Why Echo Chambers are Useful*

Ole Jann — CERGE-EI

Christoph Schottmüller — University of Cologne and TILEC

October 2, 2023

Abstract

Why do people appear to forgo information by sorting into "echo chambers"? We model a society in which information is dispersed and preferences are polarized. Segregation into small, homogeneous groups can then maximize the amount of communication that takes place, thus making such segregation both individually rational and Pareto-efficient. We examine the optimal communication structure and give sufficient conditions for when it is attainable through endogenous group formation. Players may segregate inefficiently little because their actions have an informational externality. Our framework can be extended to situations with uncertainty, public information, and different network structures.

JEL: D72 (Political Processes), D82 (Asymmetric Information), D83 (Learning, Communication), D85 (Network Formation and Analysis)

Keywords: asymmetric information, echo chambers, polarization, debate, cheap talk, information aggregation

*Jann: CERGE-EI, a joint workplace of Charles University and the Economics Institute of the Czech Academy of Sciences, Prague; ole.jann@cerge-ei.cz. Schottmüller: Department of Economics, University of Cologne; c.schottmueller@uni-koeln.de. We are grateful for helpful comments by Rachel Bernhard, James Best, Benjamin Blumenthal, Ben Brooks, Vince Crawford, Marcelo Fernandez, Ben Golub, Sanjeev Goyal, Bård Harstad, Paul Klemperer, Vasily Korovkin, Nenad Kos, Meg Meyer, Kirill Pogorelskiy, David Ronayne, Larry Samuelson, Bill Sandholm, Karl Schlag, Jakub Steiner, Peter Norman Sørensen, Kyle Woodward and Peyton Young, as well as audiences at Bar-Ilan, Bielefeld, Bonn, CERGE-EI, Cologne, Copenhagen, Groningen, Konstanz, Munich, Oxford, Royal Holloway, Sussex, Vienna, Warwick and Wisconsin-Madison and at EEA 2018 (Cologne), TTW 2018 (Northwestern), ESEWM 2018 (Naples), ECOP 2019 (Bologna), ESWM 2020 (Milan/virtual), ASSA 2021 (virtual), EPSA 2022 (Prague) and SAET 2022 (Canberra). This research was supported by Charles University Primus project 20/HUM/019 and GAČR grant 22-33162M.

Large parts of society are organized around the non-market exchange of information: People gather around breakfast and dinner tables, in meeting rooms, committees, cafés and bars, while keeping in touch with friends, co-workers and strangers through electronic messaging and social media. But while people constantly seek out others' views and knowledge, they do not seek out a wide range of different viewpoints. Instead, they tend to segregate into homogeneous communities and limit the number of views they are exposed to.[1]

This poses a theoretical puzzle: If people put so much energy into seeking and exchanging information, why do they artificially limit both the diversity and the amount of information available to themselves? Is it simply because of irrational biases or fear of confrontation, or can there be informational reasons? It also raises the urgent question of whether the resulting segregation in fact limits the spread of useful information and thus hurts society, and whether policy should seek to influence who communicates with whom.[2]

In this paper, we consider a society that faces two problems: Dispersion of information and polarization of preferences. Dispersed information means that no single individual knows enough about the world to make very successful choices. Polarized preferences mean that even if everybody had perfect knowledge about the world, people would still disagree on which action each person should take.[3] We develop a general framework to model how people rationally communicate within groups in the presence of information dispersion and preference polarization, and how they sort into groups while anticipating what communication within the groups will be like.

Our analysis shows that segregation into small, homogeneous groups can result from rational choice and *maximize* the amount of information available to any single individual. In fact, such segregation can be efficient and even Pareto-optimal for society. The optimal amount of segregation increases in the degree to which preferences are polarized, relative to the informational disagreement among individuals.

Why is that? Information dispersion and preference polarization work in opposite directions: The former makes individuals curious about learning each other's information so they can make better choices; the latter gives them an incentive to misrepresent their own information in order to influence others' choices. If the polarization of preferences increases in any given group, individuals become more and more interested in misleading

[1] See, for example, studies on segregation in blogs (Lawrence et al., 2010), on Facebook (Del Vicario et al., 2016; Quattrociocchi et al., 2016), on Twitter (Barberá et al., 2015) and in online and offline contexts in general (Gentzkow and Shapiro, 2011).

[2] Consider, for example, claims that echo chambers are "dangerous" (Grimes, 2017) and have "Balkanised society" (Itten, 2018), as well as played a role in populist insurgencies in Western democracies such as the "Brexit" referendum (Chater, 2016) or the rise of Donald Trump (Hooton, 2016).

[3] We understand "polarization" as a measure of distribution, similar to Esteban and Ray (1994) and the following literature. To avoid misunderstandings, we will use "polarization" only when talking about exogenously given preferences – i.e. a primitive of our model – and not when referring to information or beliefs.

each other and truthful communication becomes harder – even though individuals are still just as much interested in learning from one another. This means that in a highly polarized society in which everybody tries to learn from everyone, very little truthful communication may be possible because everyone also tries to mislead (almost) everyone. If we are able to split this society into segregated groups, however, people within each one of these groups may be a lot less polarized, so that for the purposes of their communication the effects of information dispersion now dominate those of preference polarization. While segregation into echo chambers limits the amount of *potential* communication, it makes *actual* communication possible.

If it is left to individuals to decide with whom to communicate, an efficient allocation may nevertheless be achievable and we give sufficient conditions for when that is the case: If individuals care sufficiently about the decisions of others, or if the polarization of preferences is very large or very small compared to the dispersion of information. But people will also sometimes segregate *too little* compared to what a social planner would choose. This is because choosing a group to communicate with has an informational externality: A person may choose a group from which she can learn the most, without fully internalizing how much others learn from her. If a society consists of two large groups with different preferences (a situation that is often synonymous with "polarization"), it can only ever be the case that in the best equilibrium, people either segregate optimally or they segregate too little.

Our theory (and our title) should not be understood to mean that echo chambers are unambiguously good for society. What we do in this paper is to identify and isolate a mechanism by which they can be useful and increase welfare. This may, in many instances, be outweighed by ways in which they can be detrimental, some of which we discuss in section 6.4. Overall, our model calls for a nuanced view: Division into small, homogeneous groups can facilitate honest debate and real mind changes just as it can narrow world views and shut out crucial information. Which of these mechanisms dominates may not be apparent from the simple fact that there are groups, and can only become clear when we understand the structures of polarization and mistrust that underlie the sorting behavior.

We also suggest that the main business model of social networking sites could be understood as providing the infrastructure for people to sort in the way that they want – and not, in fact, simply to connect them, as is commonly assumed. We discuss this idea in more detail in section 6.1.

This paper has three main contributions. First, in sections 1 to 3 we develop a tractable framework to analyze strategic communication among n players, where everybody has different information and different preferences and can talk and listen to everybody else. The framework can accommodate endogenous choice of communication structures, different communication protocols, different types of uncertainty and other modifications. Second, in section 4 we use this framework to develop the theory described in the previ-

ous paragraphs. Finally, in sections 5 and 6 we show how our model extends to various contexts, and we discuss how it can be applied to real-world situations.

An introduction to our modeling framework We analyze a model in which a number of individuals face aggregate uncertainty and have heterogeneous preferences. These individuals sort into groups, communicate within these groups, and finally each chooses an action. The state of the world and each person's preference are real numbers; a person's ideal point is simply the sum of the state of the world and her preference. Each person wants all players' actions, including her own, to be as close as possible to her own ideal point, and her payoff is concave in the distance between anyone's action and her ideal point. People may therefore take different actions based on differences in information and in preferences. Differences in either dimension are sufficient for disagreement, i.e. people would choose different actions even if they all had the same information but different preferences or vice versa.

We assume that people's preferences are common knowledge, though we relax this assumption in an extension. Before choosing whether (and what) to communicate, each person privately receives a binary signal about the state of the world. We make the simplifying assumption that each person's signal contains information about the state of the world, but no information about the information of others. Intuitively, different people observe different aspects of the world, and what one person observes does not tell her anything about what any other person knows. This assumption allows us to develop a simpler and more tractable analysis than richer models in the literature.[4]

We assume that "cheap talk" communication takes place in disjoint "rooms", where each statement by an individual can be heard by anyone else within the room, but not by people in other rooms. If a person now finds herself in a room with a mixed group of others, she faces a trade-off between wanting to correctly inform those who have preferences close to her own (as that will bring their action closer to her own ideal point), and wanting to mislead those who have very different preferences. If most of her audience has a much lower preference parameter than her, for example, she would want to make them believe that her signal says that the state of the world is a high number, to counteract the fact that her audience will always choose an action that she deems too low.

We show that the question of who tells the truth and who babbles in equilibrium in any given room has a simple solution and only depends on the difference between a person's preference and the average preference of her audience (theorem 1). This allows us to easily compute how much communication can take place in the most informative equilibrium in any room. Following the backwards-induction logic, we can then analyze how a social planner would choose to allocate people to rooms, and how people can allocate to rooms

[4]In the supplementary material, we show that our main arguments are robust to using various different assumptions and modeling techniques.

in equilibrium.[5]

We show that the most informative equilibrium in any room is always in pure strategies, and that hence we can count information in discrete "pieces": Receiving a signal, or learning another player's signal through communication, are each equivalent to getting one "piece" of information (corresponding to 1 bit in Shannon entropy). Despite the fact that people face aggregate uncertainty, have different preferences and have concave preferences about their own actions and the actions of others, we can show that all payoffs (and hence also welfare) can simply be expressed in the integer amount of pieces of information that each player has after room choice and communication have taken place (proposition 1).

We consider two different types of preference polarization. First, we analyze a society that consists of people with two types of preferences. In section 4.2, we completely characterize the optimal room allocation for all possible group sizes and magnitudes of polarization in this case. We show that either the welfare-optimum is an equilibrium of the room-choice game, or people segregate *too little* in the welfare-optimal equilibrium.

In section 4.3 we consider a more general specification of polarization as "clustering" around certain values. We introduce a parameterization of polarization in such a model and show the following result: If the relative polarization of preferences is large, full segregation by preferences is always welfare-optimal and an equilibrium, whereas integration is optimal and an equilibrium for low polarization (theorem 3).

Finally, our analysis allows us to disentangle the welfare effects of polarization and segregation. An increase in a society's polarization leads to an increased desire for segregation as well as a welfare loss in equilibrium. An observer may hence be tempted to conclude that segregation itself has caused the welfare loss, but we can show that the opposite is the case: Polarization lowers welfare (proposition 2), but segregation actually mitigates the corrosive effects of polarization. Not allowing people to segregate in the presence of polarization would lower welfare. We can hence see segregation into echo chambers as not just an individually rational action, but as society's decentralized countermeasure against the welfare losses caused by polarized preferences.

In section 5, we show that our model remains tractable if we consider several extensions to situations (i) in which there is public as well as private information, (ii) in which preferences are uncertain, and (iii) in which people can unilaterally choose whom to "follow", rather than sorting into rooms – similar to what happens on social media sites.

Relation to other research Our work is related to several methodological approaches, and ties into a wider-ranging literature on segregation, isolation and echo chambers.

As the basis of our analysis of communication, we develop a tractable model of many-to-many cheap-talk. Our geometrical solution avoids much of the exponential complexity

[5]Following the usual convention in the literature on strategic communication, we only consider the most informative equilibrium in each room and ignore less informative babbling equilibria.

that usually appears in models with multiple senders or receivers.[6]

Galeotti et al. (2013) analyze 1-to-1 communication among many agents in networks. Agents face a decision problem similar to ours, but since each agent's signal is informative about each other agent's signal, equilibria are not well-ordered in the way that our lemma 1 establishes. Even when mixed equilibria are excluded by assumption, the analysis does not consider the type of "room choice" question of our paper (and such an analysis would not be possible in this framework).

The model by Galeotti et al. (2013) has been used in the political science literature to study the formation of alliances (Penn, 2016), the existence and stability of factions within parties (Dewan and Squintani, 2016), and the effect of decision making authority on deliberation (Patty, 2022). The latter study is closest to our paper, as Patty also considers communication within a room. However, the room structure is restricted and strategic room choice is not analyzed – a simpler and more general analysis of Patty's questions might be possible when using our framework.

Hagenbach and Koessler (2010) consider 1-to-1 communication among players in a network who then play a type of coordination game. As in our model, each player's signal is uninformative about any other's signal. Mixed equilibria are considered too complex and excluded from the analysis (cf. their footnote 4), whereas in our framework we can consider all mixed equilibria but show that they are strictly less informative than the pure equilibria we focus on. Their theorem 1 proves existence of a network in which each player sends truthful messages to a set of receivers whose average bias is close enough to her own. While this may look similar to our extension in section 5.3 (or even our theorem 1), it results from the wish to coordinate the actions of various receivers, unlike in our framework.

In contrast to these and other studies, our framework also allows us to measure the "pieces of information" that each players has and show that these are a sufficient statistic for payoff and welfare (proposition 1). This, together with the ability to strictly order equilibria by informativeness, then permits the analysis of room choice both from the perspective of the individual and the social planner, which we see as our main economic contribution.

The welfare analysis of room choice in our model can also be seen as an information design problem: How can an information designer induce information exchange between several agents if commitment to a disclosure rule (as in the literature on Bayesian Persuasion) is not available? The right construction of mixed groups can induce truth-telling. Rooms endogenously create costs to lying in our model, which is the main instrument of discipline in Kartik 2009. They induce truth-telling even though different senders' infor-

[6]Some of our main arguments are not dependent on this particular setup and can be derived in a more classical cheap-talk setting akin to Crawford and Sobel (1982), as we show in the supplementary material.

mation is orthogonal to each other and there hence exists no mechanism (as in e.g. Krishna and Morgan 2001) to elicit information by playing senders off against each other. Some of the results of our paper hence allow us to analyze "communication design" without commitment, and perhaps even without a designer.

In section 5.2, we show that uncertainty about preferences has a corrosive effect on truth-telling. In contrast to earlier works such as Morgan and Stocken (2003), we allow for uncertainty of the sign as well as the size of the bias, which may be distributed continuously or discretely. Uncertainty about the bias can help or hurt information transmission (and does not necessarily help as in Li and Madarász, 2008). Our results and methods generalize without loss to large groups of players and general distributions of biases. To our knowledge, we are the first to generally analyze how bias uncertainty influences whom people want to associate and communicate with, and how it increases the appeal and the usefulness of segregation.

The debate about echo chambers has been given urgency by several studies and popular treatises on how the internet changes the way societies debate. Sunstein (2001, 2017) prominently makes the case that the internet has been increasing ideological segregation and that this endangers democracy. Gentzkow and Shapiro (2011) point out that the segregation of "offline" interactions is larger than that of "online" interactions. But while such offline segregation can happen simply because we live close to people who are like us in many socio-economic aspects, segregation on the internet is driven more directly by choice. Our model allows us to analyze the informational effects of any kind of segregation or integration, as well as predicting which communication structures arise from individual optimizing behavior, and whether they are socially optimal.

Several recent papers also mention the idea of echo chambers while focusing on exogenous news sources (or algorithm designers who do not care about the informational content of shared messages), e.g. Che and Mierendorff (2019); Martinez and Tenev (2020); Acemoglu et al. (2021). In our paper, the interplay between strategic senders and receivers is at the heart of our analysis.

Even where echo chambers would have negative consequences that are not in our model, the effects that we describe would have to be reckoned with, and a nuanced discussion of how much segregation is optimal in debate is necessary. Most importantly, we argue that those who see ideological segregation as the ruin of societies are focusing on a symptom, not the cause. Polarization of preferences and mutual mistrust are doing the real damage; informational segregation can be a rational behavior that mitigates the harm they do.

1. Model

There is an unknown state of the world $\theta = \sum_{i=1}^{n} \theta_i$. Each θ_i is independently drawn to be 0 or 1 with equal probabilities, so that θ is binomially distributed on $\{0, 1, \ldots, n\}$. n individuals each make an observation about the state. In particular, individual i receives a private signal $\sigma_i \in \left\{\sigma^l, \sigma^h\right\}$ of accuracy p about θ_i, i.e. $Pr\left(\sigma_i = \sigma^h | \theta_i = 1\right) = Pr\left(\sigma_i = \sigma^l | \theta_i = 0\right) = p > 1/2$. Before observing his signal, a player can access one of n "rooms". There are no costs to entering a room, and rooms have no capacity constraints – but each player can only be in exactly one room. After observing his signal, a player sends a cheap-talk message $m_i \in \{m^l, m^h\}$ that is received by all players in the same room.[7] Finally, each player takes an action a_i.

The payoff of player i is

$$\begin{aligned} u_i(a, b_i, \theta) &= -(a_i - b_i - \theta)^2 - \alpha \sum_{j \neq i} (a_j - b_i - \theta)^2 \\ &= -\left(a_i - b_i - \sum_{k=1}^{n} \theta_k\right)^2 - \alpha \sum_{j \neq i} \left(a_j - b_i - \sum_{k=1}^{n} \theta_k\right)^2 \end{aligned} \tag{1}$$

where a denotes the vector of actions of all players and $b_i \in \mathbb{R}$ is a commonly known "bias" of player i. That is, actions of all players affect i's payoff, and i would like that all players choose the action $b_i + \theta$. We can hence think of b_i as the *preferences* of the players, whereas θ_i is the aspect of the world that player i has *information* about. The parameter α measures the relative weight players assign to other players' behavior – in other words, the sensitivity of i's payoff to the actions of other player. If $\alpha = 0$, i only cares about his own decision; if $\alpha = 1$ then every other player's decision is just as important to i as his own decision. Players maximize their expected payoff.

The timing of the game is:

1. All biases b_i become common knowledge.
2. Players simultaneously decide which room to enter.[8]
3. Players privately observe their signals σ_i. Players simultaneously send messages m_i that are observable by everyone in the same room R_i.
4. Players simultaneously take actions a_i; payoffs are realized.

[7] Our main analysis considers binary signals and messages, but the supplementary material shows that our main results are robust to the introduction of an arbitrary finite number of states and signals.

[8] As we are only interested in equilibrium, this is equivalent to imagining a stage in which players can enter rooms, observe the composition of all rooms and then switch rooms if they like. This stage would end when no player wants to switch anymore. In this interpretation, the assumption that every player can see the composition of any room is for clarity of exposition, but for equilibrium analysis it is without loss of generality since every player will correctly anticipate the room choices of others in equilibrium and biases are common knowledge.

More formally, let B_{R_i} be the vector of biases in room R_i and denote by $\mathbb{B} \in \Re \cup \cdots \cup \Re^n$ the set of all such bias vectors. A messaging strategy in stage 2 is then a (potentially stochastic) map $m_i : \mathbb{B} \times \{\sigma^l, \sigma^h\} \to \Delta\left(\{m^l, m^h\}\right)$.[9] Stage 3 actions assign to each combination of own signal σ_i and messages sent in R_i an action in $\Re$.[10]

We analyze the model by backwards induction: First we characterize the optimal choice of action given messages, then the optimal choice of message given a room allocation, and then we analyze the game in which players choose which room to enter. The solution concept used throughout is Perfect Bayesian Equilibrium.[11]

2. Equilibrium Behavior Within a Room

2.1. Choice of Action

We can immediately see that only the first part of expression 1 matters for determining i's optimal action a_i^*. The first-order condition yields

$$a_i^* = b_i + \mathbb{E}[\theta | m_{R_i}, \sigma_i] = b_i + \sum_{j=1}^{n} \mathbb{E}[\theta_j | m_{R_i}, \sigma_i], \tag{2}$$

where m_{R_i} denotes the profile of messages sent in room R_i. In words, the optimal action is simply i's bias plus his expectation of the state, conditional on his own signal and on the messages he has received.

In the following, we will denote by $\mu_{ij} = \mathbb{E}_i[\theta_j | m_{R_i}, \sigma_i]$ i's expectation about θ_j, so that expression (2) becomes $a_i^* = b_i + \sum_{j=1}^{n} \mu_{ij}$.

2.2. Choice of Message

Now that we have established each agent's optimal action choice given expectations $(\mu_{ij})_{j=1}^n$, we can consider the optimal choice of message. For this, we focus on a single room, and consider the equilibria of the cheap talk game in this room. This means that when we speak of "equilibrium" in this section, we mean the equilibrium in a specific room (with a given set of members with given biases), and not the overall equilibrium of the game. We can do this because once players have sorted into rooms, the messages in other rooms are unobservable and the actions of players in other rooms are irrelevant to a player's optimization problem. Hence, an equilibrium of the subgame after room choice can be disassembled into one equilibrium of the cheap talk game for each room.

[9] This means that mixing is allowed, but as we will show below there exists a more informative pure-strategy equilibrium for every mixed equilibrium.

[10] In principle, strategies could also depend on the composition of other rooms. However, ignoring this possibility is without loss of generality as a player cannot gain from such a dependence (given that messages are not payoff relevant).

[11] For the equilibria that we are interested in, all messages occur in equilibrium and there is no hidden information at the time that people choose rooms, so that our results are insensitive to assumptions about off-path beliefs.

Definition 1. *We call a messaging strategy m_i ...*

- babbling *if m_i is independent of i's observed signal σ_i and therefore nobody learns anything payoff relevant from m_i.*
- truthful *if $m_i(\sigma^l) = m^l$ and $m_i(\sigma^h) = m^h$.*
- lying *if $m_i(\sigma^h) = m^l$ or $m_i(\sigma^l) = m^h$.*
- pure *if m_i is either babbling or truthful, so that m_i is either perfectly uninformative or perfectly informative about σ_i.*
- mixed *if for some signal σ^k, $k \in \{l, h\}$, both messages are sent in equilibrium with positive probability and the strategy is not babbling.*

The cheap talk game within a room can – as usual – have several equilibria. For each player i, there always exists an equilibrium in which i babbles. (Consequently, there also always exists an equilibrium in which all players babble.) In line with the cheap talk literature, we will focus on the most informative equilibrium.[12] The following lemma implies that the most informative equilibrium is in pure strategies.

Lemma 1. *Let $(m_1, \ldots, m_n)$ be equilibrium strategies. If m_i is a mixed strategy, then there also exists an equilibrium with strategies (m_i^t, m_{-i}), where m_i^t is the truthful strategy.* *(Proof on page 29.)*

What is the intuition for this result? Imagine an equilibrium in which player i mixes between messages after observing signal σ^h. That is, i is indifferent between sending a high message that induces high actions by the other players in his room and a low message that induces lower actions by the players in his room. This means that the actions induced by m^l are somewhat too low from i's point of view and the actions induced by m^h are somewhat too high. Note that i will always send the low message in case he observes a low signal in such an equilibrium because the actions i would like the other players to take are increasing in his signal. Consequently, a high message perfectly reveals i's high signal. Now consider switching to an equilibrium in which i uses the truthful strategy. When i now observes a high signal, sending the high message will lead to exactly the same actions by the other players as in the original equilibrium. However, sending a low message will lead to a lower expectation of the other players than in the original equilibrium and therefore to lower actions by the other players. Player i will then strictly prefer the high message as these lower actions are too low (given that i was indifferent in the original equilibrium).

[12] The concept of "most informative" equilibrium is not necessarily well defined in multi-sender cheap talk games. However, the following paragraphs will make clear that this concept is straightforward in our model.

The main implication of lemma 1 is that the most informative equilibrium is always in pure strategies: Starting from any mixed equilibrium we can switch the mixing players one by one to truthful strategies and the resulting strategy profile remains an equilibrium. This new equilibrium is more informative as the truthful strategy is most informative (in the Blackwell sense) and therefore best for the receivers.

Corollary 1. *The most informative equilibrium in a room is always in pure strategies.*

We can now characterize the most informative equilibrium. Intuitively, we might expect that the distance of b_i to the biases of the other players is crucial for i's incentive to tell the truth, since i becomes more interested in misleading the other players if their biases differ by a lot. We formalize this intuition and specify the most informative equilibrium in the following result, which is illustrated by figure 1:

Theorem 1. *Let $\overline{b} = \frac{\sum_{k \in R} b_k}{n_R}$ be the mean bias of players in room R. In the most informative equilibrium in this room, a player i tells the truth if and only if*

$$b_i \in \left[\overline{b} - \frac{n_R - 1}{n_R}(p - \frac{1}{2}), \overline{b} + \frac{n_R - 1}{n_R}(p - \frac{1}{2}) \right]$$

and babbles otherwise. (Proof on page 30.)

The size of the truth-telling interval increases in both n_R, the number of people in the room, and p, the precision of individual signals. The increase in n_R can be seen as a correction term: What really matters for the motivation of a player is his distance from the average bias of the *other* players in the room. Hence, if we write a symmetric interval around $\overline{b}$ (which includes b_i), we have to add this correction.[13] When p, the precision of signals, is higher, each truthful signal causes a greater change in the actions of others. People communicate truthfully if they are disciplined by the danger of influencing others' actions too much by lying. Hence, if p is higher, this disciplining force is stronger and a player can be further away from the average bias of others and still tell the truth.

3. Room Choice

We can now analyze room choice, under the assumption that the most informative equilibrium will be played in any room. We will first derive some results about the welfare-optimal room allocation, and then analyze under which conditions this optimal room allocation is in fact an equilibrium.

[13] Intuitively, one could also think that the average room bias "stabilizes" for larger n_R, so that a player can be further away from the average room bias and have the same distance from the average bias of other players in the room.

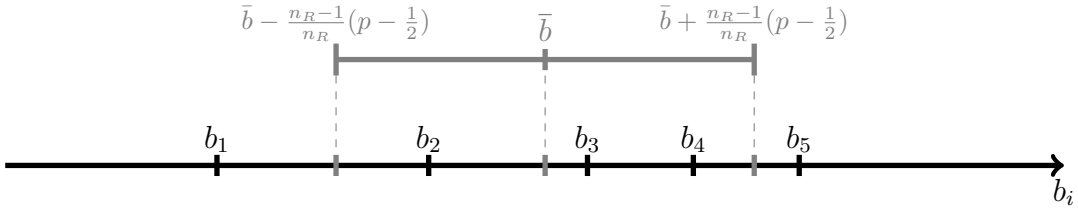

Figure 1: Finding the most informative equilibrium in a room consisting of players 1 to 5. We find the average bias and construct a symmetric interval around it. Players 1 and 5 babble in the most informative equilibrium, since their biases are too far from $\overline{b}$. Players 2, 3 and 4 tell the truth.

Given the expression for individual payoff (1), overall welfare in the model is given by

$$W(a,b,\theta) = \sum_{i=1}^{n} u_i(a,b_i,\theta) = -\sum_{i=1}^{n}\left(a_i - b_i - \sum_{k=1}^{n}\theta_k\right)^2 - \alpha\sum_{i=1}^{n}\sum_{j\neq i}\left(a_j - b_i - \sum_{k=1}^{n}\theta_k\right)^2.$$

This expression, of course, is not yet very helpful in trying to compare different room allocations. However, we can show that in our model, welfare can simply be expressed in terms of the aggregate amount of information that is held by all players after communication has taken place.

This is because, as payoffs are quadratic, we can additively separate a player's payoff into (i) losses through preference differences and (ii) losses from variance due to lack of information. In an equilibrium of the messaging game, the former losses are unavoidable, but the latter can be mitigated by increasing the flow of information between players. We can measure this flow simply by counting the pieces of information that each player has when making their decision.

Consider the information that is available to a single player. A player always receives his own signal σ_i. We can call this *one piece of information*. Assume that i also receives truthful signals from two other players; then we can say that i has three pieces of information about θ. Let $\zeta_i \in \{1,2,\ldots,n\}$ be the number of pieces of information available to player i which are either his own signal or truthful messages from other players. Given that each σ_j has two possible values (high or low), ζ_i in fact measures player i's information in *bits*, the unit of information. The following result shows that all welfare comparisons reduce to informational accounting in bits:

Proposition 1. *If the most informative equilibrium is played within all rooms,*

1. *player i's payoff is given by*

$$U_i = -\alpha\sum_{j\neq i}\{(b_j - b_i)^2\} - 1/4\,[n + \alpha(n-1)n] + (1/4 - p(1-p))\left[\zeta_i + \alpha\sum_{j\neq i}\zeta_j\right] \quad (3)$$

which is a linear and increasing function of $\zeta_i + \alpha\sum_{j\neq i}\zeta_j$.

2. welfare is given by

$$W = -\alpha \sum_{i=1}^{n} \sum_{j \neq i} \{(b_j - b_i)^2\} - \frac{1}{4} n^2 \left[1 + \alpha(n-1)\right] + (p - \frac{1}{2})^2 (1 + \alpha(n-1)) \sum_i \zeta_i$$

which is a linear and increasing function of $\sum_i \zeta_i$.

(Proof on page 31.)

In each RHS term, the first part describes the losses that occur because of payoff differences, the second part describes the (hypothetical) loss that would occur if no player had any information, and the third part describes the "gains" (compared to this hypothetical loss) from the information that becomes available to the agents through their own signals and within-room communication. The first and second part do not depend on room allocations (and hence are simple functions of model parameters), whereas the third gives us the result that individual payoffs and welfare are linear functions of the number of pieces information that are available to each player.

This redefines i's choice of room in purely informational terms: When choosing a room, i wishes to maximize a weighted sum of his own information (after communication) and that of other players. When he considers switching from, say, room R_A to R_B, i will consider how much more he can learn in room R_B, as well as how much more or less the other people in both rooms will learn after his switch. How exactly i is willing to trade off these informational effects against each other (and hence the size of a player's "informational externality") depends on α.

If $\alpha = 1$, there is no informational externality and the room choice that is optimal for a given player also maximizes welfare. This leads to the following corollary:

Corollary 2. *If* $\alpha = 1$*, the welfare-optimal room allocation is also an equilibrium of the room choice game.*

Proposition 1 means that we can quickly compare the welfare of any two room allocations. Consider, for example, the room allocation in figure 1. Having everybody in the same room generates 17 pieces of information: 3 players have 3 pieces of information each, while two players (those who babble) have 4 pieces each. Would it be possible to improve on this allocation? We can immediately see that this cannot be achieved by splitting players up into two rooms with 3 and 2 players, respectively: Even if everybody in these rooms was telling the truth, only $3^2 + 2^2 = 13$ pieces of information would be produced. The same is true for splitting them into a higher number of even smaller rooms. But even if we somehow could get 4 people in one room to tell the truth by putting one of the players into a separate room, the total number of pieces of information would be $4^2 + 1 = 17$ – the same as with full integration. Hence the room allocation shown in the figure is welfare-optimal.

Of course, we may often not be able to make such quick deductions and might have to consider many possible room allocations before concluding which one is optimal. This problem gets more complex as n grows, since the number of possible partitions of a set (given by the Bell sequence) grows quite rapidly. However, we derive general results on optimal and equilibrium room allocations in the next section.

4. Polarization and Segregation

We have now shown that the messaging problem inside each room has a simple geometrical interpretation, and that the room choice game reduces to a problem in which all players wish to minimize a weighted sum of their own uncertainty and that of the other players. In this section, we will use these results to draw a connection between the polarization of players' preferences, and the question of which room allocations are optimal, and which allocations can be achieved in equilibrium.

We will begin by giving a simple, non-technical example in which segregation is both efficient and an equilibrium. We then generalize the intuitive insights from this example to all possible models in which there are two bias types. Some insights from this model can be generalized again to all conceivable generic bias configurations with an arbitrary number of biases and players, and our framework can be used to address specific bias distributions within this general space. Finally, we show that the welfare effects of polarization work despite segregation, not through segregation.

4.1. A Non-Technical Example

Consider a set of biases as in panel (i) of figure 2: A group of 6 players, 3 of whom have relatively small biases, while the other 3 have relatively large biases. If all players are within the same room (panel i), the truth-telling interval within this fully integrated room does not cover any of the players' biases, which means that in any equilibrium none of them reveals any information. The number of pieces of information generated is 6.

Suppose the players segregate by bias type into two separate rooms – see panel (ii). The truth-telling interval in both rooms covers all the players in the respective rooms, which means that all players reveal their information truthfully. In each room, 9 pieces of information are generated, which means that overall this allocation generates 18 pieces of information.

Is this segregation an equilibrium? We can consider the most profitable deviation of player 3 (which is symmetric to the most profitable deviation of player 4 and better than the best deviations of any other players) – see panel (iii). If player 3 moves into the other room, he will move the average in this room so that players 5 and 6 no longer tell the truth in any equilibrium. He himself also does not tell the truth anymore, so that his move completely deprives society of the information of players 3, 5 and 6. (The lengthening of the truth-telling interval that results from 3's move is not enough to compensate for the

change in average bias.) The resulting room allocation generates $2^2 + 4 + 3 = 11$ pieces of information, which clearly leads to lower welfare. It is also inferior for player 3, since he now has 2 pieces of information (his own and the message from player 4) instead of 3, so that his payoff decreases. Hence this deviation is not optimal for player 3, and no player has a profitable deviation from two segregated rooms – which means that this allocation is not only welfare-optimal, but also an equilibrium.

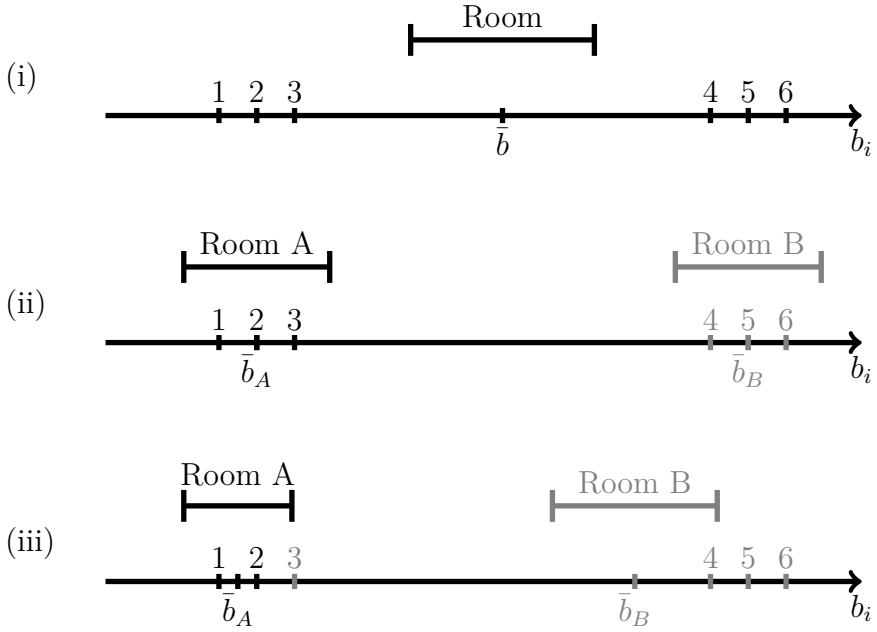

Figure 2: Truth-telling intervals for (i) the fully integrated room, (ii) two segregated rooms, (iii) player 3's best deviation from the segregated room.

4.2. Bipolar Polarization

We now focus on the case where there are two bias groups, i.e. $b_i \in \{0, b\}$ for some $b > 0$, where n_0 individuals have bias 0 and $n_b\,(= n - n_0)$ have bias b. This "bipolar polarization" is often used synonymously with the word polarization. Our results allow us to fully characterize, for the full parameter space: (i) the welfare-optimal room allocation, (ii) whether it is an equilibrium and (iii) if it is not an equilibrium, what the welfare-maximizing equilibrium looks like. This full result is given in theorem 2 below. Since this result has to consider several case-distinctions and is hence somewhat technical, we will first summarize the structure of optimal and equilibrium room allocations with two general results. A step-by-step discussion and derivation of the results is in appendix B.

The first result is that segregation is optimal and an equilibrium if polarization is high relative to differences in information (i.e. if b is large), and full integration is optimal and an equilibrium if polarization is low (if b is sufficiently low):

Result 1. *If all $b_i \in \{0, b\}$ and b is very small (large), the welfare-optimal room allocation is that all players are in the same room (are fully segregated by types); this optimal room allocation is also an equilibrium of the room-choice game.*

We can intuitively consider the two cases. In the first case, all players will send truthful messages if they are all in the same room, since if $\overline{b} = \frac{n_b b}{n} < \frac{n-1}{n}(2p-1)$ then the condition of theorem 1 is trivially fulfilled. Clearly, if b is small enough for this inequality to hold, a fully integrated room will be both welfare-maximizing and an equilibrium room allocation. At the other extreme, consider the case where the presence of one player of bias b in a room containing all n_0 players with bias 0 will lead to babbling by all players in the room. This is the case if and only if $b > n_0(p - \frac{1}{2})$. In this case (and if the analogous condition for n_b is fulfilled), any room containing players of both bias types will lead to babbling. Segregating the two groups is consequently both welfare optimal and an equilibrium.

For intermediate levels of polarization, the welfare optimal room allocation need not be an equilibrium. Perhaps counterintuitively, this will only occur because in the welfare-optimal equilibrium, the groups do not segregate enough:

Result 2. *If all $b_i \in \{0, b\}$ and the welfare-optimal room allocation is not an equilibrium, then the welfare-optimal equilibrium allocation involves too little segregation, i.e. welfare could be improved by moving players from mixed rooms into rooms that contain only their own bias type.*

Intuitively, there can be two types of informational externalities that players may ignore (if α is sufficiently small) and which cause them to segregate too little, compared to how a social planner would allocate them to rooms. To explain these externalities, we will first introduce the following theorem, which exhaustively considers all possible cases:

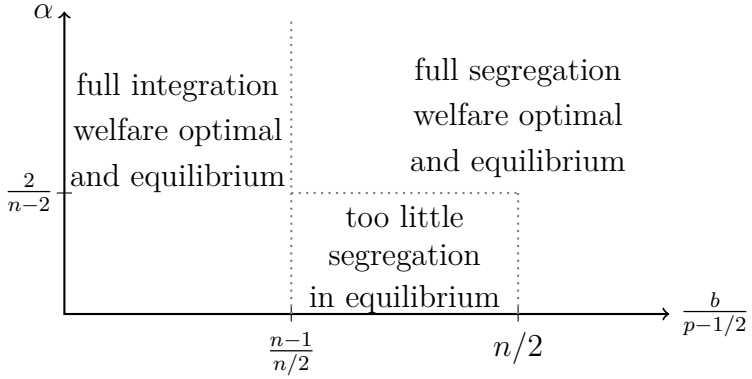

Figure 3: Welfare and equilibria for equally sized bias groups. The horizontal axis describes how polarized preferences are, relative to the difference in information between players.

Theorem 2. *Let all $b_i \in \{0, b\}$ and $n_0 \geq n_b$ (where the latter assumption is without loss of generality).*

(Welfare-optimality) *The welfare maximizing room allocation is as follows:*

1. *For $b/(p-1/2) \leq (n-1)/n_0$, all players are in one room (where their messages are truthtelling).*

2. *For $(n-1)/n_0 < b/(p-1/2) \leq (n-1)/n_b$, let $n_{m0} = \lfloor (n_b - 1)/(b/(p-1/2) - 1) \rfloor$. Then,*

   (a) *it is optimal to have all players in one room (where those with $b_i = 0$ are truthtelling and those with $b_i = b$ are babbling) if $(n_b + n_{m0})^2 + (n_0 - n_{m0})^2 \leq n_0(n_0 + n_b) + n_b$ and $n_{m0} \geq n_b$,*

   (b) *it is optimal to have one room with $n_0 - n_{m0}$ players with $b_i = 0$ and another room that contains all other players (in which everyone is truthtelling) otherwise.*

3. *For $b/(p-1/2) > (n-1)/n_b$, let $n_{mb} = \lfloor (n_0 - 1)/(b/(p-1/2) - 1) \rfloor$. Then,*

   (a) *it is optimal to have one room with $n_b - n_{mb}$ players with $b_i = b$ and another room with all other players if $n_0(n_0 + n_{mb}) + n_{mb} + (n_b - n_{mb})^2 \geq n_0^2 + n_b^2$*

   (b) *full segregation is optimal otherwise.*

(Equilibrium) *The welfare optimal room allocation is an equilibrium in cases (1), (2a) and (3a). The welfare optimal room allocation is also an equilibrium*

- *in case (2b) if*
$$\alpha \geq \frac{2n_{m0} - n_0 + 1}{n_0 - 2n_{m0} - 1 + n_b^2 + n_b(n_{m0} - 1)}$$
- *in case (3b) if $\alpha \geq (1 + n_0 - n_b)/(n_b - 1)$.*

If the welfare optimal room allocation is not an equilibrium, then the welfare optimal equilibrium features too little segregation: In case (2b), the equilibrium has all players in a single room while the welfare optimal allocation has two rooms (only one of which contains both types of players). In case (3b), full segregation is welfare-optimal but not an equilibrium. (Proofs and a step-by-step derivation of the results in part B of the appendix.)

Figures 3 and 4 depict these results and the exact relationship between welfare-optimality and equilibrium for equally-sized bias groups and one case of different-sized bias groups.

This result shows that information externalities can take two possible forms:

- In case (2b), a player destroys information in the room that she *enters*. The welfare-optimum for this case is basically achieved by starting with one fully integrated room (in which the b-types are in the minority and all babble) and removing 0-types to a separate room until the mixed room is balanced – i.e. there is no longer a large majority of 0-types that would make it impossible for b-types to tell the truth in

equilibrium. But this means that the 0-types who are in a small room by themselves could deviate and massively improve their own information by moving to the large mixed room, which would cause all the b-types there to babble but would still be more informative to the switching 0-type than staying in the separate room.

- In case (3b), a player destroys information in the room that she *leaves* (by depriving those who stay behind of the information she would otherwise have communicated to them). Full segregation by types would be welfare-optimal in this case, but b-types can again improve their own information by joining the other room. They will not do so to an extent that would cause the 0-types there to babble. But any b-type who joins the room populated mainly by 0-types will babble there, whereas she would have truthfully revealed her information in a room populated only by b-types.

One implication of theorem 2 is that no player assigned to a mixed room in the welfare optimal room allocation has an incentive to leave this mixed room in favor of a less mixed room. This is due to the fact that there is at most one mixed room in the welfare optimal room allocation and this mixed room will contain more truthtelling players than any other room.

As figure 4 shows, whether the welfare optimum is an equilibrium can be non-monotonic in b. This is because in the parameter range between cases (2b) and (3b), it is not possible to "balance" the fully integrated room by removing b-types, but a fully segregated room would be worse than a fully integrated room in which only one type communicates informatively.

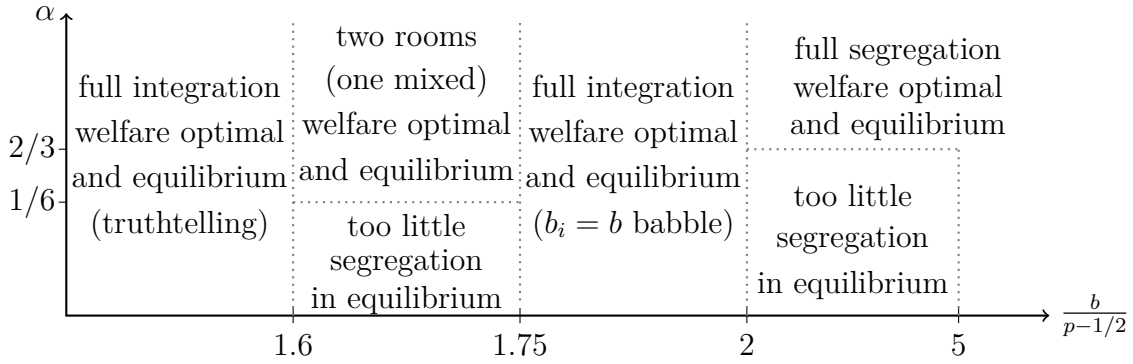

Figure 4: Welfare and equilibria when $n_0 = 5 > 4 = n_b$. (not to scale)

4.3. When is Segregation Optimal?

The previous section has shown that integration and segregation are, respectively, optimal if preferences are little polarized or very polarized. We can generalize this insight to arbitrary bias configurations with arbitrarily many biases. Let $\mathcal{B} = \langle b_1, b_2, \ldots, b_n \rangle$ be a

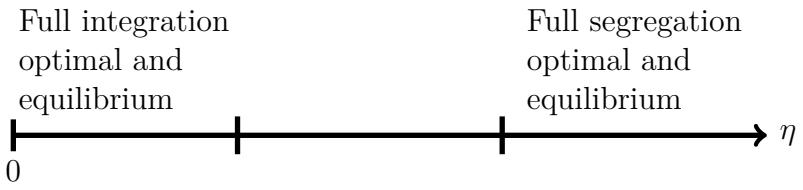

Figure 5: Welfare-optimal allocations that are also equilibria for large and small η.

bias configuration. (Note that this is not a set, as several people can have the same bias.) Assume that $\mathcal{B}$ is generic in the sense that no bias is the average of any set of other biases (except in cases where several people have the same bias).[14] Now we can consider an alternative bias configuration $\mathcal{B}_\eta = \langle \eta b_1, \eta b_2, \ldots, \eta b_n \rangle$, with $\eta \in (0, \infty)$. Intuitively, η parameterizes the polarization of preferences compared to the differences in information between players: A larger η can both mean an increase in preference polarization or a decrease in information dispersion.[15] Then the following is true:

Theorem 3. *(i) If η is sufficiently close to 0, full integration is welfare-optimal and a room-choice equilibrium for bias configuration $\mathcal{B}_\eta$.*

(ii) If η is sufficiently large, full segregation by bias types is generically welfare-optimal and a room-choice equilibrium for bias configuration $\mathcal{B}_\eta$. (Proof on page 33.)

Figure 5 summarizes the result. We can intuitively explain it in the following way: If biases are clustered very closely relative to how different the players' information is, having all players in one room would result in universal truth-telling. This cannot be improved upon in welfare terms, and it is also an equilibrium since any player would lose by leaving the fully integrated room.

On the opposite end of the spectrum, we consider the case where biases are clustered very widely compared to differences in information, and we do not assume special, non-generic properties such as that one bias is the exact average of two other biases. Then truth-telling will be impossible in any room that contains two or more players with different biases. Hence there exists no room allocation that can improve welfare compared to full segregation by bias types. Similarly, no player has an incentive to deviate from full segregation, since such a deviation cannot provide more information to the player himself or any other player.

To derive results for the area between the two cases (i.e. the center of figure 5), one would need to impose more structure, since the set of generic bias configurations is large and unordered – for example, welfare-best equilibria that are not the welfare-optimum

[14] More precisely, the assumption is that $b_i \neq \sum_{b_j \in \mathcal{B} \setminus \{b_i\}} \frac{\tilde{n}_{b_j}}{\sum_k \tilde{n}_{b_k}} * b_j$ for any vector of $\tilde{n}_{b_j} \in \{0, 1, \ldots, n_{b_j}\}$ where n_{b_j} is the number of players with bias b_j.

[15] One way to think about η is as the polarization measure proposed by Esteban and Ray (1994) (theorem 1) with an appropriate scaling parameter.

may involve too much as well as too little segregation.[16] We give an example for a case in which there is too much segregation in equilibrium in the supplementary material.

Our framework, however, can be used for the analysis of specific types of bias configurations that may be of theoretical or applied interest. In the supplementary material, we analyze two specific cases[17]: First, a world in which biases are uniformly distributed on an interval of the real line, so that there is no "massing" of biases anywhere – one might think of such a society as diverse but not polarized. In that case, (almost) full integration is welfare optimal and full integration is also an equilibrium.

In a second special case, biases are tightly clustered around a central value in a "single-peaked" configuration. We could think of this as an "anti-polarized" society in which preferences become monotonically less likely the further they are from the population average. In this case, full integration can be welfare-optimal (and is then also an equilibrium), but only if the concentration of biases around the central mode is strong enough.

4.4. Polarization Destroys Welfare

We have argued that segregation is a rational and Pareto-optimal response to polarization. This does not mean that polarization in itself increases welfare – quite the opposite. If we return to the η-parameterization under which we derived our theorem 3, we can show that both welfare and the amount of communicated information are weakly decreasing in η, i.e. our measure of polarization.

Proposition 2. *Denote expected welfare in the welfare optimal room assignment with bias configuration $\mathcal{B}_\eta$ by $W(\eta)$ and the total number of pieces of information in the welfare optimal room assignment by $\mathcal{Z}(\eta)$. $\mathcal{Z}(\eta)$ and $W(\eta)$ are both decreasing in η. (Proof on page 34.)*

To illustrate this result, consider the following thought experiment: Starting with any bias configuration and any room allocation, we increase η. This will weakly decrease communication in any room, which harms welfare. Allowing for further segregation may restore some communication, which reduces the harm – but not completely.

We should hence be very precise about the mechanism by which higher polarization decreases welfare. It is not through segregation, even though higher polarization causes more segregation, which ultimately causes less information to be exchanged. Saying "segregation lowers welfare" would ignore the crucial intermediate step, which is that polarization in itself causes an informational breakdown. In fact, segregation *mitigates* this breakdown, without of course being able to restore communication between people that are now in separate rooms.

[16] In fact, the notions of "too much" or "too little" segregation may not be well-defined if there are arbitrarily many bias groups.

[17] We have kept these cases, as well as the example from the preceding paragraph, in the supplementary material as we think of them as applications of our framework, and not as major insights themselves.

One could think of echo chambers as society's (decentralized) defense mechanism against polarization. Like fever in a human body, segregation occurs as the effect of an underlying problem, and its presence hence indicates that polarization is at problematic levels. Echo chambers, and segregation more generally, are a symptom of polarization. And just like artificially lowering fever, treating the symptom without addressing the cause can in fact exacerbate the situation. Reducing polarization will weakly improve welfare; reducing segregation may not.

5. Extensions: Uncertainty, Public Information, and Follower Networks

This section considers three extensions to our model, which show the general applicability of our framework and may be useful for specific applications. We describe each extension and present the results. All derivations and further details are, however, relegated to the supplementary material.

5.1. Public Information

Besides the private information they get from σ_i, players may also have access to common, public information that is relevant for their decision. Assume that instead of our usual assumption on θ, it was now $\theta = \tau\theta_0 + (1-\tau)\sum_{k=1}^{n}\theta_i$. In addition to the private signals σ_i that give player i information about θ_i, there is now also a public signal of accuracy p_0 that is informative about $\theta_0 \in \{0,1\}$. The parameter $\tau \in [0,1]$ gives the relative importance of public information.[18] Our main interest in this extension is the comparative static with respect to τ: Will more public information, for example due to progress in information technologies or higher-quality news outlets, lead to more or less informative communication and will this segregate society more or less?

The main mechanisms of our model remain unchanged in such an extension. Public information, however, crowds out incentives to tell the truth: If we increase the importance of public information τ, it becomes more tempting to mislead players with different biases. As private information is relatively less important for high τ, other players respond less strongly to one's message (if the message is believed to be truthful) and consequently players are less disciplined by the danger of misleading their audience "too much".

Formally, we can show that the truth-telling interval within any room (the equivalent to the interval from theorem 1 above) is

$$\left[\bar{b} - \frac{n_R - 1}{n_R}(p - \frac{1}{2})(1-\tau),\ \bar{b} + \frac{n_R - 1}{n_R}(p - \frac{1}{2})(1-\tau)\right].$$

The length of this interval is decreasing in τ, which means that for larger τ fewer players

[18] This parametrization can be interpreted as a simplified version of a model in which each of n players receives independent signals about k of in total $k * n$ substates. An increase in τ corresponds to some of these k signals now being publicly available to all players.

in a given room are truthtelling. Consequently, it is rational and efficient to segregate more if τ is higher. In particular, there exists a $\bar{\tau} < 1$ such that full segregation is optimal and an equilibrium for all $\tau \geq \bar{\tau}$.

This suggests an additional mechanism for why segregation occurs and how it may differ over time and between settings. In communication settings, both private or professional, where almost all relevant information is private information of the participants, it may be easier to achieve communication and hence segregation is less useful. But when the discussion is about national politics, for example, where almost all information is public and people's private knowledge and experiences are only a small facet of a larger whole, more segregation may be desirable.

The results also suggest that progress in information technologies, which make information publicly accessible that in earlier times was held only by experts, may lead to less (truthful) private communication and more segregation. Note, however, that this does not necessarily imply that players make less informed decision as the additional public information can more than outweigh the informational loss from less private communication.

5.2. Uncertainty

So far, we have assumed that all biases b_i are common knowledge. This may not always be the case, especially in environments where communication is somewhat anonymous, such as on the internet. In such cases, it seems reasonable to assume that both the state of the world and the types of all players are subject to uncertainty.

Assume that all biases b_i are randomly and independently distributed on $\mathbb{R}$ according to distribution F_i. Each player observes his own bias b_i, but only knows the distributions of the biases of other players. The main results of our model generalize to this setting with a few modifications. Players' motivations to tell the truth, similar to theorem 1, now depend on the distance between a player's realized bias and $\bar{b}^e$, the average of the expected biases of all players in the same room. Ex ante, the probability with which player i tells the truth hence depends on how likely it is that the realization of b_i lies within that interval around $\bar{b}^e$. An increase in mean-preserving uncertainty can increase or decrease truthtelling, depending on whether it shifts probability mass of b_i into the relevant interval around $\bar{b}^e$ or out of it. In general, however, we can show for several partial orderings of uncertainty that a sufficiently large increase in uncertainty will eventually erode all truthtelling. Such uncertainty would be most prevalent in anonymous, one-shot interactions such as in online public comment sections.

Uncertainty also has implications for whether segregation is efficient and individually optimal. Consider two bias groups (as in section 4.2 above) that are close enough to each other so that full integration is optimal and an equilibrium. Even a small increase in mean-preserving uncertainty can drastically reduce how much information is exchanged

in the fully integrated room, as players are now with a high probability too far from the average expected bias in the room to tell the truth. Segregation, however, may restore much of the information exchange (or even full truth-telling) in two segregated rooms. This may be welfare-optimal, especially given that the benefits of truth-telling are not linear in its probability.[19]

5.3. (Overlapping) Follower Networks

Our main model restricts how players can associate by only allowing players to join exactly one room, and only to communicate with the other players in that room. We can soften this assumption by considering a modified model that allows a freer choice of whom to learn from. Imagine that instead of the room choice stage, all players decide simultaneously to "follow" as many of the other players as they like. In other words, players create a directed communication network where links can be unilaterally created by the receiver of messages. In the communication stage, players then each send one message that is received by all of their followers. (This structure is close to how many social media services work.)

Many of our main results carry over to this extension in a modified way. Similarly to theorem 1, a player will now tell the truth if and only if his own bias is in the symmetric interval $\left[\overline{b} - \frac{n_{F_i}-1}{n_{F_i}}(p-\frac{1}{2}),\ \overline{b} + \frac{n_{F_i}-1}{n_{F_i}}(p-\frac{1}{2})\right]$ around the average bias of his followers. (n_{F_i} is the number of followers that i has.) This means that player i always wants to follow player j unless the very act of following makes j babble. This feature of the best response implies that the notions of most informative equilibrium and welfare optimal follower-assignment coincide. We can again show that if polarization increases, segregation becomes more desirable and it becomes optimal for players to segregate more. If polarization is low, it is efficient and an equilibrium for everyone to follow everyone – similar to the fully integrated room in our main model.

This extension has the interesting feature that there are differences between players with moderate and extreme preferences in how isolated they are from others. Players with moderate preferences can in equilibrium be followed by much of the population but still tell the truth, because different players' influences on $\overline{b}$ at least partially neutralize each other. Extremist players, however, can only be followed by other extremists of the same persuasion, as they would babble if followed by too many moderates or even by extremists at the other end of the spectrum.

6. Discussion

[19] To illustrate this point, consider a player with a relatively low bias who truthfully reveals σ^l half of the time. This means that 75% of the time, he sends the relatively uninformative message l. His messaging strategy thus partitions the state space much worse than truth-telling. This means that listening to two players who tell the truth "half of the time" in this way reduces the variance of one's belief less than listening to one players who fully tells the truth.

6.1. Who provides the Rooms?

In our model, we have assumed that the rooms are available in sufficient quantity so that players who want to segregate themselves can do so. In reality, that is of course not guaranteed. Information exchange could literally be impossible for lack of an empty room, such as when co-workers find themselves unable to discuss sensitive questions in an open-plan workspace. Bernstein and Turban (2018) have shown that the creation of open-plan offices tends to decrease the number of (public) face-to-face interactions and increase the number of (segregated) electronic interactions among colleagues. Or the shortage of rooms could be more figurative, such as when a politician may want to discuss his doubts of a policy with colleagues but cannot find a forum in which to do so without potentially giving ammunition to his political opponents.

In both cases, we have seen that segregation may be in the interest of everybody involved. It benefits not just the sender and the receiver in the segregated room, but even those who end up being excluded – since their inclusion would render communication impossible and thus not benefit anyone. Since rooms provide such clear benefits and are not automatically available, those in need of them should be willing to pay for whoever can provide them. We could imagine a group of agents who are sufficiently polarized and caught together in one place, which makes them unable to exchange any information. If now a plucky entrepreneur opened a separate room and took a small entrance fee, it would be an equilibrium for one subgroup of agents to each pay the fee, enter the room – and improve their own and everybody else's situation.

We think that this fable provides a way to understand the success of social messaging platforms such as Facebook, Twitter, WhatsApp and Snapchat. Each of these allows its users to send messages (and other content) to certain groups of others, with varying possibilities of exclusion. It can seem from the outside as if the service that is provided is to connect people with each other, but our model suggests it is just as much to exclude some people and not others, while providing sophisticated ways to determine who should and should not be excluded.[20] This has a strict economic logic to it: Once the Internet is available and ubiquitous, simply connecting people is not a scarce resource or service. But connecting them in such a way that they want to communicate truthfully, and can exchange the information they want to exchange, is much harder, and those who do it well can make a profit. Additionally, the resulting group structures are much less portable than files or contact lists (and often not at all), thus contributing to networking sites' market power.

[20]Facebook, for example, allows its users among other things to (i) choose which of their data is visible to search engines, (ii) choose for each post and image whether it is visible to everybody or just friends or friends of friends or even select group of friends (iii) block individual other users from seeing certain content (iv) create public or private events or groups to which members can be invited, (v) message directly with selected users or groups of users. All of these are tools of intelligent segregation as well as connection.

6.2. Political Parties and "Safe Spaces"

Of course, the room structure need not be provided by the market, it could be created by the agents themselves so that they can communicate with others who share their interests and world view. Besides the obvious examples of clubs and societies, we think that this is one rationale for the existence of political parties. In a society that is polarized enough, political parties can help solve the problem of aggregating political views and opinions.

We should also note that while messages are meaningless if a player is not truth-telling in equilibrium, the messages that he is most reluctant to send are those that could be seen as being counter to his own interest. For example, if an agent's b_i is much lower than the average of all b_j, he has no problem truthfully reporting σ^l, but is more reluctant after σ^h. This is how political parties can be useful: by providing a secluded forum in which, for example, members of a party can discuss the flaws and merits of their own candidates or programs. They would not be able to have this kind of discussion in the presence of members from other parties, where they would become overly defensive of "their" candidates and programs.

But the problem of defensiveness also provides an argument for so-called "safe spaces", i.e. spaces in which minorities or marginalized groups can communicate without outside interference. Informationally, such safe spaces may provide opportunities to communicate that would otherwise not exist. Consider the problem of two vegetarians who privately doubt whether vegetarianism is indeed a sensible choice – yet they find themselves defending it whenever they talk to (or in the presence of) non-vegetarians. Providing a "safe space" for vegetarians would allow them to discuss freely, and would hence provide a Pareto-improvement.

6.3. Room Choice as Communication Design

A large literature has recently analyzed the problem of designing socially optimal information structures – see, for example, Bergemann and Morris (2019). Such "information design" commonly assumes that a designer can set a rule by which messages about private information are chosen. Alternatively, players may themselves be able to commit to such a disclosure rule, which allows them to communicate truthfully despite a conflict of interest with the receiver (as in models of "Bayesian Persuasion", c.f. Kamenica and Gentzkow 2011). Any such design therefore requires that players can either be forced to follow such rules, or that rule-breaking can be monitored and punished. But in some settings, no commitment, monitoring or punishment may be available.

Our model shows that truthful communication can still be made possible even between people who prefer lying to each other, if there are other people in the same room to whom both players want to tell the truth. Crucially, room composition acts as a commitment device by making players *want* to tell the truth, which means that no objective mechanism to later compare their messages to the truth is needed. The tools we have developed in

sections 2.2 show how and when such "communication design" is possible.

The term "communication design", however, should not be understood to mean that a designer is always needed. As we have shown in section 4, players can often sort into an efficient allocation themselves (though they may need help in coordinating on one of many equilibria).

6.4. When are echo chambers bad?

Our argument that echo chambers can be useful does not necessarily mean that they are beneficial on balance and in every setting. Besides the mechanism that we analyze, echo chambers may have many other effects. Some of them can be informational, some behavioral, and some may only occur in settings that are slightly different from ours. While we believe that the mechanism we describe is very general, a complete assessment of echo chambers in a given context may well conclude that our mechanism is present but outweighed by other, detrimental effects. We will discuss some such potential effects; there are of course others and it is outside the scope of this paper to provide a full assessment of all effects that echo chambers can have.

Diversity. Our model considers gains from diversity in the sense that one's information gets more accurate (and hence one's decision better), the more people one hears from. We can thus weigh a well-known benefit of diversity (more information) against its less-discussed cost (problems with credible communication). An additional line of argument may assume that information is more closely correlated between people with similar biases – so that interaction with people with different biases becomes more valuable. Even that, however, does of course not solve the problem that communication across large preference differences may still be impossible, no matter how valuable the information that the other side holds.[21] Overall, there is simply no use in meeting people with a very diverse set of opinions and very useful information, if there is no way to get that information out of them.

Behavioral arguments. Once they hear only from people who are like them, people may fail to account for the correlation between the messages they receive.[22] Or they may fail to correctly learn in other, less well-defined ways, all of which make it harder for them to infer the state of the world from hearing only one side of the story. None of this, however, means in itself that a person would learn more if also exposed to viewpoints that they would not normally encounter, if their interlocutor rationally adjusts the informativeness of his message depending on whom he wants to inform and whom not.

[21] We consider an extension of a model in which there is only one state, and people with similar bias receive correlated information about it, in the supplementary material.

[22] C.f. the experimental work by Kallir and Sonsino (2009) and Eyster and Weizsäcker (2011) on "correlation neglect".

Endogenous Polarization One could also assume that preferences, which we take as given in our model, are actually the result of an endogenous process that depends on whom each person communicates with. Imagine, for example, that the game in this paper is played several times in a row, and between stages everybody's preferences move closer to the average of the preferences of the people they communicated with in the most recent stage. Segregated communication could then lead to further polarization, as the preferences of people who are in different rooms move further and further apart. As long as this process requires actual communication, however, some segregation may still be optimal, and such endogenous polarization would simply add another trade-off between getting people to communicate with each other (which is better than having them all babble) and causing further polarization down the line. In the long run, we might expect a stabilization of preferences around a few points, and consequently segregation into rooms, in the welfare optimum of such a repeated game.

Segregation by taste. There are two ways of applying the insights of this paper. The first, which we have used in developing our argument, is to see segregation as an informationally rational and welfare-optimal choice. Another perspective would be to assume that people segregate for exogenous or emotional reasons, or simply for reasons of taste. For example, rich people live in rich neighborhoods because of nicer houses and better infrastructure, and the segregation of types is only a secondary effect. But is such segregation necessarily informationally inefficient and bad for welfare? Our model suggests that this need not be the case. While rich people could surely learn from exchanging information with people whose lifestyle is different from theirs, it is far from given that such communication successfully takes places if we simply bring rich and poor together.[23] Even taste-based homophily can end up improving everyone's information.

Malicious actors. Our model is optimistic in the sense that while players want to mislead each other, they have an abstract interest in a well-informed society since it reduces the variance of people's mistakes. In reality, online "bots" and "trolls" may be interested in simply increasing uncertainty and chaos, both on behalf of state- and non-state actors. Similarly, (social) media companies may find that misinformation creates engagement even though it does not decrease (or even increases) the variance of people's actions. In both of these cases, segregation into echo chambers could help these actors in spreading misinformation, in particular if receivers cannot correctly deduce the motivations of a message's sender.

[23] Policies that may be more successful, following the results of our model, are: Narrowing the conflict of interest between rich and poor; convincing them that they have common goals; or reducing the uncertainty about each other's interests.

7. Conclusion

Modern democratic societies have three main mechanisms to aggregate information: Debates, markets, and votes. Of the three, debate is arguably the oldest – and while the other two require an organized framework and somebody who can enforce the rules, debate just needs an ability to speak and to listen.

But when will people speak truthfully (and hence have reason to listen)? In this paper, we have argued that if people have different preferences as well as different information, segregation into like-minded, homogeneous groups can be individually rational and Pareto-efficient. Echo chambers are not necessarily as destructive as popular discourse can make them seem. But even more importantly, we have shown that if segregation happens, it is not in itself the *cause* of an inability to debate. Instead, the existence of echo chambers is the *consequence* of differences in preferences, and of uncertainty and mistrust about other people's motives.

This has implications for how to improve debate. Society has a lot to gain from getting people with diverse backgrounds, experiences and opinions to exchange their views. But this cannot simply be achieved by forcing or cajoling people to interact who would not do so out of their own choosing. In fact, that could be counter-productive, as it could destroy disjoint groups in which communication works, in favor of large integrated groups in which it does not. Our research suggests that meaningful debate can only happen if the participants feel that they have enough in common and they trust each others' motives. Debate is more than putting people into a room and expecting them to come out smarter.

Appendix

A. Proofs

This appendix contains only the proofs for results that are explicitly given in the main text; all other results and their proofs can be found in the supplementary material.

Proof of lemma 1 on page 10.

Let $(m_1, \ldots, m_n)$ be an equilibrium. Player i's expected payoff when sending message m_i to players in room R_i can be written as

$$U_i(m_i|\sigma_i) = \mathbb{E}\left[-\left(a_i(m_{-i,R_i}, \sigma_i) - b_i - \sum_{k=1}^{n} \theta_k \right)^2 - \alpha \sum_{j \notin R_i} \left\{ \left(a_j(m_{-i,R_j}, \sigma_j) - b_i - \sum_{k=1}^{n} \theta_k \right)^2 \right\} \right.$$
$$\left. -\alpha \sum_{j \in R_i, j \neq i} \left\{ \left(a_j(m_i, m_{-i,R_i}, \sigma_j) - b_i - \sum_{k=1}^{n} \theta_k \right)^2 \right\} \middle| \sigma_i \right].$$

which can be split in a part that is independent of i's message m_i and a part that depends on m_i:

$$U_i(m_i) = \mathbb{E}\left[const - \alpha \sum_{j \in R_i, j \neq i} \left(a_j(m_i, m_{-i,R_i}, \sigma_j) - b_i - \sum_{k=1}^{n} \theta_k \right)^2 \middle| \sigma_i \right].$$

Specifically, sending message m^h gives expected payoff

$$U_i(m^h) = \mathbb{E}\left[const - \alpha \sum_{j \in R_i, j \neq i} \left(b_j - b_i + \mu_{ji}^h + \sum_{k \neq i} \mu_{jk} - \theta_i - \sum_{k \neq i} \theta_k \right)^2 \middle| \sigma_i \right]$$

where $\mu_{ji}^h = \mathbb{E}[\theta_i | m_i = m^h]$, i.e. μ_{ji}^h is the expectation of a player j in the same room as i concerning θ_i if player i sends message m^h. Note that this expectation is the same for all players $j \neq i$ in the same room as i. Sending message m^l gives

$$U_i(m^l) = \mathbb{E}\left[const - \alpha \sum_{j \in R_i, j \neq i} \left(b_j - b_i + \mu_{ji}^l + \sum_{k \neq i} \mu_{jk} - \theta_i - \sum_{k \neq i} \theta_k \right)^2 \middle| \sigma_i \right]$$

where $\mu^l_{ji} = \mathbb{E}[\theta_i | m_i = m^l]$. The difference in expected payoff is then

$$\begin{aligned}
\Delta U_i(\sigma_i) &= (U_i(m^h) - U_i(m^l))/\alpha \\
&= -\sum_{j\in R_i, j\neq i} \mathbb{E}\left[{\mu^h_{ji}}^2 - {\mu^l_{ji}}^2 + 2(\mu^h_{ji} - \mu^l_{ji}) \left(b_j - b_i + \sum_{k\neq i} \mu_{jk} - \theta_i - \sum_{k\neq i} \theta_k \right) \middle| \sigma_i \right] \\
&= -2(\mu^h_{ji} - \mu^l_{ji}) \sum_{j\in R_i, j\neq i} \left[\frac{\mu^h_{ji} + \mu^l_{ji}}{2} + b_j - b_i - \mathbb{E}\left[\theta_i | \sigma_i\right] \right] \\
&= 2(\mu^h_{ji} - \mu^l_{ji})(n_{R_i} - 1) \left[-\frac{\mu^h_{ji} + \mu^l_{ji}}{2} - \frac{\sum_{j\in R_i, j\neq i} b_j}{n_{R_i} - 1} + b_i + \mathbb{E}\left[\theta_i | \sigma_i\right] \right] \qquad (4)
\end{aligned}$$

where n_{R_i} denotes the number of players in room R_i. (For the transformation to line 3, we make use of the fact that μ_{ji} is the same for all $j \in R_i$.)

Player i is only willing to choose a mixed strategy after receiving signal σ_i if $\Delta U_i(\sigma_i) = 0$. From expression (4) it is clear that this can only be true for at most one signal as $\mathbb{E}\left[\theta_i | \sigma_i\right]$ varies in σ_i. Furthermore, $U_i(\sigma^h) = 0$ implies $U_i(\sigma^l) < 0$ and similarly $U_i(\sigma^l) = 0$ implies $U_i(\sigma^h) > 0$.

Now suppose i's equilibrium strategy m_i is mixed after signal σ^h. Then, $\Delta U_i(\sigma^h) = 0$ implies $\Delta U_i(\sigma^l) = 2(\mu^h_{ji} - \mu^l_{ji})(n_{R_i} - 1)(1 - 2p) < 0$ and therefore $m_i(\sigma^l) = m^l$ which implies $\mu^h_{ji} = p$ as a m^h is only sent by i after receiving signal σ^h. Consequently, $(\mu^h_{ji} + \mu^l_{ji})/2 \geq 1/2$ as $\mu^l_{ji} \geq 1 - p$. Now consider the equilibrium candidate (m^t_i, m_{-i}). With the truthful strategy m^t_i, $\mu^{th}_{ji} = p$ and $\mu^{tl}_{ji} = 1 - p$ and therefore $(\mu^{th}_{ji} + \mu^{tl}_{ji})/2 = 1/2$. This implies that $\Delta U_i(\sigma^h) > 0$ in the equilibrium candidate (m^t_i, m_{-i}), i.e. truthful reporting is optimal for i after receiving signal σ^h. In the equilibrium candidate (m^t_i, m_{-i}), truthful messaging is still optimal after signal σ^l as well: From $p > 1/2$, $\mu^h_{ji} \leq p$ and $\mu^l_{ji} \leq 1/2$ it follows that $-1/2 + (1 - p) < -\left(\mu^h_{ji} + \mu^l_{ji}\right)/2 + p$. As in the original equilibrium (m_i, m_{-i}) we had $\Delta U_i(\sigma^h) = 0$ and therefore $-\left(\mu^h_{ji} + \mu^l_{ji}\right)/2 + p = \sum_{j\in R_i, j\neq i} b_j/(n_{R_i} - 1) + b_i$, we get that $-1/2 + 1 - p < \sum_{j\in R_i, j\neq i} b_j/(n_{R_i} - 1) + b_i$ and therefore $U_i(\sigma^l) < 0$ in the truthful equilibrium candidate (m^t_i, m_{-i}). Hence, truthful messaging is i's best response in the equilibrium candidate (m^t_i, m_{-i}). Finally, note that the $\Delta U_j(\sigma_j)$ for $j \neq i$ is not affected by changing i's strategy from m_i to m^t_i. Hence, (m^t_i, m_{-i}) is an equilibrium.

The argument in case i's strategy is mixed after signal σ^l is analogous. □

Proof of theorem 1 on page 11.

Consider again the difference between lying and truth-telling for player i that we considered in equation (4) in the proof of lemma 1. Following corollary 1, we only consider pure strategies and therefore for every non-babbling player $\mu^h_{ji} = p$ and $\mu^l_{ji} = 1 - p$ which

implies that $\Delta U_i(\sigma^h) \geq 0$ simplifies to

$$\begin{aligned}
\frac{1}{n_R - 1} \sum_{j \in R_i, j \neq i} (b_i - b_j) &\geq \frac{1}{2} - p \\
b_i - \frac{1}{n_R - 1} \sum_{j \in R_i, j \neq i} b_j &\geq \frac{1}{2} - p \\
\frac{n_R}{n_R - 1} b_i - \frac{1}{n_R - 1} \sum_{k \in R_i} b_k &\geq \frac{1}{2} - p \\
b_i &\geq \bar{b} - \frac{n_R - 1}{n_R} \left(p - \frac{1}{2} \right).
\end{aligned}$$

If this inequality does not hold, player i will not use the truthful strategy in the most informative equilibrium and by corollary 1 this implies that he will babble in the most informative equilibrium.

We can analogously solve for $\Delta U_i(\sigma^l) \leq 0$ and get the interval used in the theorem. □

Proof of proposition 1 on page 12.

Denote the sets of babbling and truthful players in room R_j as R_j^{bab} and R_j^{truth}, respectively. For a given room allocation, the expected payoff of player i in room R_i is

$$\begin{aligned}
U_i = -\mathbb{E} \Bigg[& \left(\sum_{j \in R_i^{truth} \cup \{i\}} (\mu_{ij} - \theta_j) + \sum_{j \notin R_i^{truth} \cup \{i\}} (\frac{1}{2} - \theta_j) \right)^2 \\
& + \alpha \sum_{j \in R_i, j \neq i} \left(b_j - b_i + \sum_{k \in R_i^{truth} \cup \{j\}} (\mu_{jk} - \theta_k) + \sum_{k \notin R_i^{truth} \cup \{j\}} (\frac{1}{2} - \theta_k) \right)^2 \\
& + \alpha \sum_{j \notin R_i} \left(b_j - b_i + \sum_{k \in R_j^{truth} \cup \{j\}} (\mu_{jk} - \theta_k) + \sum_{k \notin R_i^{truth} \cup \{j\}} (\frac{1}{2} - \theta_k) \right)^2 \Bigg].
\end{aligned}$$

For any $i \neq j$, the two values of θ_i and θ_j are independent; the same is true for μ_{ij} and μ_{ik}. Hence $\mathbb{E}[\mu_{ij} - \theta_j] = 0$ and $\mathbb{E}[(\mu_{ij} - \theta_j)(\mu_{ik} - \theta_k)] = 0$, which means that the above expression can be rewritten as

$$\begin{aligned}
U_i = & - \sum_{j \in R_i^{truth} \cup \{i\}} \mathbb{E}\left[(\mu_{ij} - \theta_j)^2 \right] - \sum_{j \notin R_i^{truth} \cup \{i\}} \mathbb{E}\left[(\frac{1}{2} - \theta_j)^2 \right] \\
& - \alpha \sum_{j \in R_i, j \neq i} (b_j - b_i)^2 - \alpha \sum_{j \in R_i, j \neq i} \sum_{k \in R_i^{truth} \cup \{j\}} \mathbb{E}\left[(\mu_{jk} - \theta_k)^2 \right] - \alpha \sum_{j \in R_i, j \neq i} \sum_{k \notin R_i^{truth} \cup \{j\}} \mathbb{E}\left[(\frac{1}{2} - \theta_k)^2 \right] \\
& - \alpha \sum_{j \notin R_i} (b_j - b_i)^2 - \alpha \sum_{j \notin R_i} \sum_{k \in R_j^{truth} \cup \{j\}} \mathbb{E}\left[(\mu_{jk} - \theta_k)^2 \right] - \alpha \sum_{j \notin R_i} \sum_{k \notin R_i^{truth} \cup \{j\}} \mathbb{E}\left[(\frac{1}{2} - \theta_k)^2 \right].
\end{aligned}$$

Now note that $\mathbb{E}\left[(\mu_{jk}-\theta_k)^2\right]$ can have two possible values: If $k \in R_j^{truth} \cup \{j\}$, i.e. if j has received information about θ_k, then $\mathbb{E}\left[(\mu_{jk}-\theta_k)^2\right] = p(1-p)$. If j has not received information about θ_k, then $\mathbb{E}\left[(\mu_{jk}-\theta_k)^2\right] = \frac{1}{4}$. (We can check that information always reduces variance and increases welfare since $p > \frac{1}{2}$ and hence $p(1-p) < \frac{1}{4}$.)

This means that if i is telling the truth, we can write

$$\begin{aligned} U_i^{truth} &= -\alpha \sum_{j \neq i} \{(b_j - b_i)^2\} - \frac{1}{4}\left[n + \alpha(n-1)n\right] \\ &\quad + (\frac{1}{4} - p(1-p)) \left[n_{R_i}^{truth} + \alpha \sum_R \left\{ n_R^{truth} n_R^{truth} + (n_R - n_R^{truth})(1 + n_R^{truth}) \right\} - \alpha n_{R_i}^{truth} \right] \end{aligned} \quad (5)$$

The first term represents the loss that i suffers because other players choose a decision that is by $b_j - b_i$ too high from i's point of view. The second term represents the (theoretical) loss that would result if no player had any information and all μ's were simply $\frac{1}{2}$. The factors n and $(n-1)n$, which sum up to n^2, represent the total number of possible pieces of information in the model: If everybody's signal was available to everyone, n people would receive n pieces of information. The term hence represents, for each potential piece of information, the loss to i of that information not being available.

This loss is mitigated by information, which we see in the second line: i receives his signal and $n_{R_i}^{truth} - 1$ truthful messages, which means that instead of $\frac{1}{4}$, on each of these pieces of information i loses only $p(1-p) < \frac{1}{4}$. Other players, about whose decisions i cares with weight α, also receive some signals/messages: in any given room R, n_R^{truth} players receive their own signal and $n_R^{truth} - 1$ truthful messages while $n_R - n_R^{truth}$ players (those that babble in R) receive n_R^{truth} truthful messages and their own signal. (We have to subtract the correction term $-\alpha n_{R_i}^{truth}$ for room R_i in which there are only $n_{R_i}^{truth} - 1$ other players who tell the truth – in other words, i cannot count himself again as one of the players who receive information.) Analogously, we can write

$$\begin{aligned} U_i^{bab} &= -\alpha \sum_{j \neq i} \{(b_j - b_i)^2\} - 1/4\left[n + \alpha(n-1)n\right] \\ &\quad + (1/4 - p(1-p)) \left[1 + n_{R_i}^{truth} + \alpha \sum_R \left\{ n_R^{truth} n_R^{truth} + (n_R - n_R^{truth})(1 + n_R^{truth}) \right\} \right. \\ &\quad \left. - \alpha(1 + n_{R_i}^{truth}) \right]. \end{aligned} \quad (6)$$

In both the expressions for U_i^{truth} and U_i^{bab}, the second lines are adjusting the (pessimistic) expression in the first line for the reduction in variance by information. We can

simplify both expressions by simply writing

$$U_i = -\alpha \sum_{j \neq i} \{(b_j - b_i)^2\} - 1/4 \left[n + \alpha(n-1)n\right] + (1/4 - p(1-p)) \left[\zeta_i + \alpha \sum_{j \neq i} \zeta_j\right] \quad (7)$$

and express welfare as

$$\begin{aligned} W = \sum_i U_i &= \sum_i \left[-\alpha \sum_{j \neq i} \{(b_j - b_i)^2\} - 1/4 \left[n + \alpha(n-1)n\right] + (1/4 - p(1-p)) \left[\zeta_i + \alpha \sum_{j \neq i} \zeta_j\right] \right] \\ &= -\alpha \sum_{i=1}^{n} \sum_{j \neq i} \{(b_j - b_i)^2\} - \frac{1}{4} n^2 \left[1 + \alpha(n-1)\right] + (p - \frac{1}{2})^2 (1 + \alpha(n-1)) \sum_i \zeta_i. \end{aligned}$$

In this expression, all terms are model parameters except for the sum over all ζ_i, which shows that welfare is linearly increasing in $\sum_i \zeta_i$. □

Proof of theorem 3 on page 19.

Recall that a truth-telling equilibrium exists if and only if for every player i it is

$$\left| \sum_{k \neq i} \{b_k/(n-1)\} - b_i \right| \leq \frac{1}{2}.$$

This can be rewritten as $\left|\sum_k \{b_k\} - n b_i\right| / (n-1) \leq \frac{1}{2}$. If η is sufficiently small, this inequality holds for all players and all signals. Clearly, having all players in one room and telling the truth is welfare optimal whenever it is feasible, and no player can gain from leaving the room.

If $\left| \sum_{k \in R_i, k \neq i} \{b_k/(n-1)\} - b_i \right| > \frac{1}{2}$, then i will not be truthful when receiving either signal σ^l or σ^h. Generically, $\left| \sum_{k \in R_i, k \neq i} \{b_k/(n-1)\} - b_i \right| \neq 0$ for any room configuration containing players from more than one bias group. (This follows from the finiteness of players which implies that the number of such room configurations is finite.) Now observe that the left hand side of the non-truthtelling inequality is scaled by η while the right hand side is not. That is, for η sufficiently high, player i will report the highest (lowest) signal in all rooms in which $\sum_{k \in R_i, k \neq j} b_k < n_{R_i} b_i$ ($\sum_{k \in R_i, k \neq j} b_k > n_{R_i} b_i$). Put differently, any room that contains one or more players of a bias not equal to b_i will lead to totally uninformative messages by i if η is sufficiently high. For high enough η, this holds true for all players and it is then obvious that full separation is both welfare maximizing and an equilibrium. □

Proof of proposition 2 on page 20

Take two values of η, namely η' and $\eta'' > \eta'$. Denote a welfare optimal room assignment under η'' by R''. Consider the same room assignment R'' with biases η'. In each room the number of pieces of information is weakly higher with set of biases $B_{\eta'}$ than with set of biases $\mathcal{B}_{\eta''}$: By theorem 1 a player i is truthtelling if and only if $\eta\bar{b} - \frac{n_{R_i''}-1}{n_{R_i''}}(p-\frac{1}{2}) \leq \eta b_i \leq \eta\bar{b} + \frac{n_{R_i''}-1}{n_{R_i''}}(p-\frac{1}{2})$. Hence, player i will be truthtelling in room R_i'' with biases in $\mathcal{B}_{\eta'}$ if he is truthtelling in R_i'' with biases $\mathcal{B}_{\eta''}$ by $\eta' < \eta''$. Consequently, there is weakly more information transmitted in every room given assignment R'' under η' than under $\eta'' > \eta'$. This implies $W(\eta') \geq W(\eta'')$ by proposition 1. □

B. Bipolar polarization – step-by-step derivation of the results

We denote by n_0 (n_b) the number of people with bias 0 (b) and without loss of generality we let $n_0 \geq n_b$.

B.1. Segregation and full information as equilibrium

First, we ask the question when segregation, i.e. all players with bias $b_i = 0$ choosing room 1 and all players with bias $b_i = b$ choosing room 2, is an equilibrium. Clearly, every player is truthtelling in the most informative messaging equilibrium in this case and player i's expected payoff is

$$U_i^{fs} = -\alpha n_{-i} b^2 - 1/4\,[n + \alpha(n-1)n] + (1/4 - p(1-p))\left[(1-\alpha)n_i + \alpha n_i^2 + \alpha n_{-i}^2\right] \quad (8)$$

where n_i is the number of players with the same bias as player i and n_{-i} is the number of players with the other bias.

Proposition 3. *A segregation equilibrium exists if and only if one of the two following condition is met: (i)* $b > n_0(p-1/2)$*, (ii)* $p-1/2 < b \leq n_0(p-1/2)$ *and* $n_0 \leq (1+\alpha)(n_b-1)$.

Proof of proposition 3: Consider the incentives of player i to unilaterally deviate in his room choice in a segregation situation. There are three relevant cases that we will consider in turn: (i) after the deviation everyone including i still sends truthful messages, (ii) after the deviation i will babble but the players in the room he is switching to still send truthful messages and (iii) i's room switch leads to babbling by all players in the room he is switching to.

First, truthful messages after switch. This occurs if and only if $b \leq p - 1/2$. If i's bias group is the (weakly) smaller group, then i will benefit from a room switch in this case, see equation (8). Hence, segregation does not exist as equilibrium if b is less than $p - 1/2$.

Second, i babbles after the switch but everyone else remains truthtelling. This happens

if and only if $p - 1/2 < b \leq n_{-i}(p - 1/2)$. In this case, i's deviation payoff is

$$U_i^d = -\alpha n_{-i} b^2 - 1/4\,[n + \alpha(n-1)n] + (1/4 - p(1-p))\left[1 + n_{-i} + \alpha n_{-i}^2 + \alpha(n_i - 1)^2\right]$$

which is higher than U_i^{fs} if and only if $n_{-i} > (1+\alpha)(n_i - 1)$. Clearly, the members of the smaller group are the ones for who this constraint is more stringent. That is, a segregation equilibrium exists given $p - 1/2 < b \leq n_{-i}(p - 1/2)$ if and only if n_0 is at most $1 + \alpha(n_b - 1)$. Intuitively, members of the smaller group have a lot of information to gain form a switch if the larger group is very large. Hence, a segregation equilibrium only exists if the larger group is not too large.

Third, the switch leads to complete babbling in the room switched to. This will occur if and only if $b > n_{-i}(p-1/2)$. In this case, it is obvious that the deviation is unprofitable and a segregation equilibrium exists. □

Next we check for which parameter values a full information equilibrium, i.e. all players choosing the same room and reporting their signal truthfully, exists. Then, truthful reporting is an equilibrium of the messaging game if and only if $b(1 - (n_b - 1)/(n-1)) \leq p - 1/2$ which is equivalent to $bn_0/(n-1) \leq p - 1/2$. It is obvious that unilateral deviations in room choice from a fully integrated room are not profitable whenever truthtelling is by all players is an equilibrium in this room. We state this below as formal result and illustrate the results on full segregation and full information in figure 6.[24]

Proposition 4. *A full information equilibrium exists if and only if* $bn_0/(n-1) \leq p - 1/2$.

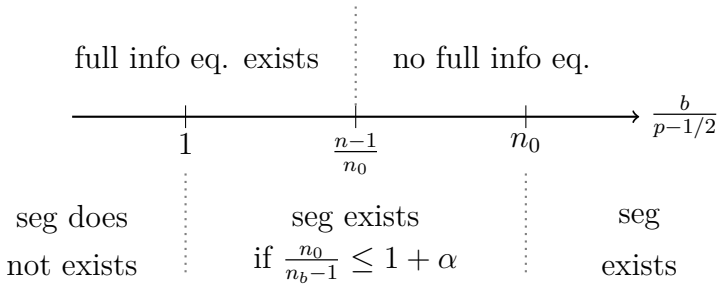

Figure 6: Segregation and full information as room choice equilibria

B.2. Intermediate results on welfare optimal room allocation

We will now consider how a planner would assign players to rooms in order to maximize the sum of players' payoffs. The planner's only tool is room assignment knowing the players' biases, i.e. the planner does not observe signals and cannot influence the messages sent or actions taken by the players.

From equation (3), it is clear that the planner's objective is equivalent to maximizing $\sum_i \zeta_i$, i.e. the total number of pieces of information by all players. We proceed in a

[24]Note that $(n-1)/n_0 \leq n_0$ as $n_0 \geq n/2$ and $n \geq 2$.

number of lemmas. To avoid case distinction, we use the convention that a player who is alone in a room will send a truthful message.

Lemma 1. *There is a welfare optimal room assignment without a room in which all players babble.*

Proof of lemma 1: In such a room there would be at least one player of each bias type. Splitting the room into two according to bias type would lead to weakly more transmitted pieces of information. □

Lemma 2. *Assume there are two rooms R_1 and R_2 such that in equilibrium players with bias $x \in \{0, b\}$ send truthful messages in both rooms. Then players with bias x will also send truthful messages in a merged room $R_1 \cup R_2$.*

Proof of lemma 2: As players send truthful messages by assumption in room R_i for $i \in \{1, 2\}$,

$$\frac{b}{p - 1/2} n_{R_i,-x} \leq n_{R_i} - 1$$

has to hold where $n_{R_i,-x}$ is the number of players in room R_i who do not have bias x. Summing this inequality over the two rooms $i = 1, 2$ yields

$$\frac{b}{p - 1/2} \left(n_{R_1,-x} + n_{R_2,-x}\right) \leq n_{R_1} + n_{R_2} - 2$$

which is sufficient for

$$\frac{b}{p - 1/2} \left(n_{R_1,-x} + n_{R_2,-x}\right) \leq n_{R_1} + n_{R_2} - 1.$$

The latter inequality implies that truthful messages in room $R_1 \cup R_2$ are optimal for players with bias x. □

Corollary 3. *The welfare maximizing room assignment will not include rooms R_1 and R_2 such that either*

- *both R_1 and R_2 are both populated exclusively by players with bias $x \in \{0, b\}$, or*
- *both R_1 and R_2 are populated by players with both biases and all players send truthful messages, or*
- *R_1 and R_2 are populated by players with both biases and only the players with bias $x \in \{0, b\}$ send truthful messages, or*
- *R_1 is exclusively populated by players with bias $x \in \{0, b\}$ and R_2 is populated by players with both biases but only players with bias x send truthful messages in R_2.*

Proof of corollary 3: In each of these cases, merging the two rooms maintains truthtelling incentives for those that originally sent truthful messages by lemma 2. As the truthful messages are received by more players, merging clearly increases the planner's objective. □

Lemma 3. *The welfare maximizing room assignment will not include rooms R_1 and R_2 such that both rooms contain players of each bias and all players in R_1 send truthful messages while only players of bias $x \in \{0, b\}$ send truthful messages in R_2.*

Proof of lemma 3: First, we consider the case where more players are in room R_1, i.e. $n_{R_1,x} + n_{R_1,-x} \geq n_{R_2,x} + n_{R_2,-x}$. Note that the number of pieces of information generated in those two rooms is $n_{R1}^2 + n_{R_2,x}(n_{R_2,x} + n_{R_2,-x})$. Note that $n_{R2,x} > n_{R_2,-x}$ as otherwise x would not be truthtelling in R_2 while $-x$ is not. Consider now an alternative room assignment that differs from the original one in the way that $n_{R2,-x}$ players of each bias are moved from R_2 to R_1. Denote everythign after the change using $\tilde{\cdot}$. That is, $\tilde{R}_2$ will contain $n_{R_2,x} - n_{R_2,-x}$ players of bias x and none of bias $-x$ while $\tilde{R}_1$ will contain $n_{R_1,x} + n_{R_2,-x}$ players of bias x and $n_{R_1,-x} + n_{R_2,-x}$ of type $-x$. The crucial result is that all players in room $\tilde{R}_1$ find truthtelling optimal: As truthtelling was optimal for players of both biases by assumption in R_1,

$$\frac{b}{p - 1/2} \leq \frac{n_{R_1} - 1}{\max\{n_{R_1,x}, n_{R_1,-x}\}}.$$

Now note that

$$\frac{n_{R_1} - 1}{\max\{n_{R_1,x}, n_{R_1,-x}\}} \leq \frac{n_{R_1} - 1 + 2n_{R_2,-x}}{\max\{n_{R_1,x} + n_{R_2,-x}, n_{R_1,-x} + n_{R2,-x}\}}$$

as the latter fraction is increasing in $n_{R_2,-x}$ and therefore truthtelling is still optimal in $\tilde{R}_1$. Clearly, truthtelling is also optimal in $\tilde{R}_2$ as only players with bias x are left there. Consequently, the total number of pieces of information in $\tilde{R}_1$ and $\tilde{R}_2$ is $(n_{R_1} + 2n_{R_2,-x})^2 + (n_{R_2,x} - n_{R_2,-x})^2$ which, by $n_{R_1} \geq n_{R_2} > n_{R_2,x}$, is strictly greater than the number of pieces of information generated by rooms R_1 and R_2. Hence, the planner prefers the room assignment $\tilde{R}_1$, $\tilde{R}_2$ over the room assignment R_1 and R_2 (keeping room assignment for players not in those rooms fixed).

Second, consider the case where more players are in room R_2, i.e. $n_{R_1,x} + n_{R_1,-x} < n_{R_2,x} + n_{R_2,-x}$. In this case, we argue that merging the two rooms to $R_1 \cup R_2$ will yield more information than keeping them separate. By lemma 2, players with bias x will still be truthtelling in room $R_1 \cup R_2$. Assume that players with bias $-x$ will not tell the truth in $R_1 \cup R_2$ (otherwise merging the two rooms is clearly optimal). Note that this implies $n_{R_1,x} + n_{R_2,x} > n_{R_1,-x}, n_{R_2,-x}$ as players with bias x tell the truth in $R_1 \cup R_2$ and players with bias $-x$ do not. The number of pieces of information generated in $R_1 \cup R_2$

is then $(n_{R_1,x} + n_{R_2,x})(n_{R_1,x} + n_{R_2,x} + n_{R_1,-x} + n_{R_2,-x})$. This is greater than the number of pieces of information generated in R_1 and R_2 separately as

$$\begin{aligned} n_{R_1,x}^2 + 2n_{R_1,x}n_{R_2,x} + n_{R_1,x}n_{R_1,-x} + n_{R_1,x}n_{R_2,-x} + n_{R_2,x}^2 + n_{R_2,x}n_{R_1,-x} + n_{R_2,x}n_{R_2,-x} \\ > n_{R_1,x}^2 + 2n_{R_1,x}n_{R_1,-x} + n_{R_1,-x}^2 + n_{R_2,x}^2 + n_{R_2,x}n_{R_2,-x} \end{aligned}$$

$$\Leftrightarrow 2n_{R_1,x}n_{R_2,x} - n_{R_1,x}n_{R_1,-x} + n_{R_1,x}n_{R_2,-x} + n_{R_2,x}n_{R_1,-x} - n_{R_1,-x}^2 > 0$$

$$\Leftrightarrow -n_{R_1,-x}n_{R_1} + n_{R_1,x}n_{R_2} + n_{R_2,x}n_{R_1} > 0$$

which holds true by $n_{R_1} < n_{R_2}$ and $n_{R_1,x} + n_{R_2,x} > n_{R_1,-x}, n_{R_2,-x}$. Consequently, the planner would prefer $R_1 \cup R_2$ to R_1 and R_2 separately. □

Lemma 4. *In a welfare maximizing room allocation, the following cannot occur: There are three rooms R_0 populated only of players with bias 0, R_b populated only with players of bias b and R_m populated with players of both biases.*

Proof of lemma 4: If R_m induces babbling, it is clearly better to assign the bias 0 (b) players in there to R_0 (R_b) instead. If only players with bias $x \in \{0, b\}$ send truthful messages in R_m, then it is clearly better to merge this room with R_x which maintains truthtelling incentives for players with bias x, see lemma 2.

Hence, we only have to consider the case where players with both biases send truthful messages in R_m. For concreteness assume there are more players in R_0 than in R_b denoted by $n_{R_0} \geq n_{R_b}$ (the reverse case is analyzed analogously). We consider two cases in turn. First, consider $n_{R_0} \geq n_{R_m} + n_{Rb}$. We claim that merging R_m and R_0 will then lead to more information than keeping these rooms separate. By lemma 2, player with bias 0 will still be truthtelling in $R_0 \cup R_m$ and the number of pieces of information generated by rooms R_b and $R_0 \cup R_m$ is at least $n_{R_b}^2 + (n_{R_0} + 1)(n_{R_m} + n_{R_0})$ which is strictly higher than the number of pieces of information generated by R_0, R_b and R_m, i.e. $n_{R_b}^2 + n_{R_0}^2 + n_{R_m}^2$, as $n_{R_0} + 1 > n_m$.

Second, consider $n_{R_0} < n_{R_m} + n_{R_b}$. The following change in the room allocation creates more information: Move n_{R_b} players from R_b and n_{R_b} players from R_0 to R_m. This leaves no one in R_b, $n_{R_m} + 2n_{R_b}$ in R_m and $n_{R_0} - n_{R_b}$ players in R_0. As in the proof of lemma 3, the move maintains truthtelling incentives for players in R_m. The number of pieces of information generated after the move is $(n_{R_m} + 2n_{R_b})^2 + (n_{R_0} - n_{R_b})^2 = n_{R_m}^2 + 4n_{R_b}n_{R_m} + 5n_{R_b}^2 + n_{R_0}^2 - 2n_{R_0}n_{R_b}$ which is higher than the number of pieces of information generated by R_0, R_b and R_m without the move, i.e. $n_{R_b}^2 + n_{R_0}^2 + n_{R_m}^2$, by $n_{R_0} < n_{R_b} + n_{R_m}$. □

Corollary 4. *The welfare optimal room assignment consists of at most two rooms and at most one room in which players of both biases are present.*

Proof of corollary 4: This follows from the combination of corollary 3 and lemmas 1, 3 and 4. □

B.3. Welfare optimal room allocation and equilibrium

Corollary 4 (and the preceding lemmas) leave the following possibilities for welfare optimal room assignment:

- segregation: each bias group has its own exclusive room.
- full integration: one room for all players
- mix: one room exclusively with players of bias x and one room with both bias types in which either
  - only players of type $-x$ send truthful messages
  - all players send truthful messages.

The idea behind the mix situation is the following: If there are more players with bias $-x$ than players with bias x and assigning all players to one room would lead to babbling, then it can be optimal not to separate completely but to assign some players of the minority x to the majority room $-x$. The x players will babble there but they receive a lot of information (truthful messages of all the $-x$ players). If such a situation is optimal, then clearly it must be the case that assigning one more minority player to the mixed room would lead to babbling. This is the first mixed assignment possibility.

The second mixed assignment possibility refers to a situation where the minority bias group would babble if all players were in one room while the majority would send truthful messages. In this case, taking some players of the majority to a separate room can restore truthtelling by both groups in the mixed room and might be optimal.

As should be clear from the discussion above, the mixed scenario is only welfare optimal if the group sizes differ. With equal group sizes either segregation or full integration is optimal. Which of the two is optimal depends on whether the bias difference b is sufficiently small to obtain truthtelling in a fully integrated room. While it is straightforward to prove this directly, we will here prove the more general theorem 2 and come back to the special case of equal group sizes afterwards.

Proof of theorem 2 on page 16.

By corollary 4, we can focus on at most two rooms in the welfare optimal room assignment.

If $b/(p-1/2) \leq (n_0+n_b-1)/n_0$, then full information, i.e. all players in one room and every player sends a truthful message, is feasible and a single room is obviously welfare optimal.

If $(n_0 + n_b - 1)/n_0 < b/(p - 1/2) \leq (n_0 + n_b - 1)/n_b$ (which is only possible if $n_0 > n_b$), then players with bias b would babble in a single room. It is important to notice that segregation cannot be optimal in this situation: segregation leads to $n_b^2 + n_0^2$ pieces of information while in one single room there would still be $n_0 * (n_0 + n_b)$ pieces of information which is greater than $n_b^2 + n_0^2$ by $n_0 > n_b$.[25] Consequently, there are two options for the welfare maximal room assignment. Either all players are in one single room and babbling by players with bias b is tolerated or some players with bias 0 are assigned to a separate room in order to balance the mixed room and restore truthtelling incentives for the bias b players. The maximal number of bias 0 players that can remain in the mixed room without inducing babbling by bias b players is $n_{m0} = \lfloor (n_b - 1)/(b/(p - 1/2) - 1) \rfloor$. Furthermore, truthtelling by the majority group clearly also requires that players with bias 0 are (weakly) in the majority. Hence, consider for now $n_{m0} \geq, n_b$. The number of pieces of information in the scenario with n_{m0} bias 0 players and all bias b players in one room and $n_0 - n_{m0}$ bias 0 players in a separate room is $(n_b + n_{m0})^2 + (n_0 - n_{m0})^2$. Whether this is higher or lower than the number of pieces in a fully integrated room, i.e. $n_0(n_0 + n_b) + n_b$, depends on the parameters. In particular, if $b/(p - 1/2)$ is close to the lower boundary (and n_b is not too small), two rooms will be optimal as n_{m0} will be relatively high. If $b/(p - 1/2)$ is close to the upper boundary, however, n_{m0} will be low and one room with babbling by the minority will be optimal. Finally, if $n_{m0} < n_b$, then babbling by one group occurs in any mixed room and therefore it is optimal to have one integrated room in which the minority players babble.

Finally, we consider $b/(p-1/2) > (n_0+n_b-1)/n_b$. In this case, even the bias 0 players will babble if all players are in one room. This implies that putting some bias zero players in an own separate room will no longer help: All the remaining bias zero players would have even higher incentives to babble and it would be more informative to fully separate the two bias groups. Consequently, only two options remain: Either the just mentioned segregation or enough bias b players are assigned to an own separate room to restore truthtelling incentives for the bias 0 players in the mixed room. The maximum number of bias b players that can remain in the mixed room without destroying truthtelling by bias 0 players is $n_{mb} = \lfloor (n_0-1)/(b/(p-1/2)-1) \rfloor$. This yields $n_0(n_0+n_{mb})+n_{mb}+(n_b-n_{mb})^2$ pieces of information which can, depending on the parameters, be higher or lower than the $n_0^2 + n_b^2$ pieces of information created by segregation.

It remains to check when the welfare optimal room assignment constitutes an equilibrium of the game. If full information is feasible, i.e. if $b/(p-1/2) \leq (n_0+n_b-1)/n_0$, then it is clearly an equilibrium. For $(n_0 + n_b - 1)/n_0 < b/(p - 1/2) \leq (n_0 + n_b - 1)/n_b$, the welfare optimal room assignment is definitely an equilibrium if it is optimal to keep all players in the same room (and have the smaller group babbling): The point is that this can

[25] Similarly, it is not optimal to put only some bias b players into their own room: As $n_b \leq n_0$, they would have less information there than they would have if they were babbling in one single room.

only be optimal if isolating one player of the larger group would not lead to truthtelling in a room with all other players. (If this was the case, then isolating one player of the larger group would be optimal.) But this implies that any deviation in room choice from the situation with one big mixed room will lead to less information for the deviating player and less (or the same) information for the other players. Following (3), such a deviation will therefore reduce the deviating player's expected payoff. Hence, the welfare maximizing room assignment is an equilibrium if $(n_0+n_b-1)/n_0 < b/(p-1/2) \leq (n_0+n_b-1)/n_b$ and $(n_b+n_{m0})^2+(n_0-n_{m0})^2 \leq n_0(n_0+n_b)$. For the case $(n_b+n_{m0})^2+(n_0-n_{m0})^2 > n_0(n_0+n_b)$ where two rooms are optimal – one mixed room with truthtelling by all players and one with only players of the larger group – the relevant question is whether the n_0-n_{m0} players in the room only for bias 0 players would want to deviate to the mixed room. Note that this deviation will destroy truthtelling by the bias b players. However, a unilaterally deviating player will obtain information from n_{m0} other players which is more information than the $n_0 - n_{m0} - 1$ truthful messages he obtains when not deviating.[26] It is therefore unsurprising that the welfare optimal room assignment will only be an equilibrium if α is sufficiently high. Using (3), the deviation will not be profitable if and only if it decreases $\zeta_i + \alpha \sum_{j \neq i} \zeta_j$. This is the case if and only if

$$n_{m0} + 1 + \alpha \left((n_0 - n_{m0} - 1)^2 + (n_{m0} + 1)(n_b + n_{m0}) \right) \leq n_0 - n_{m0} + \alpha \left((n_0 - n_{m0} - 1)(n_0 - n_{m0}) + (n_{m0} + n_b)^2 \right).$$

$$\Leftrightarrow \alpha \geq \frac{2n_{m0} - n_0 + 1}{n_0 - 2n_{m0} - 1 + n_b^2 + n_b(n_{m0} - 1)}. \tag{9}$$

Hence, if $(n_0+n_b-1)/n_0 < b/(p-1/2) \leq (n_0+n_b-1)/n_b$ and $(n_b+n_{m0})^2+(n_0-n_{m0})^2 > n_0(n_0+n_b)$ as well as $n_{m0} \geq n_b$, the welfare optimal room assignment is an equilibrium if and only if (9) holds.

For $b/(p-1/2) > (n_0+n_b-1)/n_b$, the welfare optimal room assignment can be either a mixed room combined with a room exclusively for the smaller group or segregation. In the first case it is straightforward to see that this is an equilibrium. Recall that moving one more player from the separate room to the mixed room would lead to babbling by all players in the mixed room. Clearly, such a deviation is not profitable. In the second case, we already showed before that total separation is an equilibrium if either $b/(p-1/2) \geq n_0$ or $\alpha \geq (1 + n_0 - n_b)/(n_b - 1)$. □

Note that for $b/(p-1/2) > n_0$ we get $n_{bm} = 0$ and in this case segregation is always

[26] To see this, first note that $n_{m0} \geq n_b$ as no truthtelling mixed room exists if bias b players were not truthtelling in a perfectly balanced, i.e. same number of players from each bias group, room. Then note that $(n_b+n_{m0})^2+(n_0-n_{m0})^2 > n_0(n_0+n_b)$ is more demanding for higher n_0 (keeping other parameters fixed). It is straightforward to calculate that the inequality cannot be satisfied (given $n_b \leq n_{m0}$) for $n_0 \geq 5n_{m0}/3$. Hence, $n_{m0} \geq 4n_0/5$ whenever this room assignment is welfare maximizing which implies $n_{m0} > n_0 - n_{m0} - 1$.

welfare optimal. Furthermore, an increase in $b/(p-1/2)$ will decrease both n_{m0} and n_{mb}. This implies that we can simply step through the different cases of theorem 2 as $b/(p-1/2)$ increases. However, depending on parameter values, in particular n_0 and n_b, some of the cases may be skipped. However, the first and last case will never be skipped as they always occur for sufficiently low, respectively high, values of $b/(p-1/2)$.

If the welfare optimal room allocation is not an equilibrium, then the welfare optimal equilibrium features too little segregation: In case (2b), the equilibrium has all players in a single room while the welfare optimal allocation has two rooms. In case (3b), the welfare optimal room allocation is full segregation which is not an equilibrium.

To illustrate, consider the special case of equal group sizes, i.e. $n_0 = n_b = n/2$. This rules out case (2) of theorem 2. Furthermore, the condition defining case (3a) is not met as it can be rewritten as $n_{mb}+1-n/2 \geq 0$ but given the condition $b/(p-1/2) > (n-1)/n_b$ of case (3) $n_{mb} < n/2-1$ and therefore case (3a) is not possible if $n_b = n_0 = n/2$. Hence, only case (1) and (3b) remain with equal group sizes and therefore the welfare optimal room allocation is either full segregation or full integration in this case. Full integration is welfare optimal if $b/(p-1/2) \leq (n-1)/(n/2)$ and is also an equilibrium in this case. If $b/(p-1/2) > (n-1)/(n/2)$, segregation is welfare optimal but only an equilibrium if $\alpha \geq 1/(n/2-1)$. Otherwise, there is too little segregation in equilibrium.

References

Acemoglu, D., A. Ozdaglar, and J. Siderius (2021). Misinformation: Strategic sharing, homophily, and endogenous echo chambers. National Bureau of Economic Research Working Paper, No. 28884.

Barberá, P., J. T. Jost, J. Nagler, J. A. Tucker, and R. Bonneau (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science 26*(10), 1531–1542.

Bergemann, D. and S. Morris (2019). Information design: A unified perspective. *Journal of Economic Literature 57*(1), 44–95.

Bernstein, E. S. and S. Turban (2018). The impact of the 'open' workspace on human collaboration. *Philosophical Transactions of the Royal Society B 373*(1753), 20170239.

Chater, J. (2016). What the EU referendum result teaches us about the dangers of the echo chamber. `https://www.newstatesman.com/2016/07/what-eu-referendum-result-teaches-us-about-dangers-echo-chamber`. Accessed: 2021-07-02.

Che, Y.-K. and K. Mierendorff (2019). Optimal dynamic allocation of attention. *American Economic Review 109*(8), 2993–3029.

Crawford, V. P. and J. Sobel (1982). Strategic information transmission. *Econometrica 50*(6), 1431–1451.

Del Vicario, M., A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences 113*(3), 554–559.

Dewan, T. and F. Squintani (2016). In defense of factions. *American Journal of Political Science 60*(4), 860–881.

Esteban, J.-M. and D. Ray (1994). On the measurement of polarization. *Econometrica 62*(4), 819–851.

Eyster, E. and G. Weizsäcker (2011). Correlation neglect in financial decision-making. *DIW Discussion Papers 1104*.

Galeotti, A., C. Ghiglino, and F. Squintani (2013). Strategic information transmission networks. *Journal of Economic Theory 148*(5), 1751–1769.

Gentzkow, M. and J. M. Shapiro (2011). Ideological segregation online and offline. *Quarterly Journal of Economics 126*(4), 1799–1839.

Grimes, D. R. (2017). Echo chambers are dangerous – we must try to break free of our online bubbles. https://www.theguardian.com/science/blog/2017/dec/04/echo-chambers-are-dangerous-we-must-try-to-break-free-of-our-online-bubbles. Accessed: 2021-07-02.

Hagenbach, J. and F. Koessler (2010). Strategic communication networks. *Review of Economic Studies 77*(3), 1072–1099.

Hooton, C. (2016). Social media echo chambers gifted Donald Trump the presidency. https://www.independent.co.uk/voices/donald-trump-president-social-media-echo-chamber-hypernormalisation-adam-curtis-pro html. Accessed: 2021-07-02.

Itten, A. (2018). Coming undone: How echo chambers balkanised society. https://www.politics.co.uk/comment-analysis/2018/08/16/coming-undone-how-echo-chambers-balkanised-society. Accessed: 2021-07-02.

Kallir, I. and D. Sonsino (2009). The neglect of correlation in allocation decisions. *Southern Economic Journal 75*(4), 1045–1066.

Kamenica, E. and M. Gentzkow (2011). Bayesian persuasion. *American Economic Review 101*(6), 2590–2615.

Kartik, N. (2009). Strategic communication with lying costs. *Review of Economic Studies 76*(4), 1359–1395.

Krishna, V. and J. Morgan (2001). A model of expertise. *Quarterly Journal of Economics 116*(2), 747–775.

Lawrence, E., J. Sides, and H. Farrell (2010). Self-segregation or deliberation? Blog readership, participation, and polarization in American politics. *Perspectives on Politics 8*(1), 141–157.

Li, M. and K. Madarász (2008). When mandatory disclosure hurts: Expert advice and conflicting interests. *Journal of Economic Theory 139*(1), 47–74.

Martinez, G. and N. H. Tenev (2020). Optimal echo chambers. arXiv preprint arXiv:2010.01249.

Morgan, J. and P. C. Stocken (2003). An analysis of stock recommendations. *RAND Journal of Economics 34*(1), 183–203.

Patty, J. W. (2022). Designing deliberation for decentralized decisions. *American Journal of Political Science*, forthcoming.

Penn, E. M. (2016). Engagement, disengagement, or exit: A theory of equilibrium associations. *American Journal of Political Science 60*(2), 322–336.

Quattrociocchi, W., A. Scala, and C. R. Sunstein (2016). Echo chambers on Facebook. Available on SSRN.

Sunstein, C. R. (2001). *Republic.com*. Princeton University Press.

Sunstein, C. R. (2017). *#Republic: Divided democracy in the age of social media*. Princeton University Press.