\frac{1}{2}$. We call the above setup the *continuous signals environment*. We call *simple disclosure equilibrium* (SDE) an equilibrium featuring two thresholds $\underline{s} < s_1 \leq s_2 < \overline{s}$ such that S discloses s if and only if $s \leq s_1$ or $s \geq s_2$. We obtain the following equilibrium characterization.

Proposition 6 *Assume the continuous signals environment:*

1. There always exists an SDE and any equilibrium is an SDE.

2. If $\alpha_S \in \{\alpha_R, 1 - \alpha_R\}$ there is a unique SDE featuring $s_1 = s_2$, i.e. full disclosure. If $\alpha_S \notin \{\alpha_R, 1 - \alpha_R\}$, then any SDE features $s_1 < s_2$. Furthermore, all signals congruent with the bias of the player with the most extreme prior are disclosed.

3. If there exist multiple equilibria, then they are ordered in terms of Blackwell informativeness. When φ increases, the most Blackwell informative equilibrium becomes strictly more Blackwell informative.

The fundamental qualitative features of equilibrium echo those arising under binary signals. Only signals that are congruent with the prior of the most extreme player are fully

revealed. Furthermore, with $\alpha_S = 1 - \alpha_R$, 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 1. S would strictly prefer to commit to full disclosure ex ante if $\alpha_S > \alpha_R$. If $\alpha_S < \alpha_R$, 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 0. S would strictly prefer to commit to full disclosure ex ante if $\alpha_S < \alpha_R$. If $\alpha_S > \alpha_R$, she ex ante strictly prefers any equilibrium over full disclosure.

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

3.2 Instrumental disagreement aversion

S's aversion to perceived disagreement might stem from the fact that it adversely affects subsequent interaction with R. We here consider two-stage games in which S may disclose her private information in stage 1 while in stage 2, R makes a decision which is payoff-relevant to both S and R and which depends on R's first- and second-order beliefs. We consider two setups matching this description in what follows. In both games considered, stage 1 essentially coincides with the binary disclosure problem considered in our main setup.

3.2.1 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 β_i.

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 R is not hired, the problem remains untackled and R simply obtains a payoff of 0. If S is hired, the contract proposed by S specifies a reward of 1 if the agent tackles the problem successfully (this outcome being observable). By hiring S, R incurs a privately observed and random (transaction) cost c, which is drawn from a uniform distribution on $[0,1]$. Let $I(k)$ be an indicator function, where $k = 1(0)$ indicates success (failure), $I(1) = 1$ and $I(0) = 0$. Conditional on S being hired and outcome k, the payoff of R is $I(k)\tau - c$. The agent 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. Conditional on hiring, efforts e_0 and e_1 and outcome k, the payoff obtained by R 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 disclosure strategy FD 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 backwards 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 $\sigma \in \{0,1,\emptyset\}$, in a putative full disclosure (FD) equilibrium, where $\emptyset$ 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 thus rewrites as:

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

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 good.

We now consider the principal's hiring decision after observing the disclosure $d \in \{0, 1, \emptyset\}$. If she decides to hire, the principal obtains an expected payoff of $\tau\Pi(\widetilde{\beta}_S(d), \widetilde{\beta}_R(d))$, where

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

She thus hires if and only if c is smaller than the above (i.e. iff hiring yields a net benefit). 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 increase in R's subjective expected payoff from hiring occasioned by S disclosing σ in a putative FD equilibrium. Clearly, in a putative full disclosure 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

$$p \geq \max \left\{ \frac{-\beta_S\beta_R}{\beta_S + \beta_R - 2\beta_S\beta_R - 1}, \frac{\beta_S + \beta_R - \beta_S\beta_R - 1}{\beta_S + \beta_R - 2\beta_S\beta_R - 1} \right\}.$$

It can be verified that this condition is equivalent to the one appearing in Proposition 1.

3.2.2 Collective decision-making by compromise

Consider the following simple game of decision making by compromise. 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 $x = \frac{1}{2}(x_S + x_R)$. 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 given by $-(\beta_i - x)^2$, 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 2 below, FD is feasible above the solid black curve. Below (strictly), only either D0 or D1 is feasible. We assume $\alpha_S = .55$. 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. Increasing prior misalignment can thus be locally beneficial to disclosure, which echoes our main characterization.

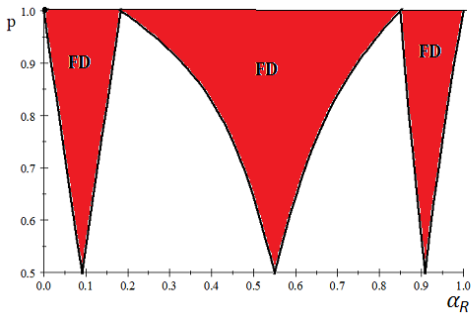

Figure 2: Partial equilibrium characterization.

Our next Proposition provides a partial formal description of the qualitative features of the above figure. In what follows, 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 10 *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.

3.3 Joint observation of public signals

3.3.1 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 observing her cost c_i of participating. If both decide to participate, they observe a randomly drawn public binary signal which is correct with probability p. If at least one of the agents opts against participating, the game ends: No signal is observed and players incur no cost. We call agents x and y, where agent $z \in \{x,y\}$ assigns prior probability z to state 0 and $x > y$. 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 each possible public signal:

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

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

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

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

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

Clearly, player z decides to participate if and only if $c_z \leq V^z(x, y, p)$. We obtain the following characterization of the value of participating for each player.

Proposition 11 *1. For given x, $V^x(x, y) \geq 0$ for any y, while $V^x(x, y) = 0$ if and only if $y \in \{0, x, 1\}$.*

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

3. For given x, $V^x(x, y)$ 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. 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. Some degree of heterogeneity thus 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 12 *There is a unique pair $\{x^*, y^*\}$ maximizing the probability of signal acquisition. For this pair, it holds true that $y^* = 1 - x^*$ and $x^* \notin \{0, 1/2, 1\}$.*

3.3.2 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 very 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. #

We make two simplifying assumptions that embody forms of myopia. First, a player aims only at maximizing the perceived disagreement of 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.

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?

Figure 3, 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 $\omega = 0$. The process always 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.

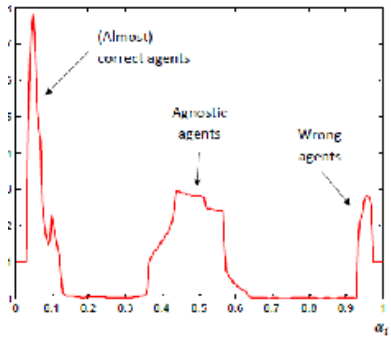

Figure 3: Asymptotic population distribution of beliefs.

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 wrong beliefs. 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.

4 Conclusion

This paper introduces a new type of belief-dependent preferences reflecting an aversion to perceived disagreement. We identify three main implications of disagreement aversion. First, the sender withholds information that contradicts the prior of the party with the most extreme prior belief. Second, larger differences in priors can imply better incentives for disclosure and joint information acquisition. Finally, avoiding perceived disagreement can be counterproductive from ex ante point of view, thereby revealing a hidden cost of "political correctness". Importantly, aversion to perceived disagreement endogenously arises from strategic concerns in a variety of environments. 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, 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.

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*, rds030.

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 (2017a). From extreme to mainstream: How social norms unravel. Technical report, National Bureau of Economic Research.

Bursztyn, L., G. Egorov, and S. Fiorin (2017b). From extreme to mainstream: How social norms unravel. Technical report.

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

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

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

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

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

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.

Golman, R., D. Hagmann, and G. Loewenstein (2017). Information avoidance. *Journal of Economic Literature 55*(1), 96–135.

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.

Kajackaite, A. and U. Gneezy (2015). Lying costs and incentives. Technical report, Mimeo.

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

Kartik, N. and E. Zanardo (2016). When does information reduce disagreement?

Khalmetski, K. and D. Sliwka (2017). Disguising lies-image concerns and partial lying in cheating games. Technical report.

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

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

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*, 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.

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

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.

Zanardo, E. (2017). How to measure disagreement? Technical report, Mimeo., Columbia University.

5 Technical Appendix

5.1 Appendix I: Disclosure with binary signals (Proof of Proposition 1)

Proposition 1 follows from a set of Lemmas, which are stated and proved in what follows. For Lemmas I.B to I.E, we assume $\alpha_R < \frac{1}{2}$, which is wlog given the symmetric nature of the model. For simplicity, we occasionally use the notation $x = \alpha_S$ and $y = \alpha_R$.

Lemma I.A *Let* $D(\sigma) = \left|\widetilde{\alpha}_S(\sigma) - \widetilde{\alpha}_R(\sigma)\right|$, *for* $\sigma \in \{0, 1, \emptyset\}$, *where* $\emptyset$ *stands for "S holds no-signal".*

a) There exists no equilibrium in which S always omits to disclose with probability one.

b) If there exists no signal $\sigma^* \in \{0, 1\}$ *s.t.* $D(\sigma^*) = D(\emptyset)$, *then there exists no equilibrium in which S uses a mixed strategy.*

Proof:

Step 1 This proves Point a). We here prove that there cannot be an equilibrium in which S omits to disclose with positive probability after both signals. Assume that S mixes between no disclosure and disclosure given both $\sigma = 0$ and $\sigma = 1$. Then, R's perceived disagreement after no disclosure (denoted $D(nd)$), satisfies

$$D(nd) = q_1 D(1) + q_2 D(0) + q_3 D(\emptyset),$$

where $q_1, q_2, q_3 \in (0, 1)$ and $q_1 + q_2 + q_3 = 1$. Note that such an equilibrium requires $D(nd) \geq D(0)$ and $D(nd) \geq D(1)$. Note that either $D(1)$ or $D(0)$ is strictly smaller than the two remaining elements in $\{D(0), D(1), D(\emptyset)\}$. It follows that at least one element of $\{D(0), D(1)\}$ is strictly smaller than $D(nd)$, implying a strict deviation incentive.

Step 2 From step 1, we know that mixing between disclosure and non-disclosure cannot occur for both signals and can thus happen at most for one signal $\sigma^* \in \{0, 1\}$. Then, $D(nd) = qD(\sigma^*) + (1 - q)D(\emptyset)$. Such an equilibrium requires the indifference condition $D(\sigma^*) = D(nd)$. It follows that such an equilibrium cannot exist if there is no $\sigma^* \in \{0, 1\}$ s.t. $D(\sigma^*) = D(\emptyset)$. Combining the results proved in steps 1 and 2, Point b) is hereby proved. ■

Lemma I.B *Let* $y < \frac{1}{2}$. *Denote by* $\vartheta_i(x, y, p)$ *the difference between ex ante and ex post perceived disagreement given disclosure of a signal* σ, *for* $\sigma \in \{0, 1\}$.

a) If $x \leq 1-y$, *then* $\vartheta_0(x,y,p) \geq (>)0$ *iff* $p \geq (>)P_0(x,y)$, *where* $P_0(x,y) \in \left[\frac{1}{2},1\right]$. *If instead* $x > 1-y$, *then* $\vartheta_0(x,y,p) > 0$ *for any* $p \geq \frac{1}{2}$.

b) If $x < 1-y$, *then* $\vartheta_1(x,y,p) > 0$ *for any* $p \geq \frac{1}{2}$. *If instead* $x \geq 1-y$, *then* $\vartheta_1(x,y,p) \geq (>)0$ *if and only if* $p \geq (>)P_1(x,y)$, *where* $P_1(x,y) \in \left[\frac{1}{2},1\right]$.

c) $P_0(1-y,y) = P_1(1-y,y) = \frac{1}{2}$ *and* $P_0(0,y) = P_1(1,y) = 1$. *Also,* $\frac{\partial P_0(x,y)}{\partial x} < 0$, $\frac{\partial P_0(x,y)}{\partial y} < 0$, $\frac{\partial P_1(x,y)}{\partial x} > 0$ *and* $\frac{\partial P_1(x,y)}{\partial y} > 0$.

Proof:

Step 0 Assume a putative full-disclosure equilibrium. We prove a sequence of substatements which together yield the above Lemma.

Step 1 This proves a) and part of c). Note that

$$\vartheta_0(x,y,p) = (2p-1)\frac{(x-y)(p+x+y-px-py-xy+2pxy-1)}{(p+x-2px-1)(p+y-2py-1)}.$$

Solving

$$p+x+y-px-py-xy+2pxy-1=0$$

yields the solution

$$p = P_0(x,y) \equiv \frac{x+y-xy-1}{x+y-2xy-1}.$$

Note that

$$\begin{aligned}\frac{\partial P_0(x,y)}{\partial x} &= y\frac{y-1}{(x+y-2xy-1)^2} < 0,\\ \frac{\partial P_0(x,y)}{\partial y} &= x\frac{x-1}{(x+y-2xy-1)^2} < 0.\end{aligned}$$

$P_0(x,y)$ is thus a decreasing function of x. Solving $P_0(x,y) = \frac{1}{2}$ yields $x = 1-y$. In other words, it holds true that $P_0(x,y) < \frac{1}{2}$ if $y > 1-x$ and $P_0(x,y) \geq \frac{1}{2}$ if $y \leq 1-x$. Note also that $P_0(0,y) = \frac{y-1}{y-1} = 1$.

Step 2 This proves b) and part of c). Note that

$$\vartheta_1(x,y,p) = -(2p-1)\frac{(x-y)(px-p+py+xy-2pxy)}{(p+x-2px)(p+y-2py)}.$$

Solving

$$px-p+py+xy-2pxy=0$$

yields the solution:

$$p = P_1(x,y) \equiv -x\frac{y}{x+y-2xy-1}.$$

Note that

$$\begin{aligned}\frac{\partial P_1(x,y)}{\partial x} &= -y\frac{y-1}{(x+y-2xy-1)^2} > 0,\\ \frac{\partial P_1(x,y)}{\partial y} &= -x\frac{x-1}{(x+y-2xy-1)^2} > 0.\end{aligned}$$

$P_1(x,y)$ is thus an increasing function of x. Solving $P_1(x,y) = \frac{1}{2}$ yields $x = 1-y$. In other words, it holds true that $P_1(x,y) \leq \frac{1}{2}$ if $x \leq 1-y$ and $P_1(x,y) > \frac{1}{2}$ if $x > 1-y$. Note also that $P_1(1,y) = -\frac{y}{1+y-2y-1} = 1$. ■

Lemma I.C Let $\alpha_R < \frac{1}{2}$.

a) Given $\alpha_S \in \{\alpha_R, 1-\alpha_R\}$, the FD equilibrium exists.

b) Let $\alpha_S \notin \{\alpha_R, 1-\alpha_R\}$. If $\alpha_S \in \left\{\alpha_S^*(\alpha_R,p), \alpha_S^{**}(\alpha_R,p)\right\}$, then the FD equilibrium exists.

c) Let $\alpha_S \notin \{\alpha_R, 1-\alpha_R\}$. The unique equilibrium is FD if $\alpha_S \in \left(\alpha_S^*(\alpha_R,p), \alpha_S^{**}(\alpha_R,p)\right)$.

Proof:

Step 0 For fixed α_R, $P_0(\alpha_S,\alpha_R)$ and $P_1(\alpha_S,\alpha_R)$ are functions of one variable (α_S) mapping into values of $p \in [0,1]$, and are both trivially invertible. For fixed α_R, define $\alpha_S^*(\alpha_R,p)$ as the inverse function of $P_0(\alpha_S,\alpha_R)$ and $\alpha_S^{**}(\alpha_R,p)$ as the inverse function of $P_1(\alpha_S,\alpha_R)$. For fixed α_R, these map from values of p into values of α_S. We have:

$$\begin{aligned}\alpha_S^*(\alpha_R,p) &\equiv P_0^{-1}(p,\alpha_R) = \frac{(1-\alpha_R)(1-p)}{1-p+\alpha_R(2p-1)},\\ \alpha_S^{**}(\alpha_R,p) &\equiv P_1^{-1}(p,\alpha_R) = \frac{p(1-\alpha_R)}{\alpha_R+p(1-2\alpha_R)}.\end{aligned}$$

Step 1 This proves Point a). If $\alpha_S = \alpha_R$, simply note that ex post perceived disagreement always equals zero, just as the ex ante perceived disagreement. Consider now the case of $\alpha_S = 1-\alpha_R$. Lemma I.B states that for any $p \geq \frac{1}{2}$, in a putative FD equilibrium, S weakly prefers disclosing given any signal.

Step 2 This proves Point b). It follows from Lemma I.B that for any signal, S weakly favours disclosing.

Step 3 This proves Point c). It follows from Lemma I.B that in the putative FD equilibrium, S strictly favors disclosing any signal. Formally, for every $\sigma^* \in \{0,1\}$ we have $D(\sigma^*) < D(\emptyset)$. The FD equilibrium thus exists as S has no deviation incentive. Furthermore, Lemma I.A implies that this is the only equilibrium. ■

Lemma I.D *Let $y \leq \frac{1}{2}$. There exists a D1-equilibrium if and only if $x \leq 1-y$* and $p \leq P_0(x,y)$.

Proof:

Step 0 Assume a putative D1-equilibrium. We prove a sequence of substatements which together yield the above Lemma. Denote by $f_i(x,y,p,\varphi)$ the difference between ex ante and ex post perceived disagreement given disclosure of an i-signal, for $i \in \{0,1\}$.

Step 1 When holding a 0-signal, S should prefer to omit disclosing. Note that:

$$\begin{aligned} & f_0(x,y,p,\varphi) \equiv \\ & \left(\begin{array}{c} \left(\frac{\varphi(yp+(1-y)(1-p))}{\varphi(yp+(1-y)(1-p))+(1-\varphi)}\right)\left(\frac{xp}{xp+(1-x)(1-p)}\right) \\ + \left(\frac{(1-\varphi)}{y\varphi p+(1-y)\varphi(1-p)+(1-\varphi)}\right) x \end{array} \right) \\ & - \left(\frac{y(\varphi p+1-\varphi)}{y(\varphi p+1-\varphi)+(1-y)(\varphi(1-p)+1-\varphi)} \right) \\ & - \left(\frac{xp}{xp+(1-x)(1-p)} - \frac{yp}{yp+(1-y)(1-p)} \right), \end{aligned}$$

which simplifies to

$$(\varphi-1)(2p-1)\frac{(x-y)(p+x+y-px-py-xy+2pxy-1)}{(p+x-2px-1)(p+y-2py-1)(p\varphi+y\varphi-2py\varphi-1)}.$$

Solving for $f_0(x,y,p,\varphi) = 0$, the (unique) solution is given by $p = P_0(x,y)$. We may state that $f_0(x,y,p,\varphi) \leq 0$ if and only $p \leq P_0(x,y)$.

Step 2 After a 1-signal, S should prefer to disclose. Note that:

$$\begin{array}{l} f_1(x,y,p,\varphi) \equiv \\ \left(\begin{array}{c} \left(\frac{\varphi(yp+(1-y)(1-p))}{\varphi(yp+(1-y)(1-p))+(1-\varphi)}\right)\left(\frac{xp}{xp+(1-x)(1-p)}\right) \\ + \left(\frac{(1-\varphi)}{y\varphi p+(1-y)\varphi(1-p)+(1-\varphi)}\right) x \end{array} \right) \\ - \left(\frac{y(\varphi p+1-\varphi)}{y(\varphi p+1-\varphi)+(1-y)(\varphi(1-p)+1-\varphi)}\right) \\ - \left(\frac{x(1-p)}{x(1-p)+(1-x)p} - \frac{y(1-p)}{y(1-p)+(1-y)p}\right). \end{array}$$

The argument is in two steps. Define the following function:

$$\begin{array}{l} \widetilde{f}_1(x,y,p,\varphi) \\ \equiv \quad \left(\begin{array}{c} \left(\frac{\varphi(yp+(1-y)(1-p))}{\varphi(yp+(1-y)(1-p))+(1-\varphi)}\right)\left(\frac{xp}{xp+(1-x)(1-p)}\right) \\ + \left(\frac{(1-\varphi)}{y\varphi p+(1-y)\varphi(1-p)+(1-\varphi)}\right) x \end{array} \right) \\ \qquad - \left(\frac{y(\varphi p+1-\varphi)}{y(\varphi p+1-\varphi)+(1-y)(\varphi(1-p)+1-\varphi)}\right) \\ - (x-y), \end{array}$$

which simplifies to

$$-\varphi(2p-1)\frac{(x-y)(p+x+y-px-py-xy+2pxy-1)}{(p+x-2px-1)(p\varphi+y\varphi-2py\varphi-1)}.$$

Note that this expression is positive for any $p \leq P_0(x,y)$ and recall that $P_0(x,y) > \frac{1}{2}$ iff $x < 1-y$. Note finally that given $x < 1-y$,

$$(x-y) > \frac{x(1-p)}{x(1-p)+(1-x)p} - \frac{y(1-p)}{y(1-p)+(1-y)p}.$$

We may thus conclude that a fortiori, for any $p \leq P_0(x,y)$ it also holds true that $f_1(x,y,p,\varphi) \geq 0$, implying that after a 1-signal, S prefers to disclose.■

Lemma I.E *Let $y \leq \frac{1}{2}$. There exists a D0-equilibrium if and only if $x \geq 1-y$* and *$p \leq P_1(x,y)$. If these conditions hold with strict inequality, it is furthermore the only equilibrium.*

Proof:

Step 0 Assume a putative D0 equilibrium. We prove a sequence of substatements which together yield the above Lemma. Denote by $g_i(x, y, p, \varphi)$ the difference between ex ante and ex post perceived disagreement given disclosure of an i-signal, for $i \in \{0, 1\}$.

Step 1 After a 0-signal, S should prefer to disclose. Note that:

$$\begin{aligned} & g_0(x, y, p, \varphi) \\ \equiv & \left(\begin{array}{c} \left(\frac{\varphi(y(1-p)+(1-y)p)}{\varphi(y(1-p)+(1-y)p)+(1-\varphi)}\right)\left(\frac{x(1-p)}{x(1-p)+(1-x)p}\right) + \left(\frac{(1-\varphi)}{\varphi(y(1-p)+(1-y)p)+(1-\varphi)}\right) x \\ - \left(\frac{y(\varphi(1-p)+1-\varphi)}{y(\varphi(1-p)+1-\varphi)+(1-y)(\varphi p+1-\varphi)}\right) \end{array} \right) \\ & - \left(\frac{xp}{xp+(1-x)(1-p)} - \frac{yp}{yp+(1-y)(1-p)}\right). \end{aligned}$$

Here, the argument is in two steps. Define the following function:

$$\begin{aligned} \widetilde{g}_0(x, y, p, \varphi) \equiv & \left(\begin{array}{c} \left(\frac{\varphi(y(1-p)+(1-y)p)}{\varphi(y(1-p)+(1-y)p)+(1-\varphi)}\right)\left(\frac{x(1-p)}{x(1-p)+(1-x)p}\right) + \left(\frac{(1-\varphi)}{\varphi(y(1-p)+(1-y)p)+(1-\varphi)}\right) x \\ - \left(\frac{y(\varphi(1-p)+1-\varphi)}{y(\varphi(1-p)+1-\varphi)+(1-y)(\varphi p+1-\varphi)}\right) \end{array} \right) \\ & - (x - y), \end{aligned}$$

which simplifies to

$$\varphi\,(2p-1)\,\frac{x-y}{p+x-2px}\,\frac{px-p+py+xy-2pxy}{p\varphi-\varphi+y\varphi-2py\varphi+1}$$

Note that $\widetilde{g}_0(x, y, p, \varphi) \geq 0$ for any $p \leq P_1(x, y)$. Recall furthermore that $P_1(x, y) > \frac{1}{2}$ iff $x > 1 - y$. Now, simply note that given $x > 1 - y$,

$$\frac{xp}{xp+(1-x)(1-p)} - \frac{yp}{yp+(1-y)(1-p)} < (x - y).$$

We may conclude that a fortiori for any $p \leq P_1(x, y)$, it holds true that $g_0(x, y, p, \varphi) \geq 0$, implying that after a 0-signal, S prefers to disclose.

Step 2 After a 1-signal, S should prefer to omit disclosing. Note that:

$$\begin{aligned} & g_1(x, y, p, \varphi) \\ \equiv & \left(\begin{array}{c} \left(\frac{\varphi(y(1-p)+(1-y)p)}{\varphi(y(1-p)+(1-y)p)+(1-\varphi)}\right)\left(\frac{x(1-p)}{x(1-p)+(1-x)p}\right) + \left(\frac{(1-\varphi)}{\varphi(y(1-p)+(1-y)p)+(1-\varphi)}\right) x \\ - \left(\frac{y(\varphi(1-p)+1-\varphi)}{y(\varphi(1-p)+1-\varphi)+(1-y)(\varphi p+1-\varphi)}\right) \end{array} \right) \\ & - \left(\frac{x(1-p)}{x(1-p)+(1-x)p} - \frac{y(1-p)}{y(1-p)+(1-y)p}\right). \end{aligned}$$

Note that $g_1(x, y, p, \varphi)$ simplifies to

$$(\varphi - 1)(2p - 1)\frac{x - y}{(p + x - 2px)(p + y - 2py)}\frac{px - p + py + xy - 2pxy}{p\varphi - \varphi + y\varphi - 2py\varphi + 1}.$$

Now, simply note that $g_1(x, y, p, \varphi) \leq 0$ for any $p \leq P_1(x, y)$.■

Lemma I.F *Let* $y \leq \frac{1}{2}$.
a) If $x < 1 - y$ *and* $p < P_0(x, y)$, *the D1 equilibrium is the only equilibrium.*
b) If $x > 1 - y$ *and* $p < P_1(x, y)$, *the D0 equilibrium is the only equilibrium.*
Proof:

Step 1 This proves Point a). Given the stated conditions, there exists no FD equilibrium and no D0 equilibrium. Invoking Lemma I.A, we may furthermore conclude that there exists no equilibrium featuring a mixed disclosure strategy.

Step 2 This proves Point b). Given the stated conditions, there exists no FD equilibrium and no D1 equilibrium. Invoking Lemma I.A, we may furthermore conclude that there exists no equilibrium featuring a mixed disclosure strategy. ■

5.2 Appendix II: Hidden cost of PC with binary signals

5.2.1 Preliminaries

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

In an FD equilibrium:

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

In a D0 equilibrium:

$$\begin{aligned}
\widetilde{\alpha}_R^{D0}(nd) &= \frac{\Pr[nd|\omega=0,D0]\alpha_R}{\Pr[nd|\omega=0,D0]\alpha_R+\Pr[nd|\omega=1,D0](1-\alpha_R)} \\
&= \frac{(\varphi(1-p)+(1-\varphi))\alpha_R}{(\varphi(1-p)+(1-\varphi))\alpha_R+(\varphi p+(1-\varphi))(1-\alpha_R)}, \\
E_R^{D0}[\widetilde{\alpha}_S|nd] &= \Pr[\sigma=1|nd,D0]\widetilde{\alpha}_S(1)+\Pr[\varnothing|nd,D0]\alpha_S \\
&= \frac{\Pr[\sigma=1]}{\Pr[\sigma=1]+\Pr[\varnothing]}\widetilde{\alpha}_S(1)+\frac{\Pr[\varnothing]}{\Pr[\sigma=1]+\Pr[\varnothing]}\alpha_S \\
&= \frac{\varphi((1-p)\alpha_R+p(1-\alpha_R))}{\varphi((1-p)\alpha_R+p(1-\alpha_R))+(1-\varphi)}\widetilde{\alpha}_S(1) \\
&\quad +\frac{1-\varphi}{\varphi((1-p)\alpha_R+p(1-\alpha_R))+(1-\varphi)}\alpha_S.
\end{aligned}$$

In a D1 equilibrium:

$$\begin{aligned}
\widetilde{\alpha}_R^{D1}(nd) &= \frac{\Pr[nd|\omega=0,D1]\alpha_R}{\Pr[nd|\omega=0,D1]\alpha_R+\Pr[nd|\omega=1,D1](1-\alpha_R)} \\
&= \frac{(\varphi p+(1-\varphi))\alpha_R}{(\varphi p+(1-\varphi))\alpha_R+(\varphi(1-p)+(1-\varphi))(1-\alpha_R)}, \\
E_R^{D1}[\widetilde{\alpha}_S|nd] &= \Pr[\sigma=0|nd,D1]\widetilde{\alpha}_S(0)+\Pr[\varnothing|nd,D1]\alpha_S \\
&= \frac{\Pr[\sigma=0]}{\Pr[\sigma=0]+\Pr[\varnothing]}\widetilde{\alpha}_S(0)+\frac{\Pr[\varnothing]}{\Pr[\sigma=0]+\Pr[\varnothing]}\alpha_S \\
&= \frac{\varphi(p\alpha_R+(1-p)(1-\alpha_R))}{\varphi(p\alpha_R+(1-p)(1-\alpha_R))+(1-\varphi)}\widetilde{\alpha}_S(0) \\
&\quad +\frac{1-\varphi}{\varphi(p\alpha_R+(1-p)(1-\alpha_R))+(1-\varphi)}\alpha_S.
\end{aligned}$$

5.2.2 Proof of Proposition 2

Step 1 Consider the case $\alpha_S > \alpha_R$ in D1 equilibrium. Using the expressions from section 5.2.1, the expected perceived disagreement for the sender can be derived as follows:

$$\begin{aligned}
E_S[\Delta^{D1}] &= \Pr[\sigma=0](E_R^{D1}[\widetilde{\alpha}_S|nd] - \widetilde{\alpha}_R^{D1}(nd)) + \Pr[\sigma=1](\widetilde{\alpha}_S(1) - \widetilde{\alpha}_R(1)) \\
+\Pr[\sigma &= \varnothing](E_R^{D1}[\widetilde{\alpha}_S|nd] - \widetilde{\alpha}_R^{D1}(nd)) \\
&= (\varphi(\alpha_S p + (1-\alpha_S)(1-p)) + 1 - \varphi) \\
&\quad \times \left(\begin{array}{c} \left(\frac{\varphi(\alpha_R p + (1-\alpha_R)(1-p))}{\varphi(\alpha_R p + (1-\alpha_R)(1-p)) + (1-\varphi)}\right)\left(\frac{\alpha_S p}{\alpha_S p + (1-\alpha_S)(1-p)}\right) \\ + \left(\frac{(1-\varphi)}{\alpha_R \varphi p + (1-\alpha_R)\varphi(1-p) + (1-\varphi)}\right)\alpha_S - \left(\frac{\alpha_R(\varphi p + 1 - \varphi)}{\alpha_R(\varphi p + 1 - \varphi) + (1-\alpha_R)(\varphi(1-p) + 1 - \varphi)}\right) \end{array} \right) \\
&\quad + \varphi(\alpha_S(1-p) + (1-\alpha_S)p)\left(\frac{\alpha_S(1-p)}{\alpha_S(1-p) + (1-\alpha_S)p} - \frac{\alpha_R(1-p)}{\alpha_R(1-p) + (1-\alpha_R)p}\right).
\end{aligned}$$

At the same time, under full disclosure

$$\begin{aligned}
E_S[\Delta^{FD}] &= \Pr[\sigma=0](\widetilde{\alpha}_S(0) - \widetilde{\alpha}_R(0)) + \Pr[\sigma=1](\widetilde{\alpha}_S(1) - \widetilde{\alpha}_R(1)) \\
+\Pr[\sigma &= \varnothing](\alpha_S - \alpha_R) \\
&= \varphi(\alpha_S p + (1-\alpha_S)(1-p))\left(\frac{\alpha_S p}{\alpha_S p + (1-\alpha_S)(1-p)} - \frac{\alpha_R p}{\alpha_R p + (1-\alpha_R)(1-p)}\right) \\
&\quad + \varphi(\alpha_S(1-p) + (1-\alpha_S)p)\left(\frac{\alpha_S(1-p)}{\alpha_S(1-p) + (1-\alpha_S)p} - \frac{\alpha_R(1-p)}{\alpha_R(1-p) + (1-\alpha_R)p}\right) \\
&\quad + (1-\varphi)(\alpha_S - \alpha_R).
\end{aligned}$$

Then,

$$\begin{aligned}
& E_S[\Delta^{D1}] - E_S[\Delta^{FD}] \\
= \; & (\varphi(\alpha_S p + (1-\alpha_S)(1-p)) + 1 - \varphi) \\
& \times \left(\begin{array}{c} \left(\frac{\varphi(\alpha_R p + (1-\alpha_R)(1-p))}{\varphi(\alpha_R p + (1-\alpha_R)(1-p)) + (1-\varphi)}\right)\left(\frac{\alpha_S p}{\alpha_S p + (1-\alpha_S)(1-p)}\right) \\ + \left(\frac{(1-\varphi)}{\alpha_R \varphi p + (1-\alpha_R)\varphi(1-p) + (1-\varphi)}\right)\alpha_S - \left(\frac{\alpha_R(\varphi p + 1 - \varphi)}{\alpha_R(\varphi p + 1 - \varphi) + (1-\alpha_R)(\varphi(1-p) + 1 - \varphi)}\right) \end{array} \right) \\
& - \varphi(\alpha_S p + (1-\alpha_S)(1-p))\left(\frac{\alpha_S p}{\alpha_S p + (1-\alpha_S)(1-p)} - \frac{\alpha_R p}{\alpha_R p + (1-\alpha_R)(1-p)}\right) \\
& - (1-\varphi)(\alpha_S - \alpha_R) \\
= \; & \Phi_1 \Phi_2
\end{aligned}$$

where

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

Note that Φ_2 is an increasing function of α_S. At the same time, by Proposition 1, $\alpha_S < \alpha_S^*$ in D1 equilibrium. Consequently,

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

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

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

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

Step 2 Consider the case $\alpha_S > \alpha_R$ in D0 equilibrium. The expected perceived disagreement in equilibrium for the sender is:

$$\begin{aligned}E_S[\Delta^{D0}] &= \Pr[\sigma=1](E_R^{D0}[\widetilde{\alpha}_S|nd]-\widetilde{\alpha}_R^{D0}(nd))+\Pr[\sigma=0](\widetilde{\alpha}_S(0)-\widetilde{\alpha}_R(0))\\ +\Pr[\sigma &= \varnothing](E_R^{D0}[\widetilde{\alpha}_S|nd]-\widetilde{\alpha}_R^{D0}(nd))\\ &= (\varphi(\alpha_S(1-p)+(1-\alpha_S)p)+1-\varphi)\\ &\quad\times\left(\begin{array}{c}\left(\frac{\varphi(\alpha_R(1-p)+(1-\alpha_R)p)}{\varphi(\alpha_R(1-p)+(1-\alpha_R)p)+(1-\varphi)}\right)\left(\frac{\alpha_S(1-p)}{\alpha_S(1-p)+(1-\alpha_S)p}\right)\\ +\left(\frac{(1-\varphi)}{\varphi(\alpha_R(1-p)+(1-\alpha_R)p)+(1-\varphi)}\right)\alpha_S\\ -\left(\frac{\alpha_R(\varphi(1-p)+1-\varphi)}{\alpha_R(\varphi(1-p)+1-\varphi)+(1-\alpha_R)(\varphi p+1-\varphi)}\right)\end{array}\right)\\ &\quad+\varphi(\alpha_S p+(1-\alpha_S)(1-p))\left(\frac{\alpha_S p}{\alpha_S p+(1-\alpha_S)(1-p)}-\frac{\alpha_R p}{\alpha_R p+(1-\alpha_R)(1-p)}\right).\end{aligned}$$

Then,

$$\begin{aligned}
& E_S[\Delta^{D0}] - E_S[\Delta^{FD}] \\
= \; & (\varphi(\alpha_S(1-p)+(1-\alpha_S)p)+1-\varphi) \\
& \times \left(\begin{array}{c} \left(\frac{\varphi(\alpha_R(1-p)+(1-\alpha_R)p)}{\varphi(\alpha_R(1-p)+(1-\alpha_R)p)+(1-\varphi)}\right)\left(\frac{\alpha_S(1-p)}{\alpha_S(1-p)+(1-\alpha_S)p}\right)+\left(\frac{(1-\varphi)}{\varphi(\alpha_R(1-p)+(1-\alpha_R)p)+(1-\varphi)}\right)\alpha_S \\ -\left(\frac{\alpha_R(\varphi(1-p)+1-\varphi)}{\alpha_R(\varphi(1-p)+1-\varphi)+(1-\alpha_R)(\varphi p+1-\varphi)}\right) \end{array} \right) \\
& -\varphi(\alpha_S(1-p)+(1-\alpha_S)p)\left(\frac{\alpha_S(1-p)}{\alpha_S(1-p)+(1-\alpha_S)p}-\frac{\alpha_R(1-p)}{\alpha_R(1-p)+(1-\alpha_R)p}\right) \\
& -(1-\varphi)(\alpha_S-\alpha_R) \\
= \; & \Phi_3\Phi_4,
\end{aligned}$$

where

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

Function Φ_4 is decreasing in α_S. At the same time, by Proposition 1 in D0-equilibrium we have $\alpha_S > \alpha_S^{**}$. Consequently,

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

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

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

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

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

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

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

5.2.3 Proof of Proposition 3

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

$$\begin{aligned}
\widetilde{\Theta}(\text{Partial},\widehat{\alpha}) &= E_{\widehat{\alpha}}\left[|\widetilde{\alpha}_S(\sigma,\text{Partial}) - \widetilde{\alpha}_R(\sigma,\text{Partial})|\right] \\
&\geq E_{\widehat{\alpha}}\left[\widetilde{\alpha}_S(\sigma,\text{Partial}) - \widetilde{\alpha}_R(\sigma,\text{Partial})\right] \\
&= E_{\widehat{\alpha}}[\widetilde{\alpha}_S(\sigma,\text{Partial})] - E_{\widehat{\alpha}}[\widetilde{\alpha}_R(\sigma,\text{Partial})] \\
&= E_{\widehat{\alpha}}\left[\widetilde{\alpha}_S(\sigma,\text{Full})\right] - E_{\widehat{\alpha}}\left[\widetilde{\alpha}_R(\sigma,\text{Partial})\right]. \qquad (2)
\end{aligned}$$

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

$$\begin{aligned}
\widetilde{\Theta}(\text{Full},\widehat{\alpha}) &= E_{\widehat{\alpha}}\left[|\widetilde{\alpha}_S(\sigma,\text{Full}) - \widetilde{\alpha}_R(\sigma,\text{Full})|\right] \\
&= E_{\widehat{\alpha}}\left[\widetilde{\alpha}_S(\sigma,\text{Full})\right] - E_{\widehat{\alpha}}\left[\widetilde{\alpha}_R(\sigma,\text{Full})\right]. \qquad (3)
\end{aligned}$$

It follows from the above that

$$\widetilde{\Theta}(\text{Partial},\widehat{\alpha}) - \widetilde{\Theta}(\text{Full},\widehat{\alpha}) \geq E_{\widehat{\alpha}}\left[\widetilde{\alpha}_R(\sigma,\text{Full})\right] - E_{\widehat{\alpha}}\left[\widetilde{\alpha}_R(\sigma,\text{Partial})\right]. \qquad (4)$$

Step 2 We now show that $E_{\widehat{\alpha}}\left[\widetilde{\alpha}_R(\sigma,\text{Full})\right] - E_{\widehat{\alpha}}\left[\widetilde{\alpha}_R(\sigma,\text{Partial})\right] > 0$ if and only if $\widehat{\alpha} > \alpha_R$. 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

$$\widetilde{\alpha}_R(\sigma) = \frac{\widehat{\alpha}(\sigma)\frac{\alpha_R}{\widehat{\alpha}}}{\widehat{\alpha}(\sigma)\frac{\alpha_R}{\widehat{\alpha}} + (1-\widehat{\alpha}(\sigma))\frac{1-\alpha_R}{1-\widehat{\alpha}}},$$

where $\widehat{\alpha}(\sigma)$ is the posterior belief of the receiver had she had a prior $\alpha_R = \widehat{\alpha}$. One can verify that the above function is concave in $\widehat{\alpha}(\sigma)$ if $\widehat{\alpha} < \alpha_R$ 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,[6] we obtain that

$$E_{\widehat{\alpha}}\left[\widetilde{\alpha}_R(\sigma, \text{Partial})\right] < (>) E_{\widehat{\alpha}}\left[\widetilde{\alpha}_R(\sigma, \text{Full})\right] \text{ if } \widehat{\alpha} > (<)\alpha_R \tag{5}$$

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

Step 3 (4) and (5) together imply

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

Thus, the third party would prefer full disclosure over partial disclosure whenever $\widehat{\alpha} > \alpha_R$, i.e., whenever $\widehat{\alpha}$ is either inbetween α_R and α_S or $\widehat{\alpha} > \alpha_S > \alpha_R$.

Step 4 Consider $\widehat{\alpha} < \alpha_R < \alpha_S$ with α_S being sufficiently close to 1. We have

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

(i.e., we have equalities at all stages in contrast to (2)). This together with (3) and (5) implies

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

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 $\alpha_S < \alpha_R$ is conceptually identical, and is hence omitted. In particular, we obtain that

$$\begin{aligned}
\widetilde{\Theta}(\text{Partial},\widehat{\alpha}) - \widetilde{\Theta}(\text{Full},\widehat{\alpha}) &> 0 \text{ if } \widehat{\alpha} < \alpha_R, \\
\widetilde{\Theta}(\text{Partial},\widehat{\alpha}) - \widetilde{\Theta}(\text{Full},\widehat{\alpha}) &< 0 \text{ if } \alpha_S < a_R < \widehat{\alpha} \text{ and } a_S \text{ is close to } 0.
\end{aligned}$$

■

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

5.3 Appendix III: Disclosure with binary signals and prior uncertainty

5.3.1 Proof of Proposition 4.a)

Step 1 Consider a putative FD equilibrium. Let G denote the (symmetric) cumulative distribution function of players' prior beliefs. Then, if the sender discloses 1-signal, the expected perceived disagreement is

$$\Delta(1) = \int_{\alpha_R=0}^{1}\int_{\alpha_S=0}^{1} |\widetilde{\alpha}_S(1,\alpha_S) - \widetilde{\alpha}_R(1,\alpha_R)|\, dG(\alpha_S)dG(\alpha_R),$$

where $\widetilde{\alpha}_i(1,\alpha_i)$ denotes the posterior belief of player i conditional on 1-signal and prior belief α_i. If the sender does not disclose, the expected perceived disagreement is

$$\Delta(nd) = \int_{\alpha_R=0}^{1}\int_{\alpha_S=0}^{1} |\alpha_S - \alpha_R| dG(\alpha_S)dG(\alpha_R).$$

In FD equilibrium we must have $\Delta(1) - \Delta(nd) < 0$. We have

$$\begin{aligned} & \Delta(1) - \Delta(nd) \\ = & \int_{\alpha_R=0}^{1}\int_{\alpha_S=0}^{1} \left(|\widetilde{\alpha}_S(1,\alpha_S) - \widetilde{\alpha}_R(1,\alpha_R)| - |\alpha_S - \alpha_R|\right) dG(\alpha_S)dG(\alpha_R). \end{aligned}$$

Denote $\kappa(\alpha_S,\alpha_R) = |\widetilde{\alpha}_S(1,\alpha_S) - \widetilde{\alpha}_R(1,\alpha_R)| - |\alpha_S - \alpha_R|$. Then,

$$\begin{aligned} & \int_{\alpha_R=0}^{1}\int_{\alpha_S=0}^{1} \left(|\widetilde{\alpha}_S(1,\alpha_S) - \widetilde{\alpha}_R(1,\alpha_R)| - |\alpha_S - \alpha_R|\right) dG(\alpha_S)dG(\alpha_R) \\ = & \int_{\alpha_R=0}^{1}\int_{\alpha_S=0}^{1} \kappa(\alpha_S,\alpha_R) dG(\alpha_S)dG(\alpha_R) \\ = & \int_{\alpha_R=0}^{0.5}\int_{\alpha_S=0}^{1} \left(\kappa(\alpha_S,\alpha_R) + \kappa(\alpha_S,1-\alpha_R)\right) dG(\alpha_S)dG(\alpha_R), \end{aligned}$$

where the last equality follows due to symmetry of G. Next, denote $\lambda(\alpha_S,\alpha_R) = \kappa(\alpha_S,\alpha_R) + \kappa(\alpha_S,1-\alpha_R)$. Then,

$$\begin{aligned} & \int_{\alpha_R=0}^{0.5}\int_{\alpha_S=0}^{1} \left(\kappa(\alpha_S,\alpha_R) + \kappa(\alpha_S,1-\alpha_R)\right) dG(\alpha_S)dG(\alpha_R) \qquad (6) \\ = & \int_{\alpha_R=0}^{0.5}\int_{\alpha_S=0}^{1} \lambda(\alpha_S,\alpha_R) dG(\alpha_S)dG(\alpha_R) \\ = & \int_{\alpha_R=0}^{0.5}\int_{\alpha_S=0}^{0.5} (\lambda(\alpha_S,\alpha_R) + \lambda(1-\alpha_S,\alpha_R)) dG(\alpha_S)dG(\alpha_R). \qquad (7) \end{aligned}$$

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

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

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

Step 2 By symmetry considerations, the same property holds for 0-signals, i.e. $\Delta(0) - \Delta(nd) < 0$. Formally, the proof proceeds analogously redefining $\kappa = |\widetilde{\alpha}_S(0, \alpha_S) - \widetilde{\alpha}_R(0, \alpha_R)| - |\alpha_S - \alpha_R|$. ■

5.3.2 Proof of Proposition 4.b)

Suppose that S's prior α_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

$$\begin{aligned}
\Delta(1) - \Delta(nd) &= \int_{\alpha_R=0}^{1} \left(|\widetilde{\alpha}_S(1,\alpha_S) - \widetilde{\alpha}_R(1,\alpha_R)| - |\alpha_S - \alpha_R|\right) dG(\alpha_R) \\
= \; & \int_{\alpha_R=0}^{0.5} \left(\kappa(\alpha_S, \alpha_R) + \kappa(\alpha_S, 1-\alpha_R)\right) dG(\alpha_R). \qquad (8)
\end{aligned}$$

Consider $\alpha_R < \frac{1}{2}$ such that $1 - \alpha_R > \alpha_S > \alpha_R$. For such α_R it holds

$$\begin{aligned}
\kappa(\alpha_S, \alpha_R) + \kappa(\alpha_S, 1-\alpha_R) &= (\widetilde{\alpha}_S(1,\alpha_S) - \widetilde{\alpha}_R(1,\alpha_R)) - (\alpha_S - \alpha_R) \\
&\quad + (\widetilde{\alpha}_R(1,1-\alpha_R) - \widetilde{\alpha}_S(1,\alpha_S)) - (1 - \alpha_R - \alpha_S) \\
&= \widetilde{\alpha}_R(1,1-\alpha_R) - \widetilde{\alpha}_R(1,\alpha_R) + 2\alpha_R - 1 \\
&= \frac{(1-\alpha_R)(1-p)}{(1-\alpha_R)(1-p) + \alpha_R p} - \frac{\alpha_R(1-p)}{\alpha_R(1-p) + (1-\alpha_R)p} + 2\alpha_R - 1 \\
&= -\frac{(1-2p)^2(1-\alpha_R)(1-2\alpha_R)\alpha_R}{(1-p+\alpha_R(2p-1))(\alpha_R + p(1-2\alpha_R))} < 0.
\end{aligned}$$

Since the probability mass of $\alpha_R < 0.5$ such that the condition $1 - \alpha_R > \alpha_S > \alpha_R$ is satisfied is sufficiently large for α_S sufficiently close to 0.5, the right-hand side of (8) is negative as well. Hence, the sender would prefer to disclose 1-signal over no disclosure. The same claim for 0-signals follows by symmetry considerations. Consequently, the FD equilibrium exists. ■

5.3.3 Proof of Proposition 5

Step 1 Note first that the expected payoff of R, given beliefs defined by the distribution $\{\alpha, 1-\alpha\}$ over $\{0,1\}$, is given by minus the variance of ω, which equals $\alpha(1-\alpha)$.

Step 2 We first consider the experiment, corresponding to a D1 equilibrium, in which only 1-signals are disclosed (denoted E_1). We denote by $d = \emptyset$ the event in which no signal is disclosed by S. Denote by $\Pi^R(1, E_1)$ ($\Pi^R(\emptyset, E_1)$) the expected payoff of R given disclosure of a 1-signal (no signal). The expected utility of R conditional on facing the E_1 experiment is given by:

$$P(d=1)\Pi^R(1,E_1) + P(d=\emptyset)\Pi^R(\emptyset,E_1).$$

Note first that if R's prior is α, then

$$P(d=1) = \varphi\left(\alpha(1-p) + (1-\alpha)p\right)$$

and

$$P(d=\emptyset) = \varphi\left(\alpha p + (1-\alpha)(1-p)\right) + (1-\varphi).$$

R's posterior distribution after a 1-signal is given by:

$$\frac{\alpha(1-p)}{\alpha(1-p)+(1-\alpha)p}, 1-\frac{\alpha(1-p)}{\alpha(1-p)+(1-\alpha)p}.$$

It follows that

$$\Pi^R(1,E_1) = -\left(\frac{\alpha(1-p)}{\alpha(1-p)+(1-\alpha)p}\right)\left(1-\frac{\alpha(1-p)}{\alpha(1-p)+(1-\alpha)p}\right).$$

R's posterior distribution after no disclosure is given by:

$$P(\omega=0\,|\emptyset,E_1\,) \equiv \left(\frac{\alpha\left(\varphi p+(1-\varphi)\right)}{\alpha\left(\varphi p+(1-\varphi)\right)+(1-\alpha)\left(\varphi(1-p)+(1-\varphi)\right)}\right), 1-P(\omega=0\,|\emptyset,E_1\,).$$

It follows that

$$\begin{aligned}&\Pi^R(\emptyset,E_1)\\ =\;& -P(\omega=0\,|\emptyset,E_1\,)\left(1-P(\omega=0\,|\emptyset,E_1\,)\right).\end{aligned}$$

The value of the experiment E_1 for R if the latter has prior α is thus:

$$\begin{aligned}&\Pi^R_{E_1}(\alpha,\varphi,p)\\ =\;& -\varphi\left(\alpha(1-p)+(1-\alpha)p\right)\left(\frac{\alpha(1-p)}{\alpha(1-p)+(1-\alpha)p}\right)\left(1-\frac{\alpha(1-p)}{\alpha(1-p)+(1-\alpha)p}\right)\\ &-(1-\varphi\left(\alpha(1-p)+(1-\alpha)p\right))\\ &\times\left(\begin{array}{c}\left(\frac{\alpha(\varphi p+(1-\varphi))}{\alpha(\varphi p+(1-\varphi))+(1-\alpha)(\varphi(1-p)+(1-\varphi))}\right)\\ \left(1-\left(\frac{\alpha(\varphi p+(1-\varphi))}{\alpha(\varphi p+(1-\varphi))+(1-\alpha)(\varphi(1-p)+(1-\varphi))}\right)\right)\end{array}\right).\end{aligned}$$

This simplifies significantly to:

$$\Pi^R_{E_1}(\alpha,\varphi,p) = -\alpha\frac{(\alpha-1)\left(p+\alpha-2p\alpha-\alpha\varphi-p^2\varphi+2p\alpha\varphi\right)}{(p+\alpha-2p\alpha)\left(p\varphi+\alpha\varphi-2p\alpha\varphi-1\right)}.$$

Step 3 We now consider the experiment, corresponding to a D0 equilibrium, in which only 0-signals are disclosed (denoted E_0). Denote by $\Pi^R(0,E_0)$ ($\Pi^R(\emptyset,E_0)$) the expected payoff of R given disclosure of a 0-signal (no signal). The expected utility of R conditional on facing the E_0 experiment is given by:

$$P(d=0)\Pi^R(0,E_0)+P(d=\emptyset)\Pi^R(\emptyset,E_0).$$

Note first that if R's prior is α, then

$$P(d=0) = \varphi\left(\alpha p + (1-\alpha)(1-p)\right)$$

and

$$P(d=\emptyset) = \varphi + (1-\varphi)\alpha(1-p) + (1-\alpha)p.$$

R's posterior distribution after a 0-signal is given by:

$$\frac{\alpha p}{\alpha p + (1-\alpha)(1-p)}, 1 - \frac{\alpha p}{\alpha p + (1-\alpha)(1-p)}.$$

It follows that:

$$\Pi^R(0, E_0) = -\left(\frac{\alpha p}{\alpha p + (1-\alpha)(1-p)}\right)\left(1 - \frac{\alpha p}{\alpha p + (1-\alpha)(1-p)}\right).$$

R's posterior distribution after no disclosure is given by:

$$P(\omega = 0 \,|\emptyset, E_0) \equiv \left(\frac{\alpha\left(\varphi(1-p) + (1-\varphi)\right)}{\alpha\left(\varphi(1-p) + (1-\varphi)\right) + (1-\alpha)(\varphi p + (1-\varphi))}\right), 1 - P(\omega = 0 \,|\emptyset, E_0).$$

It follows that:

$$\begin{aligned} & \Pi^R(\emptyset, E_0) \\ = \; & -\left(\frac{\alpha\left(\varphi(1-p) + (1-\varphi)\right)}{\alpha\left(\varphi(1-p) + (1-\varphi)\right) + (1-\alpha)(\varphi p + (1-\varphi))}\right) \\ & \left(1 - \left(\frac{\alpha\left(\varphi(1-p) + (1-\varphi)\right)}{\alpha\left(\varphi(1-p) + (1-\varphi)\right) + (1-\alpha)(\varphi p + (1-\varphi))}\right)\right). \end{aligned}$$

The value of the experiment E_0 to R if her prior is α is thus:

$$\begin{aligned} & \Pi^R_{E_0}(\alpha, \varphi, p) \\ = \; & -\varphi\left(\alpha p + (1-\alpha)(1-p)\right)\left(\frac{\alpha p}{\alpha p + (1-\alpha)(1-p)}\right)\left(1 - \frac{\alpha p}{\alpha p + (1-\alpha)(1-p)}\right) \\ & -(1 - \varphi\left(\alpha p + (1-\alpha)(1-p)\right)) \\ & \times \left(\begin{array}{c} \left(\frac{\alpha(\varphi(1-p)+(1-\varphi))}{\alpha(\varphi(1-p)+(1-\varphi))+(1-\alpha)(\varphi p+(1-\varphi))}\right) \\ \left(1 - \left(\frac{\alpha(\varphi(1-p)+(1-\varphi))}{\alpha(\varphi(1-p)+(1-\varphi))+(1-\alpha)(\varphi p+(1-\varphi))}\right)\right) \end{array} \right). \end{aligned}$$

This simplifies significantly to:

$$\Pi_0^R(\alpha,\varphi,p)=\alpha\frac{(\alpha-1)\left(p+\alpha+\varphi-2p\alpha-2p\varphi-\alpha\varphi+p^2\varphi+2p\alpha\varphi-1\right)}{(p+\alpha-2p\alpha-1)\left(p\varphi-\varphi+\alpha\varphi-2p\alpha\varphi+1\right)}.$$

Step 4 Using the obtained formulas, we have:

$$\begin{aligned}&\Pi_{E_0}^R(\alpha,\varphi,p)-\Pi_{E_1}^R(\alpha,\varphi,p)\\ =\;&-\alpha^2\varphi\left(\alpha-1\right)^2\left(\varphi-1\right)\left(2p-1\right)^3\\ &\frac{2\alpha-1}{\left(-4p^2\alpha^2+4p^2\alpha-p^2+4p\alpha^2-4p\alpha+p-\alpha^2+\alpha\right)}.\\ &\left(4p^2\alpha^2\varphi^2-4p^2\alpha\varphi^2+p^2\varphi^2-4p\alpha^2\varphi^2+4p\alpha\varphi^2-p\varphi^2+\alpha^2\varphi^2-\alpha\varphi^2+\varphi-1\right)\end{aligned}$$

We now show that the above expression is positive if $2\alpha-1<0$ and negative if the reverse inequality holds. Note that for any admissible p,φ,α,

$$\left(-4p^2\alpha^2+4p^2\alpha-p^2+4p\alpha^2-4p\alpha+p-\alpha^2+\alpha\right)$$

has the same sign. Similarly, for any admissible p,φ,α,

$$\left(4p^2\alpha^2\varphi^2-4p^2\alpha\varphi^2+p^2\varphi^2-4p\alpha^2\varphi^2+4p\alpha\varphi^2-p\varphi^2+\alpha^2\varphi^2-\alpha\varphi^2+\varphi-1\right)$$

has the same sign. The argument is as follows. Solving

$$-4p^2\alpha^2+4p^2\alpha-p^2+4p\alpha^2-4p\alpha+p-\alpha^2+\alpha=0$$

for α yields the solutions $\frac{p}{2p-1}$ and $\frac{1}{2p-1}\left(p-1\right)$, the first of which is <0 for any p and the second of which is >1 for any p. Similarly, solving

$$4p^2\alpha^2\varphi^2-4p^2\alpha\varphi^2+p^2\varphi^2-4p\alpha^2\varphi^2+4p\alpha\varphi^2-p\varphi^2+\alpha^2\varphi^2-\alpha\varphi^2+\varphi-1=0$$

for α yields the solutions $-\frac{1}{\varphi-2p\varphi}\left(-\varphi+p\varphi+1\right)$ and $-\frac{p\varphi-1}{\varphi-2p\varphi}$. Define the functions:

$$t_1(p,\varphi)=-\frac{1}{\varphi-2p\varphi}\left(-\varphi+p\varphi+1\right),$$

$$t_2(p,\varphi)=-\frac{p\varphi-1}{\varphi-2p\varphi}.$$

Note that $t_1(p,\varphi) > 1$ for any p, φ. To see this, note that $\frac{\partial t_1(p,\varphi)}{\partial \varphi} = -\frac{1}{\varphi^2(2p-1)}$ and that $t_1(p,1) = \frac{p}{2p-1} > 1$. Note on the other hand that $t_2(p,\varphi) < 0$ for any p, φ. To see this, note that $\frac{\partial t_2(p,\varphi)}{\partial \varphi} = \frac{1}{\varphi^2(2p-1)}$ and that $t_2(p,1) = \frac{1}{2p-1}(p-1) < 0$. We may conclude that the sign of $\Pi_0^R(\alpha,\varphi,p) - \Pi_1^R(\alpha,\varphi,p)$ is determined by the sign of $2\alpha - 1$. Specifically, it holds true that $\Pi_{E_0}^R(\alpha,\varphi,p) - \Pi_{E_1}^R(\alpha,\varphi,p) > (<)0$ if $\alpha < (>)\frac{1}{2}$. In words, if R's prior is biased towards state 0, she prefers to face D1 communication at the disclosure stage (and vice versa).

Step 5 Assume that the set of possible values of α_R (denoted Σ) is such that any possible value yields either the D0 or the D1 as the only incentive compatible strategy in the disclosure subgame. The set of priors Σ thus divides into two distinct sets of priors, Σ^- and Σ^+ satisfying the following. The subset Σ^- contains the values s.t. $\alpha_R < 1 - \alpha_S$ and Σ^+ contains the values s.t. $\alpha_R > 1 - \alpha_S$. Furthermore, all priors in Σ^- yield equilibrium D1 in the disclosure subgame while all priors in Σ^+ yield equilibrium D0 in the disclosure subgame. Assume a putative equilibrium featuring essentially truthful communication. So all members of Σ^- have an incentive to truthfully reveal that they belong to Σ^-. If this is the case, given the findings of previous steps, it must be that all elements of Σ^- are larger than $\frac{1}{2}$ since all types in Σ^- must prefer D1 over D0. This in turn implies that it must be true that $1 - \alpha_S > \frac{1}{2}$. This then implies that all priors in Σ^+ are strictly larger than $\frac{1}{2}$, implying that prior types belonging to Σ^+ strictly prefer D1 over D0. But this means that these will want to deviate to announcing that they belong to Σ^-, thereby ensuring that they face D1 at the disclosure stage. This yields a contradiction and thus proves that there cannot be an equilibrium in which R (essentially) truthfully reveals her prior. ■

5.4 Appendix IV: Disclosure with continuous signals (Proof of Proposition 6)

Proposition 6 follows from a set of Lemmas, which are stated and proved in what follows.

Lemma II.A *If $\alpha_S \neq \alpha_R$, then $\Delta(s) := |\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)|$ satisfies the following. i) $\Delta(\underline{s}) = \Delta(\overline{s}) = 0$. ii) There exists $\widehat{s}$ such that $\Delta(s)$ is increasing in s for all $s < \widehat{s}$ and decreasing in s for all $s > \widehat{s}$. iii) $\widetilde{s} < (>)\widehat{s}$ if the player with the lower prior is less (more) extreme. Instead, $\widetilde{s} = \widehat{s}$ if*

$\alpha_S = 1 - \alpha_R$, *i.e. if players are equally extreme.*

Proof:

Step 1 i) is immediate. To show ii) we show that there is a unique s such that $\frac{d}{ds}\left(\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)\right) = 0$ (this assumes strict MLRP). Note that

$$\begin{aligned}\frac{d}{ds}\left(\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)\right) &= \frac{d}{ds}\left(\frac{\alpha_S}{\alpha_S + (1-\alpha_S)\frac{f(s|h)}{f(s|l)}} - \frac{\alpha_R}{\alpha_R + (1-\alpha_R)\frac{f(s|h)}{f(s|l)}}\right) \\ &= \left(\frac{\alpha_R(1-\alpha_R)}{\left(\alpha_R + (1-\alpha_R)\frac{f(s|h)}{f(s|l)}\right)^2} - \frac{\alpha_S(1-\alpha_S)}{\left(\alpha_S + (1-\alpha_S)\frac{f(s|h)}{f(s|l)}\right)^2}\right)\frac{d}{ds}\frac{f(s|h)}{f(s|l)}.\end{aligned}$$

Consider the solution to

$$\alpha_R(1-\alpha_R)\left(\alpha_S + (1-\alpha_S)\frac{f(s|h)}{f(s|l)}\right)^2 = \alpha_S(1-\alpha_S)\left(\alpha_R + (1-\alpha_R)\frac{f(s|h)}{f(s|l)}\right)^2.$$

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

$$\begin{aligned}&\frac{d}{ds}\alpha_R(1-\alpha_R)\left(\alpha_S + (1-\alpha_S)\frac{f(s|h)}{f(s|l)}\right)^2 \\ =\ & 2\alpha_R(1-\alpha_R)(1-\alpha_S)\left(\alpha_S + (1-\alpha_S)\frac{f(s|h)}{f(s|l)}\right)\frac{d}{ds}\frac{f(s|h)}{f(s|l)}, \\ &\frac{d}{ds}\alpha_S(1-\alpha_S)\left(\alpha_R + (1-\alpha_R)\frac{f(s|h)}{f(s|l)}\right)^2 \\ =\ & 2\alpha_S(1-\alpha_R)(1-\alpha_S)\left(\alpha_R + (1-\alpha_R)\frac{f(s|h)}{f(s|l)}\right)\frac{d}{ds}\frac{f(s|h)}{f(s|l)}.\end{aligned}$$

The result follows by assumption as

$$\begin{aligned}&2\alpha_R(1-\alpha_R)(1-\alpha_S)\left(\alpha_S + (1-\alpha_S)\frac{f(s|h)}{f(s|l)}\right)\frac{d}{ds}\frac{f(s|h)}{f(s|l)} \\ \gtreqless\ & 2\alpha_S(1-\alpha_R)(1-\alpha_S)\left(\alpha_R + (1-\alpha_R)\frac{f(s|h)}{f(s|l)}\right)\frac{d}{ds}\frac{f(s|h)}{f(s|l)},\end{aligned}$$

which is equivalent to

$$\alpha_R\alpha_S + \alpha_R(1-\alpha_S)\frac{f(s|h)}{f(s|l)} \gtreqless \alpha_R\alpha_S + \alpha_S(1-\alpha_R)\frac{f(s|h)}{f(s|l)}$$

which is equivalent to $\alpha_R \gtreqless \alpha_S$. Hence, $\widehat{s}$ must be unique. Existence follows from continuity and (i) together with $\Delta(\widetilde{s}) = |\alpha_S - \alpha_R| > 0$.

Step 2 To show (iii), define $\alpha_{\max} = \max\{\alpha_S, \alpha_R\}$ and $\alpha_{\min} = \min\{\alpha_S, \alpha_R\}$ such that $\Delta(s) = \widetilde{\alpha}_{\max}(s) - \widetilde{\alpha}_{\min}(s)$. We then have:

$$\begin{aligned}\frac{d}{ds}\Delta(\widetilde{s}) &= \left(\frac{\alpha_{\min}(1-\alpha_{\min})}{(\alpha_{\min}+(1-\alpha_{\min}))^2} - \frac{\alpha_{\max}(1-\alpha_{\max})}{(\alpha_{\max}+(1-\alpha_{\max}))^2}\right)\frac{d}{ds}\frac{f(\widetilde{s}|h)}{f(\widetilde{s}|l)} \gtreqless 0 \\ &\iff \alpha_{\min}(1-\alpha_{\min}) \gtreqless \alpha_{\max}(1-\alpha_{\max}),\end{aligned}$$

and $\widetilde{s} < (>)\widehat{s}$ if $\alpha_{\min}$ is less extreme than $\alpha_{\max}$.■

Lemma II.B *There exists a simple disclosure equilibrium and any equilibrium is a simple disclosure equilibrium.*

Proof:

Step 0 Steps 1-2 introduce key equilibrium conditions. In steps 3-4, we show that there exists at least one SDE. Step 5 proves that any equilibrium is an SDE. In what follows, we assume $\alpha_S > \alpha_R$. The proof for the reverse case follows the same steps and omitted.

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

$$\Pr_R(s \in \Psi^{nd}) = \alpha_R \int_{\Psi^{nd}} f(s|l)ds + (1-\alpha_R)\int_{\Psi^{nd}} f(s|h)ds.$$

Hence, when S does not disclose, R's posterior is

$$\begin{aligned}\widetilde{\alpha}_R(nd) &= \frac{\varphi}{(1-\varphi)+\varphi\Pr_R(s\in\Psi^{nd})}\int_{\Psi^{nd}}\left(\alpha_R f(s|l)+(1-\alpha_R)f(s|h)\right)\widetilde{\alpha}_R(s)ds \\ &\quad + \frac{(1-\varphi)}{(1-\varphi)+\varphi\Pr_R(s\in\Psi^{nd})}\alpha_R \\ &= \frac{\varphi\int_{\Psi^{nd}} f(s|l)ds + (1-\varphi)}{(1-\varphi)+\varphi\Pr_R(s\in\Psi^{nd})}\alpha_R.\end{aligned}$$

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

$$\begin{aligned}\widetilde{\alpha}_{RS}(nd) &= \frac{\varphi}{(1-\varphi)+\varphi\Pr_R(s\in\Psi^{nd})}\int_{\Psi^{nd}}\left(\alpha_R f(s|l)+(1-\alpha_R)f(s|h)\right)\widetilde{\alpha}_S(s)ds \\ &\quad + \frac{(1-\varphi)}{(1-\varphi)+\varphi\Pr_R(s\in\Psi^{nd})}\alpha_S.\end{aligned}$$

Step 2 The discrepancy in beliefs upon disclosure is given by $|\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)|$ and at s_1 and s_2, S must be indifferent between disclosure and non-disclosure. Hence, we require

$$|\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)| = |\widetilde{\alpha}_{RS}(nd) - \widetilde{\alpha}_R(nd)| \text{ for } s = s_1, s_2. \tag{9}$$

Step 3 Recall that $|\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)|$ is concave and single peaked in s over $[\underline{s}, \overline{s}]$ and that $\widehat{s}$ is the signal that maximizes $|\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)|$. For each $\varepsilon \in [\alpha_S - \alpha_R, \varepsilon_{\max} = |\widetilde{\alpha}_S(\widehat{s}) - \widetilde{\alpha}_R(\widehat{s})|]$, let $\mathbf{s}(\varepsilon) = \{s_1(\varepsilon), s_2(\varepsilon)\}$ denote the unique pair of thresholds such that $|\widetilde{\alpha}_S(s_i(\varepsilon)) - \widetilde{\alpha}_R(s_i(\varepsilon))| = \varepsilon$, for $i = 1, 2$. Note that by definition, $\lim_{\varepsilon \to \varepsilon_{\max}} s_1(\varepsilon) = \lim_{\varepsilon \to \varepsilon_{\max}} s_2(\varepsilon) = \widehat{s}$. Denote by

$$|E_R\widetilde{\alpha}_S(nd, s_1, s_2) - \widetilde{\alpha}_R(nd, s_1, s_2)|$$

the perceived difference in posteriors given no disclosure, given a simple disclosure rule specified by the no-disclosure interval $[s_1, s_2]$. Recall finally that a simple disclosure rule (s_1, s_2) constitutes an equilibrium disclosure rule iff:

$$|\widetilde{\alpha}_S(s_i) - \widetilde{\alpha}_R(s_i)| = |E_R\widetilde{\alpha}_S(nd, s_1, s_2) - \widetilde{\alpha}_R(nd, s_1, s_2)|, \text{ for } i = 1, 2.$$

Step 4 Note first that

$$|\widetilde{\alpha}_S(s_1(\varepsilon)) - \widetilde{\alpha}_R(s_1(\varepsilon))|$$

and

$$|E_R\widetilde{\alpha}_S(nd, s_1(\varepsilon), s_2(\varepsilon)) - \widetilde{\alpha}_R(nd, s_1(\varepsilon), s_2(\varepsilon))|$$

are both continuous in ε for $\varepsilon \in [\alpha_S - \alpha_R, \varepsilon_{\max} = \widetilde{\alpha}_S(\widehat{s}) - \widetilde{\alpha}_R(\widehat{s})]$. Second, note that it is trivially true that:

$$\begin{aligned} & |E_R\widetilde{\alpha}_S(nd, s_1(\varepsilon_{\max}), s_2(\varepsilon_{\max})) - \widetilde{\alpha}_R(nd, s_1(\varepsilon_{\max}), s_2(\varepsilon_{\max}))| \\ < \; & |\widetilde{\alpha}_S(s_1(\varepsilon_{\max})) - \widetilde{\alpha}_R(s_1(\varepsilon_{\max}))|. \end{aligned}$$

Third, note that.

$$\begin{aligned} & |E_R\widetilde{\alpha}_S(nd, s_1(\alpha_S - \alpha_R), s_2(\alpha_S - \alpha_R)) - \widetilde{\alpha}_R(nd, s_1(\alpha_S - \alpha_R), s_2(\alpha_S - \alpha_R))| \\ > \; & |\widetilde{\alpha}_S(s_1(\alpha_S - \alpha_R)) - \widetilde{\alpha}_R(s_1(\alpha_S - \alpha_R))| = \alpha_S - \alpha_R. \end{aligned}$$

It follows that there must exist some $\varepsilon \in [\alpha_S - \alpha_R, \varepsilon_{\max} = |\widetilde{\alpha}_S(\widehat{s}) - \widetilde{\alpha}_R(\widehat{s})|$ such that

$$\begin{aligned} & |\widetilde{\alpha}_S(s_1(\varepsilon)) - \widetilde{\alpha}_R(s_1(\varepsilon))| \\ = \; & |E_R\widetilde{\alpha}_S(nd, s_1(\varepsilon), s_2(\varepsilon)) - \widetilde{\alpha}_R(nd, s_1(\varepsilon), s_2(\varepsilon))|\,. \end{aligned}$$

Step 5 We prove by contradiction that any equilibrium is a simple disclosure equilibrium. Assume thus an equilibrium which is not an SDE. Upon non-disclosure, let the perceived difference in beliefs be denoted by C and this is by definition ≥ 0. Suppose first that $C > 0$. Clearly, conditional on obtaining a signal, S wants to disclose if the resulting discrepancy $\Delta(s)$ is smaller than C. Recall now that $\Delta(s)$ is single peaked and concave. It follows that for any $C > 0$, there are s_1, s_2 satisfying $\underline{s} < s_1 < s_2 < \overline{s}$ such that the actual disagreement in beliefs is strictly higher than C for $\sigma \in (s_1, s_2)$ and strictly lower than C if $\sigma < s_1$ and $\sigma > s_2$. 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. Suppose now that $C = 0$. In such a case, S would strictly prefer not to disclose any signal $s \in (\underline{s}, \overline{s})$. So the equilibrium should feature no disclosure of any $s \in (\underline{s}, \overline{s})$. But then, it must be the case that $C > 0$, a contradiction. ■

Lemma II.C *a) If $\alpha_R = 1 - \alpha_S$ then $s_1 = s_2$, i.e. there is full disclosure.*

b) If $\alpha_R \neq 1 - \alpha_S$ then $s_1 < s_2$, i.e. a non-empty set of signals $s \in (s_1, s_2)$ is not disclosed.

c) Assume that $\alpha_S > \alpha_R$. If $\alpha_R < 1 - \alpha_S$, 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 $\alpha_R > 1 - \alpha_S$, 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.

d) Assume that $\alpha_S < \alpha_R$. If $\alpha_R > 1 - \alpha_S$, 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 $\alpha_R < 1 - \alpha_S$, 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:

Step 0 We focus throughout on the case of $\alpha_S > \alpha_R$. The proof for the reverse case follows the same steps and omitted.

Step 1 The discrepancy in beliefs upon disclosure is given by $|\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)|$ and at s_1 and s_2, S must be indifferent between disclosure and non-disclosure. Hence, we require

$$|\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)| = |\widetilde{\alpha}_{RS}(nd) - \widetilde{\alpha}_R(nd)| \text{ for } s = s_1, s_2, \tag{10}$$

which directly implies from the preceding Lemma that $s_1 \leq \widehat{s} \leq s_2$, strictly if $s_1 < s_2$. Further, it then follows directly from

$$\begin{aligned} & \widetilde{\alpha}_{RS}(nd) - \widetilde{\alpha}_R(nd) \\ = & \frac{\varphi}{(1-\varphi) + \varphi \Pr_R(s \in \Psi^{nd})} \int_{s_1}^{s_2} (\alpha_R f(s|l) + (1-\alpha_R) f(s|h)) (\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)) \, ds \\ & + \frac{(1-\varphi)}{(1-\varphi) + \varphi \Pr_R(s \in \Psi^{nd})} (\alpha_S - \alpha_R) \end{aligned}$$

together with (10) that under the optimal disclosure rule we must have

$$\begin{aligned} \widetilde{\alpha}_S(s_1) - \widetilde{\alpha}_R(s_1) &= \widetilde{\alpha}_S(s_2) - \widetilde{\alpha}_R(s_2) = \widetilde{\alpha}_S(nd) - \widetilde{\alpha}_R(nd) \\ &> \alpha_S - \alpha_R, \end{aligned}$$

implying that the uninformative signal $\widetilde{s}$ is always disclosed. Thus, if $\widetilde{s} = \widehat{s}$, then there is full disclosure conditional on an available signal.

Step 2 It remains to be shown that for $\widetilde{s} \neq \widehat{s}$, we have $s_1 < \widehat{s} < s_2$. We will argue by contradiction. Suppose that $\widetilde{s} \neq \widehat{s}$ and that there is an equilibrium with full disclosure conditional on an available signal. In such an equilibrium, disclosing $s = \widehat{s}$ leads to the a perceived disagreement of $\widetilde{\alpha}_S(\widehat{s}) - \widetilde{\alpha}_R(\widehat{s})$ which, from $\widetilde{s} \neq \widehat{s}$, is strictly greater than the perceived disagreement without disclosure, the latter being given by $|\widetilde{\alpha}_{RS}(nd) - \widetilde{\alpha}_R(nd)| = \alpha_S - \alpha_R$. Hence, disclosure of $s = \widehat{s}$ cannot be optimal. The remaining results then follow from Lemma II.A. ■

Lemma II.D *a) If there exist multiple equilibria, then the equilibria are ordered in terms of Blackwell informativeness. b) If φ increases, the most Blackwell informative equilibrium becomes more informative.*

Proof:

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

Step 1 Note that if $\mathbf{s}(\varepsilon)$ and $\mathbf{s}(\varepsilon')$ are two equilibrium disclosure rules and $\varepsilon' > \varepsilon$, then $s_1(\varepsilon') > s_1(\varepsilon)$ and $s_2(\varepsilon') < s_2(\varepsilon)$ so that $(s_1(\varepsilon'), s_2(\varepsilon')) \subset (s_1(\varepsilon), s_2(\varepsilon))$. This is true because $|\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)|$ is concave and single peaked in s over $[\underline{s}, \overline{s}]$. This furthermore implies that $\mathbf{s}(\varepsilon')$ is more Blackwell informative than $\mathbf{s}(\varepsilon)$.

Step 2 Note that

$$\begin{aligned} & \widetilde{\alpha}_{RS}(nd, s_1(\varepsilon), s_2(\varepsilon)) - \widetilde{\alpha}_R(nd, s_1(\varepsilon), s_2(\varepsilon)) \\ = & \frac{\varphi \int_{s_1(\varepsilon)}^{s_2(\varepsilon)} (\alpha_R f(s|l) + (1-\alpha_R) f(s|h)) (\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)) ds + (1-\varphi)(\alpha_S - \alpha_R)}{(1-\varphi) + \varphi \Pr_R(s \in \Psi^{nd})}, \end{aligned}$$

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

$$A = \int_{s_1(\varepsilon)}^{s_2(\varepsilon)} (\alpha_R f(s|l) + (1-\alpha_R) f(s|h)) (\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)) ds$$

and $\delta = (\alpha_S - \alpha_R)$, it follows that :

$$\frac{\partial \widetilde{\alpha}_{RS}(nd, s_1(\varepsilon), s_2(\varepsilon)) - \widetilde{\alpha}_R(nd, s_1(\varepsilon), s_2(\varepsilon))}{\partial \varphi} = \frac{\begin{array}{c}(A-\delta)\left[(1-\varphi) + \varphi \Pr_R(s \in \Psi^{nd})\right] \\ -\left[\varphi A + (1-\varphi)\delta\right]\left[-1 + \Pr_R(s \in \Psi^{nd})\right]\end{array}}{\left[(1-\varphi) + \varphi \Pr_R(s \in \Psi^{nd})\right]^2}.$$

The above expression rewrites and simplifies as follows:

$$\begin{aligned} & \frac{\begin{array}{c}(A - (\alpha_S - \alpha_R))\left[(1-\varphi) + \varphi \Pr_R(s \in \Psi^{nd})\right] \\ -\left[\varphi A + (1-\varphi)\delta\right]\left[-1 + \Pr_R(s \in \Psi^{nd})\right]\end{array}}{\left[(1-\varphi) + \varphi \Pr_R(s \in \Psi^{nd})\right]^2} \\ = & \frac{\begin{array}{c}A(1-\varphi) + A\varphi \Pr_R(s \in \Psi^{nd}) - \delta(1-\varphi) - \delta\varphi \Pr_R(s \in \Psi^{nd}) \\ +\varphi A - \varphi A \Pr_R(s \in \Psi^{nd}) + (1-\varphi)\delta - (1-\varphi)\delta \Pr_R(s \in \Psi^{nd})\end{array}}{\left[(1-\varphi) + \varphi \Pr_R(s \in \Psi^{nd})\right]^2} \\ = & \frac{A - \Pr_R(s \in \Psi^{nd})\delta}{\left[(1-\varphi) + \varphi \Pr_R(s \in \Psi^{nd})\right]^2} \\ = & \frac{\int_{s_1(\varepsilon)}^{s_2(\varepsilon)} (\alpha_R f(s|l) + (1-\alpha_R) f(s|h)) (\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)) ds - \Pr_R(s \in \Psi^{nd})(\alpha_S - \alpha_R)}{\left[(1-\varphi) + \varphi \Pr_R(s \in \Psi^{nd})\right]^2} > 0 \end{aligned}$$

To see that the last inequality holds, remember that we are only looking at the set of εs satisfying $\varepsilon > \alpha_S - \alpha_R$, which implies that

$$\begin{aligned} & \int_{s_1(\varepsilon)}^{s_2(\varepsilon)} \left(\alpha_R f(s|l) + (1-\alpha_R) f(s|h)\right) \left(\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)\right) ds \\ > \quad & \alpha_S - \alpha_R. \end{aligned}$$

Step 3 Let us slightly abuse notation and write

$$\widetilde{\alpha}_{RS}(nd, s_1(\varepsilon), s_2(\varepsilon), \varphi) - \widetilde{\alpha}_R(nd, s_1(\varepsilon), s_2(\varepsilon), \varphi)$$

for the perceived difference in beliefs given no-disclosure, assuming disclosure rule $\{s_1(\varepsilon), s_2(\varepsilon)\}$ and parameter φ. Let $\mathbf{s}^{\varphi}(\varepsilon) = \left\{s_1^{\varphi}(\varepsilon), s_2^{\varphi}(\varepsilon)\right\}$ be the most informative disclosure rule under φ. Note that given our previous step, assuming $\varphi' > \varphi$ it holds true that

$$\begin{aligned} & \widetilde{\alpha}_{RS}(nd, s_1^{\varphi}(\varepsilon), s_2^{\varphi}(\varepsilon), \varphi') - \widetilde{\alpha}_R(nd, s_1^{\varphi}(\varepsilon), s_2^{\varphi}(\varepsilon), \varphi') \\ > \quad & \\ & \widetilde{\alpha}_{RS}(nd, s_1^{\varphi}(\varepsilon), s_2^{\varphi}(\varepsilon), \varphi) - \widetilde{\alpha}_R(nd, s_1^{\varphi}(\varepsilon), s_2^{\varphi}(\varepsilon), \varphi). \end{aligned}$$

It follows that there exists some $\varepsilon' > \varepsilon$ such that given φ', the disclosure rule $\{s_1(\varepsilon'), s_2(\varepsilon')\}$ constitutes an equilibrium disclosure rule. Given that $|\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)|$ is concave and single peaked in s over $[\underline{s}, \overline{s}]$, it follows that $s_1(\varepsilon') > s_1(\varepsilon)$ and $s_2(\varepsilon') < s_2(\varepsilon)$ so that $(s_1(\varepsilon'), s_2(\varepsilon')) \subset (s_1(\varepsilon), s_2(\varepsilon))$. This implies that $\{s_1(\varepsilon'), s_2(\varepsilon')\}$ is more Blackwell informative than $\{s_1(\varepsilon), s_2(\varepsilon)\}$. ■

5.5 Appendix V: Hidden cost of PC with continuous signals

5.5.1 Proof of Proposition 7

Step 0 We prove 1. in what follows. By assumption, it holds true that $\widetilde{s} < s_1 < s_2$. It follows that the most extreme player is biased towards state 0, omitted signals indicate state 1 and $\alpha_S \neq 1 - \alpha_R$. We focus on proving that S would strictly prefer to commit to full disclosure if $\alpha_S > \alpha_R$. Note that combining the assumptions $\alpha_S > \alpha_R$ and $\widetilde{s} < s_1 < s_2$ implies that $\alpha_R \in (1 - \alpha_S, \alpha_S)$. The proof that S instead prefers equilibrium disclosure

given $\alpha_S < \alpha_R$ 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 a simple disclosure equilibrium (SDE) featuring thresholds $\{s_1, s_2\}$ is given by:

$$\begin{aligned}
&(1-\varphi)\left[\widetilde{\alpha}_{RS}(nd) - \widetilde{\alpha}_R(nd)\right] \\
&+\varphi \int_{s_1}^{s_2} \left(\alpha_S f(s|l) + (1-\alpha_S) f(s|h)\right) \left[\widetilde{\alpha}_{RS}(nd) - \widetilde{\alpha}_R(nd)\right] ds \\
&+\varphi \int_{s \notin \Psi^{nd}} \left(\alpha_S f(s|l) + (1-\alpha_S) f(s|h)\right) \left[\widetilde{\alpha}_{RS}(s) - \widetilde{\alpha}_R(s)\right] ds.
\end{aligned}$$

Recall also that we know from previous derivations that

$$\begin{aligned}
&\widetilde{\alpha}_{RS}(nd) - \widetilde{\alpha}_R(nd) \\
= \;& \frac{\varphi}{(1-\varphi) + \varphi \Pr_R(s \in \Psi^{nd})} \int_{s_1}^{s_2} \left(\alpha_R f(s|l) + (1-\alpha_R) f(s|h)\right) \left(\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)\right) ds \\
&+ \frac{(1-\varphi)}{(1-\varphi) + \varphi \Pr_R(s \in \Psi^{nd})} \left(\alpha_S - \alpha_R\right).
\end{aligned}$$

Step 2 We here consider a putative full disclosure equilibrium. Given disclosure s, $\widetilde{\alpha}_{RS}(s) = \widetilde{\alpha}_S(s)$. From S's perspective, the ex ante perceived disagreement in an equilibrium with full disclosure is thus:

$$\begin{aligned}
&\varphi \int_{s_1}^{s_2} \left(\alpha_S f(s|l) + (1-\alpha_S) f(s|h)\right) \left[\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)\right] ds \\
&+\varphi \int_{s \notin \Psi^{nd}} \left(\alpha_S f(s|l) + (1-\alpha_S) f(s|h)\right) \left[\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)\right] ds \\
&+ (1-\varphi)\left[\alpha_S - \alpha_R\right].
\end{aligned}$$

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. We have:

$$\begin{aligned}
\Theta(\text{Partial}) &= \left[\varphi \Pr_S(s \in \Psi^{nd}) + (1-\varphi)\right] \left[\widetilde{\alpha}_{RS}(nd) - \widetilde{\alpha}_R(nd)\right] \\
&= \left[(1-\varphi) + \varphi \Pr_R(s \in \Psi^{nd})\right] \\
&\quad \times \left[\begin{array}{c} \frac{\varphi}{(1-\varphi)+\varphi \Pr_R(s \in \Psi^{nd})} \int_{s_1}^{s_2} \left(\alpha_R f(s|l) + (1-\alpha_R) f(s|h)\right) \left(\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)\right) ds \\ + \frac{(1-\varphi)}{(1-\varphi)+\varphi \Pr_R(s \in \Psi^{nd})} \left(\alpha_S - \alpha_R\right) \end{array}\right]
\end{aligned}$$

and

$$\Theta(\text{Full}) = \varphi \int_{s_1}^{s_2} \left[\alpha_S f(s|l) + (1-\alpha_S) f(s|h)\right] \left[\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)\right] ds + (1-\varphi)(\alpha_S - \alpha_R).$$

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

$$\left[\varphi \Pr_S(s \in \Psi^{nd}) + (1-\varphi)\right] \left[\widetilde{\alpha}_{RS}(nd) - \widetilde{\alpha}_R(nd)\right] \tag{11}$$

$$> \quad \varphi \int_{s_1}^{s_2} \left[\alpha_S f(s|l) + (1-\alpha_S) f(s|h)\right] \left[\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)\right] ds + (1-\varphi)(\alpha_S - \alpha_R). \tag{12}$$

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

$$\Pr_{\widehat{\alpha}_R}(s \in \Psi^{nd}) = \int_{s_1}^{s_2} (\widehat{\alpha}_R f(s|l) + (1-\widehat{\alpha}_R) f(s|h)) ds.$$

We define $\Delta(\alpha_S, \alpha_R, \widehat{\alpha}_R)$ as a slightly modified version of $\widetilde{\alpha}_{RS}(nd) - \widetilde{\alpha}_R(nd)$. We let

$$\begin{aligned} & \Delta(\alpha_S, \alpha_R, \widehat{\alpha}_R) \\ = \quad & \frac{\varphi}{(1-\varphi) + \varphi \Pr_{\widehat{\alpha}_R}(s \in \Psi^{nd})} \int_{s_1}^{s_2} (\widehat{\alpha}_R f(s|l) + (1-\widehat{\alpha}_R) f(s|h)) \left(\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)\right) ds \\ & + \frac{(1-\varphi)}{(1-\varphi) + \varphi \Pr_{\widehat{\alpha}_R}(s \in \Psi^{nd})} (\alpha_S - \alpha_R). \end{aligned}$$

Let us finally define

$$\widetilde{\Theta}(\text{Partial}, \widehat{\alpha}_R) = \left[\varphi \Pr_S(s \in \Psi^{nd}) + (1-\varphi)\right] \left[\Delta(\alpha_S, \alpha_R, \widehat{\alpha}_R)\right]$$

and note that $\widetilde{\Theta}(\text{Partial}, \alpha_R) = \Theta(\text{Partial})$.

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

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

$$\begin{aligned}
& \widetilde{\Theta}(\text{Partial}, \widehat{\alpha}_R) && (13)\\
= & \left[\varphi \Pr_S(s \in \Psi^{nd}) + (1-\varphi)\right] \left[\Delta\left(\alpha_S, \alpha_R, \alpha_S\right)\right] && (14)\\
= & \left[\begin{array}{c} \varphi \Pr_S(s \in \Psi^{nd}) \\ +(1-\varphi) \end{array}\right] && (15)\\
& \times \left[\begin{array}{c} \frac{\varphi}{(1-\varphi)+\varphi \Pr_{\alpha_S}(s \in \Psi^{nd})} \int_{s_1}^{s_2} \left(\alpha_S f(s|l) + (1-\alpha_S) f(s|h)\right) \left(\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)\right) ds \\ + \frac{(1-\varphi)}{(1-\varphi)+\varphi \Pr_{\alpha_S}(s \in \Psi^{nd})} \left(\alpha_S - \alpha_R\right) \end{array}\right] && (16)\\
= & \varphi \int_{s_1}^{s_2} \left[\alpha_S f(s|l) + (1-\alpha_S) f(s|h)\right] \left[\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)\right] ds + (1-\varphi)\left(\alpha_S - \alpha_R\right) && (17)\\
= & \Theta(\text{Full}). && (18)
\end{aligned}$$

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

$$\begin{aligned}
& \Delta\left(\alpha_S, \alpha_R, \widehat{\alpha}_R\right) \\
= & \left[\begin{array}{c} \frac{\varphi \Pr_{\widehat{\alpha}_R}(s \in \Psi^{nd})}{(1-\varphi)+\varphi \Pr_{\widehat{\alpha}_R}(s \in \Psi^{nd})} \int_{s_1}^{s_2} \frac{\left(\widehat{\alpha}_R f(s|l) + (1-\widehat{\alpha}_R) f(s|h)\right)}{\Pr_{\widehat{\alpha}_R}(s \in \Psi^{nd})} \left(\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)\right) ds \\ + \frac{(1-\varphi)}{(1-\varphi)+\varphi \Pr_{\widehat{\alpha}_R}(s \in \Psi^{nd})} \left(\alpha_S - \alpha_R\right) \end{array}\right].
\end{aligned}$$

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

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

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

$$\Delta\left(\alpha_S, \alpha_R, \widehat{\alpha}_R\right) = p(\widehat{\alpha}_R) A(\widehat{\alpha}_R) + (1 - p(\widehat{\alpha}_R))(\alpha_S - \alpha_R),$$

where we let

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

and we let

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

The derivative of $\Delta\left(\alpha_S, \alpha_R, \widehat{\alpha}_R\right)$ w.r.t. $\widehat{\alpha}_R$ is thus given by

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

In order to prove that $\frac{\partial \Delta(\alpha_S, \alpha_R, \widehat{\alpha}_R)}{\partial \widehat{\alpha}_R} < 0$, it thus suffices to show that $\frac{\partial A(\widehat{\alpha}_R)}{\partial \widehat{\alpha}_R} < 0$,

$$\left[A(\widehat{\alpha}_R) - (\alpha_S - \alpha_R)\right] > 0$$

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

Note first that $\frac{\partial \Pr_{\widehat{\alpha}_R}(s \in \Psi^{nd})}{\partial \widehat{\alpha}_R} = \int_{s_1}^{s_2}(f(s|l) - f(s|h))ds$, which is strictly negative given that we know that $f(s|h) > f(s|l)$ for any $s \in [s_1, s_2]$, recalling that $\widetilde{s} < s_1 < s_2$. It follows immediately that $\frac{(1-\varphi)}{(1-\varphi)+\varphi \Pr_{\widehat{\alpha}_R}(s \in \Psi^{nd})} = 1 - p(\widehat{\alpha}_R)$ increases in $\widehat{\alpha}_R$ and that $\frac{\varphi \Pr_{\widehat{\alpha}_R}(s \in \Psi^{nd})}{(1-\varphi)+\varphi \Pr_{\widehat{\alpha}_R}(s \in \Psi^{nd})} = p(\widehat{\alpha}_R)$ decreases in $\widehat{\alpha}_R$. Second, note that $A(\widehat{\alpha}_R) - (\alpha_S - \alpha_R) > 0$ is a property of equilibrium that we have already established (see proof of Lemma II.C). Third, we now show

that $A(\widehat{\alpha}_R) = E_{\widehat{\alpha}_R}\left[\left(\widetilde{\alpha}_S(s) - \widetilde{\alpha}_R(s)\right)|s \in [s_1, s_2]\right]$ decreases as $\widehat{\alpha}_R$ increases. Note that:

$$
\begin{aligned}
& \frac{\partial\left[\int_{s_1}^{s_2} \frac{\left(\widehat{\alpha}_R f(s|l)+(1-\widehat{\alpha}_R) f(s|h)\right)}{\Pr_{\widehat{\alpha}_R}(s\in\Psi^{nd})}\left(\widetilde{\alpha}_S(s)-\widetilde{\alpha}_R(s)\right) ds\right]}{\partial \widehat{\alpha}_R} \\
= & \int_{s_1}^{s_2} \frac{\left(\begin{array}{c}\left(f(s|l)-f(s|h)\right)\left[\int_{s_1}^{s_2} \widehat{\alpha}_R f(s|l)+\left(1-\widehat{\alpha}_R\right) f(s|h) ds\right] \\ -\left[\widehat{\alpha}_R f(s|l)+\left(1-\widehat{\alpha}_R\right) f(s|h)\right]\left[\int_{s_1}^{s_2}(f(s|l)-f(s|h)) ds\right]\end{array}\right)}{\left[\Pr_{\widehat{\alpha}_R}(s \in \Psi^{nd})\right]^2}\left(\widetilde{\alpha}_S(s)-\widetilde{\alpha}_R(s)\right) ds \\
= & \frac{\left(\begin{array}{c}\left[\int_{s_1}^{s_2} \widehat{\alpha}_R f(s|l)+\left(1-\widehat{\alpha}_R\right) f(s|h) ds\right]\left[\int_{s_1}^{s_2}\left(f(s|l)-f(s|h)\right)\left(\widetilde{\alpha}_S(s)-\widetilde{\alpha}_R(s)\right) ds\right] \\ -\left[\int_{s_1}^{s_2}(f(s|l)-f(s|h)) ds\right]\left[\int_{s_1}^{s_2}\left(\widehat{\alpha}_R f(s|l)+\left(1-\widehat{\alpha}_R\right) f(s|h)\right)\left(\widetilde{\alpha}_S(s)-\widetilde{\alpha}_R(s)\right) ds\right]\end{array}\right)}{\left[\Pr_{\widehat{\alpha}_R}(s \in \Psi^{nd})\right]^2} \\
= & \frac{\left(\begin{array}{c}-\left[\int_{s_1}^{s_2}(f(s|l)-f(s|h)) ds\right]\left[\int_{s_1}^{s_2}\left(\widehat{\alpha}_R f(s|l)+\left(1-\widehat{\alpha}_R\right) f(s|h)\right)\left(\widetilde{\alpha}_S(s)-\widetilde{\alpha}_R(s)\right) ds\right] \\ +\left[\int_{s_1}^{s_2} \widehat{\alpha}_R f(s|l)+\left(1-\widehat{\alpha}_R\right) f(s|h) ds\right]\left[\int_{s_1}^{s_2}(f(s|l)-f(s|h))\left(\widetilde{\alpha}_S(s)-\widetilde{\alpha}_R(s)\right) ds\right]\end{array}\right)}{\left[\Pr_{\widehat{\alpha}_R}(s \in \Psi^{nd})\right]^2} \\
< & \frac{\left(\begin{array}{c}-\left[\int_{s_1}^{s_2}(f(s|l)-f(s|h)) ds\right]\left[\int_{s_1}^{s_2}\left(\widehat{\alpha}_R f(s|l)+\left(1-\widehat{\alpha}_R\right) f(s|h)\right)\left(\widetilde{\alpha}_S(s)-\widetilde{\alpha}_R(s)\right) ds\right] \\ +\left[\int_{s_1}^{s_2} \widehat{\alpha}_R f(s|l)+\left(1-\widehat{\alpha}_R\right) f(s|h) ds\right]\left[\int_{s_1}^{s_2}(f(s|l)-f(s|h))\right]\left[\int_{s_1}^{s_2}\left(\widetilde{\alpha}_S(s)-\widetilde{\alpha}_R(s)\right) ds\right]\end{array}\right)}{\left[\Pr_{\widehat{\alpha}_R}(s \in \Psi^{nd})\right]^2} \\
= & \frac{-\left[\int_{s_1}^{s_2}(f(s|l)-f(s|h)) ds\right]\left(\begin{array}{c}\left[\int_{s_1}^{s_2}\left(\widehat{\alpha}_R f(s|l)+\left(1-\widehat{\alpha}_R\right) f(s|h)\right)\left(\widetilde{\alpha}_S(s)-\widetilde{\alpha}_R(s)\right) ds\right] \\ -\left[\int_{s_1}^{s_2} \widehat{\alpha}_R f(s|l)+\left(1-\widehat{\alpha}_R\right) f(s|h) ds\right]\left[\int_{s_1}^{s_2}\left(\widetilde{\alpha}_S(s)-\widetilde{\alpha}_R(s)\right) ds\right]\end{array}\right)}{\left[\Pr_{\widehat{\alpha}_R}(s \in \Psi^{nd})\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.

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

5.5.2 Proof of Proposition 8

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

5.6 Appendix VI: Joint observation of public signals

5.6.1 Proof of Proposition 11

Step 1 This proves Point 1 of the proposition. Assume without loss of generality that $x \geq y$. Given a 0-signal, the change in disagreement is

$$\omega_0(x,y,p) = (x-y) - \left(\frac{xp}{xp+(1-x)(1-p)} - \frac{yp}{yp+(1-y)(1-p)} \right).$$

Given a 1-signal, we instead have:

$$\omega_1(x,y,p) = (x-y) - \left(\frac{x(1-p)}{x(1-p)+(1-x)p} - \frac{y(1-p)}{y(1-p)+(1-y)p} \right).$$

The expected changes in disagreement from the perspective of agents with priors x and y are given by, respectively:

$$\begin{aligned} V^x(x,y,p) &= (xp+(1-x)(1-p))\,\omega_0(x,y,p) \\ &\quad + (1-(xp+(1-x)(1-p)))\,\omega_1(x,y,p), \qquad (19) \\ V^y(x,y,p) &= (yp+(1-y)(1-p))\,\omega_0(x,y,p) \\ &\quad + (1-(yp+(1-y)(1-p)))\,\omega_1(x,y,p). \qquad (20) \end{aligned}$$

These expressions further simplify to

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

and

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

It can be trivially shown that these expressions are always positive no matter the values of x, y and p, where $V^x(V^y)$ equals to 0 if and only if $x(y) \in \{0, 1, y(x)\}$.

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

$$\frac{\partial^3 V^x(x, y, p)}{\partial y^3} = \frac{6(1-2p)^2(1-p)p}{(1-y-p(1-2y))^4(y+p(1-2y))^4} M, \tag{21}$$

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,

$$M_{|x=y} > 0. \tag{22}$$

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 (22) and the fact that M is linear in x implies that $M > 0$. Consequently, by (21)

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

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

$$y > x : \frac{\partial V^x(x,y,p)}{\partial x} = \frac{(1-y)y(1-2p)^2}{(y-1+p-2yp)(y+p-2yp)} < 0, \tag{23}$$

$$y < x : \frac{\partial V^x(x,y,p)}{\partial x} = -\frac{(1-y)y(1-2p)^2}{(y-1+p-2yp)(y+p-2yp)} > 0, \tag{24}$$

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 12

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 11.1).

Step 2 By (23) and (24) 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 expressions in (19) and (20), 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 11.1, we obtain that there is unique $x^* \in (0, 1/2)$ maximizing $\mu(x, 1 - x)$.