A Multi-agent Model for Opinion Evolution in Social Networks Under Cognitive Biases

Mário Alvim, Artur Gaspar da Silva, Sophia Knight, Frank Valencia

A Multi-Agent Model for Opinion Evolution under Cognitive Biases

Mário S. Alvim[1], Artur Gaspar da Silva[1], Sophia Knight[2], and Frank Valencia[3,4⋆]

[1] Department of Computer Science, UFMG, Brazil
[2] Department of Computer Science, University of Minnesota Duluth, USA
[3] CNRS-LIX, École Polytechnique de Paris, France
[4] Pontificia Universidad Javeriana Cali, Colombia

Abstract. We generalize the DeGroot model for opinion dynamics to better capture realistic social scenarios. We introduce a model where each agent has their own individual *cognitive biases*. Society is represented as a directed graph whose edges indicate how much agents influence one another. Biases are represented as the functions in the square region $[-1, 1]^2$ and categorized into four sub-regions based on the potential reactions they may elicit in an agent during instances of *opinion disagreement*. Under the assumption that each bias of every agent is a *continuous* function within the region of receptive but resistant reactions ($\mathbf{R}$), we show that the society converges to a consensus if the graph is strongly connected. Under the same assumption, we also establish that the entire society converges to a unanimous opinion if and only if the *source components* of the graph—namely, strongly connected components with no external influence—converge to that opinion. We illustrate that convergence is not guaranteed for strongly connected graphs when biases are either discontinuous functions in $\mathbf{R}$ or not included in $\mathbf{R}$. We showcase our model through a series of examples and simulations, offering insights into how opinions form in social networks under cognitive biases.

Keywords: Cognitive bias, Multi-Agent Systems, Social Networks

1 Introduction

In recent years, the significance and influence of social networks have experienced a remarkable surge, capturing widespread attention and shaping users' opinions in substantial ways. The *dynamics of opinion/belief formation* in social networks involves individuals expressing their opinions, being exposed to the opinions of others, and adapting or reinforcing their own views based on these interactions.

⋆ Mário S. Alvim and Artur Gaspar da Silva were partially supported by CNPq, CAPES, and FAPEMIG. Frank Valencia's contribution to this work is partially supported by the SGR project PROMUEVA (BPIN 2021000100160) supervised by Minciencias.

Modeling these dynamics allows us to gain insights into how opinions form, spread, and evolve within social networks.

The DeGroot multi-agent model [6] is one of most prominent formalisms for opinion formation dynamics in social networks. Society is represented as a directed graph whose edges indicate how much individuals (called *agents*) influence one another. Each agent has an opinion represented as a value in $[0, 1]$ indicating the strength of their agreement with an underlying proposition (e.g., "*vaccines are safe*"). They repeatedly update their opinions with the weighted average of their opinion differences (level of *disagreement*) with those who influence them. The DeGroot model is valued for its tractability, derived from its connection with matrix powers and Markov chains, and it remains a significant focus of study providing a comprehensive understanding of opinion evolution [7].

Nevertheless, the DeGroot model has an important caveat: It assumes *homogeneity* and *linearity* of opinion update. In social scenarios, however, two agents may update their opinions differently depending on their individual *cognitive biases* on disagreement—i.e., how they interpret and react towards the level of disagreements with others. This results in more complex updates that may involve non-linear even non-monotonic functions. For example, an individual under *confirmation (cognitive) bias* [3] may ignore the opinion of those whose level of disagreement with them is over a certain threshold. In fact, much of the unpredictability in opinion formation is due to users' *biases* in their belief updates, where users sometimes tend to reinforce their original beliefs, instead of questioning and updating their opinions upon encountering new information. Indeed, rather than perfect rational agents, users are often subject to cognitive biases.

In an earlier FORTE paper [1], we introduced a DeGroot-like model with a *non-linear* update mechanism tailored for a specific type of confirmation bias. The model was shown to be tractable and it provides insights into the effect of this cognitive bias in opinion dynamics. Nevertheless, it also assumes homogeneity of opinion update, and choosing a particular function to represent the bias, although natural, may seem somewhat ad-hoc.

To address the above-mentioned caveat, in this paper we introduce a generalization of the DeGroot model that allows for *heterogeneous* and *non-linear* opinion updates. Each agent has their own individual cognitive biases on levels of disagreement. These biases are represented as arbitrary functions in the square region $[-1, 1]^2$. The model then unifies disparate belief update styles with bias into a single framework which takes *disagreement* between agents as the central parameter. Indeed, standard cognitive biases of great importance in social networks such as backfire effect [9], authority bias [10], and confirmation bias [3], among others, can be represented in the framework.

We classify the biases in $[-1, 1]^2$ into four sub-regions $(\mathbf{M}, \mathbf{R}, \mathbf{B}, \mathbf{I})$ based on the cognitive reactions they may cause in an agent during instances of *opinion disagreement*. For example, agents that are malleable, easily swayed, exhibit fanaticism or prompt to follow authoritative figures can be modelled with biases in $\mathbf{M}$. Agents that are receptive to other opinions, but unlike malleable ones, can exhibit some skepticism to fully accepting them can be modelled with biases

in the region **R**. Individuals that become more extreme when confronted with opposing opinions can be modelled by biases in **B**. Finally insular agents can be modelled with the bias in **I**.

Consensus is a central property for opinion formation dynamics. Indeed the inability to converge to consensus is often a sign of a polarized society. In this paper we use the above-mentioned region classification to provide the following insightful theoretical results for consensus.

- Assuming that each bias of every agent is a *continuous* function in **R**, the society converges to a consensus if that society is strongly connected. This implies that a strongly connected society can converge to a consensus if its members are receptive but resistant to the opinions of others.
- Under the same assumption, we also establish that the entire society converges to a unanimous opinion if and only if the *source components* of the graph, i.e., strongly connected components with no external influence, converge to that opinion. This implies that upon agreeing on an opinion, closed and potentially influential groups, can make all individuals converge to that opinion in a society whose members are receptive but resistant.
- We show that convergence is not guaranteed for strongly connected graphs when biases are either discontinuous functions in **R** or not included in **R**.

We also demonstrate our model with examples and computer simulations that provide insights into opinion formation under cognitive biases. The open code for these simulations can be found at https://github.com/bolaabcd/polarization2.

2 An Opinion Model with Cognitive Biases

The DeGroot model [6] is a well-known model for social learning. In this formalism each individual (*agent*) repeatedly updates their current opinion by averaging the opinion values of those who influence them. But one of its limitation is that the model does not provide a mechanism for capturing the *cognitive biases* under which each individual may interpret and react to the opinion of others.

In this section we introduce a generalization of the DeGroot model with a mechanism to express arbitrary cognitive bias based on opinion disagreement.

2.1 Influence Graph

In social learning models, a *community/society* is typically represented as a directed weighted graph with edges between individuals (agents) representing the direction and strength of the influence that one carries over the other. This graph is referred to as the *Influence Graph*.

Definition 1 (Influence Graph). *An (n-agent)* influence graph *is a directed weighted graph* $G = (A, E, I)$ *with* $A = \{1, \ldots, n\}$ *the vertices,* $E \subseteq A \times A$ *the edges, and* $I : A \times A \to [0, 1]$ *a weight function s.t.* $I(i, j) = 0$ *iff* $(i, j) \notin E$.

The vertices in A represent n agents of a given community or network. The set of edges $E \subseteq A \times A$ represents the (direct) influence relation between these agents; i.e., $(i, j) \in E$ means that agent i influences agent j. The value $I(i, j)$, for simplicity written $I_{i,j}$, denotes the strength of the influence: 0 means no influence and a higher value means stronger influence. We use A_i to denote the set $\{j \mid (j, i) \in E\}$ of agents that have a direct influence over agent i.

Remark 1. In contrast to [1], we do not require agents to have nonzero self-influence. Furthermore, since we do not require the sum of influences over a given agent to be 1 (unlike [6]), we will use the following notation for *proportional influence* of j over i: $\overline{I_{j,i}} = \frac{I_{j,i}}{\sum_{k \in A_i} I_{k,i}}$ if $(j, i) \in E$, else $\overline{I_{j,i}} = 0$.

2.2 General Opinion Update

Similar to the DeGroot-like models in [7], we model the evolution of agents' opinions about some underlying *statement* or *proposition*. For example, such a proposition could be "*vaccines are unsafe*," "*human activity has little impact on climate change*,", "*AI poses a threat to humanity*", or "*Reviewer 2 is wonderful*".

The *state of opinion* (or *belief state*) of all the agents is represented as a vector in $[0, 1]^{|A|}$. If B is a state of opinion, $B_i \in [0, 1]$ denotes the *opinion* (*belief*, or *agreement*) value of agent $i \in A$ regarding the underlying proposition. If $B_i = 0$, agent i completely disagrees with the underlying proposition; if $B_i = 1$, agent i completely agrees with the underlying proposition. Furthermore, the higher the value of B_i, the stronger the agreement with such a proposition.

At each time unit $t \in \mathbb{N}$, every agent $i \in A$ updates their opinion. We shall use B^t to denote the state of opinion at time $t \in \mathbb{N}$. We can now define a general DeGroot-like opinion model as follows.

Definition 2 (Opinion Model). *An* Opinion Model *is a tuple* (G, B^0, μ_G) *where* G *is an* n*-agent influence graph,* B^0 *is the initial state of opinion, and* $\mu_G : [0, 1]^n \to [0, 1]^n$ *is a state-transition function, called* update function. *For every* $t \in \mathbb{N}$*, the state of opinion at time* $t + 1$ *is given by* $B^{t+1} = \mu_G(B^t)$.

The update functions can be used to express any deterministic and discrete transition from one opinion state to the next, possibly taking into account the influence graph. This work singles out and characterizes a meaningful family of update functions extending the basic DeGroot model with cognitive biases that are based on opinion *disagreement*. Intuitively, these update functions specify the reaction of an agent to the opinion disagreements with each of their influencers. To build some intuition, we first recall the update function of the DeGroot model.

Below we omit the index from the update function μ_G if no confusion arises.

2.3 DeGroot Update

The standard DeGroot model [6] is obtained by the following update function:

$$\mu(B)_i = \sum_{j \in A_i} \overline{I_{j,i}} B_j \tag{1}$$

for every $i \in A$. Thus, in the DeGroot model each agent updates their opinion by taking the weighted average of the opinions of those who influence them. We can rewrite Eq. 1 as follows:

$$\mu(B)_i = B_i + \sum_{j \in A_i} \overline{I_{j,i}}(B_j - B_i). \tag{2}$$

Notice that DeGroot update is *linear* in the agents' opinions and can be expressed in terms of *disagreement*: The opinion of every agent i is updated taking into account the weighted average of their *opinion disagreement* or *opinion difference* with those who influence them.

Intuitively, if j influences i, then i's opinion would tend to move closer to j's. The *disagreement term* $(B_j - B_i) \in [-1, 1]$ in Eq. 2 realizes this intuition. If $(B_j - B_i)$ is a negative term in the sum, the disagreement can be thought of as contributing with a magnitude of $|B_j - B_i|$ (multiplied by $\overline{I_{j,i}}$) to *decreasing* i's belief in the underlying proposition. Similarly, if $(B_j - B_i)$ is positive, the disagreement contributes with the same magnitude but to *increasing* i*'s* belief.

2.4 Disagreement-Bias Update

Now we generalize DeGroot updates by defining a class of update functions that also allows for *non-linear* updates, and for each agent to react differently to opinion disagreement with distinct agents. We capture this reaction by means of bias functions $\beta_{i,j} : [-1, 1] \to [-1, 1]$, where $(j, i) \in E$, on opinion disagreement stating how the bias of i towards the opinion of j, $\beta_{i,j}$, affects i's new opinion.

In the following definition we use the clamp function for the interval $[0, 1]$ which is defined as $[r]_0^1 = \min(\max(r, 0), 1)$ for any $r \in \mathbb{R}$.

Definition 3 (Bias Update). *Let (G, B^0, μ_G) be an opinion model with $G = (A, E, I)$. The function μ_G is a* (disagreement) bias update *if for every $i \in A$,*

$$\mu_G(B)_i = \left[B_i + \sum_{j \in A_i} \overline{I_{j,i}} \beta_{i,j}(B_j - B_i) \right]_0^1 \tag{3}$$

where each $\beta_{i,j}$ with $(j, i) \in E$, called the (disagreement) bias *from i towards j, is an endo-function*[5] *on $[-1, 1]$. The model (G, B^0, μ_G) is a* (disagreement) bias *opinion model if μ_G is a disagreement bias update function.*

The clamp function $[\cdot]_0^1$ guarantees that the right-hand side of Eq. 3 yields a valid belief value (a value in $[0, 1]$). Intuitively, the function $\beta_{i,j}$ represents the direction and magnitude of how agent i reacts to their disagreement $B_j - B_i$ with agent j. If $\beta_{i,j}(B_j - B_i)$ is a negative term in the sum of Eq. 3, then the bias of

[5] The biases we wish to capture can be seen as distortions of disagreements, themselves disagreements. It seems then natural to choose $[-1, 1]$ as the domain and co-domain of the bias function.

agent i towards j contributes with a magnitude of $|\beta_{i,j}(B_j - B_i)|$ (multiplied by $\overline{I_{j,i}}$) to *decreasing* i's belief in the underlying proposition. Conversely, if $\beta_{i,j}(B_j - B_i)$ is positive, it contributes to *increasing* i's belief with the same magnitude.

Below we identify some particular examples of the cognitive biases that can be captured with disagreement-bias opinion models.

Example 1 (Some Cognitive Biases). Clearly, the classical DeGroot update function Eq. 2 can be recovered from Eq. 3 by letting every bias $\beta_{i,j}$ be the identity on disagreement: i.e., $\beta_{i,j} = \texttt{degroot}$ where $\texttt{degroot}(x) = x$.

Confirmation Bias. We now illustrate some form of *confirmation bias* [3] where agents are more *receptive* to opinions that are closer to theirs. An example of confirmation bias can be obtained by letting $\beta_{i,j} = \texttt{conf}(x) = x(1 + \delta - |x|)/(1 + \delta)$ for a very small non-negative constant δ.[6] In the following plots and simulations we fix $\delta = 1 \times 10^{-4}$. This bias causes i to pay less attention to the opinion of j as their opinion distance $|x| = |B_j - B_i|$ tends to 1.

Backfire Effect. Let us now consider another important cognitive bias called backfire effect [9]. Under this effect an agent strengthens their position of disagreement with another agent if their opinions are significantly distant. A form of backfire effect can be obtained by letting $\beta_{i,j} = \texttt{backf}$ where $\texttt{backf}(x) = -x^3$. Notice that unlike the DeGroot update, this bias contributes to changing i's opinion with a *magnitude* of $|\texttt{backf}(B_j - B_i)|$ (multiplied by $\overline{I_{j,i}}$) *but in the opposite direction* of the opinion of j. This potentially makes the new opinion of agent i *more* distant from that of j.

Authority Bias. Another common cognitive bias in social networks is the authority bias [10] under which individuals tend to blindly follow authoritative or influential figures often to the extreme. Let $\beta_{i,j} = \texttt{fan}$ be the *sign* function, i.e., $\texttt{fan}(x) = x/|x|$ if $x \neq 0$, otherwise $\texttt{fan}(x) = 0$. This bias illustrates a case of die-hard *fanaticism* of i towards j. Intuitively, when confronted with any disagreement $x = B_j - B_i \neq 0$, this bias contributes to changing i's opinion with the *highest magnitude*, i.e., $|\beta_{i,j}(x)| = 1$, in the *direction* of the opinion of j.

Finally we illustrate a bias that, unlike the previous, causes agents to ignore opinions of others. We call it the *insular* bias $\beta_{i,j} = \texttt{ins}$ and it is defined as the zero function $\texttt{ins} : x \mapsto 0$. □

The particular bias function examples of Ex. 1 are depicted in the *square region* $[-1, 1]^2$ in Fig. 1. The functions may seem somewhat ad hoc but in Section 3 we identify a broad *family* of *bias functions* in the region $[-1, 1]^2$ that guarantees a property of central interest in multi-agent opinion evolution; namely, whether all the agents will converge to the same opinion, i.e. *convergence to consensus*.

Remark 2. We conclude this section by noting that unlike the DeGroot model, in Eq. 3 we allow agents to react with a distinct bias function to each of their influencers. This broadens the range of captured opinion dynamics and we illustrate this in the next section with an example exhibiting agents with different bias functions including those in Ex. 1. This, however, comes at a price; the

[6] The confirmation bias function from [1] uses $\delta = 0$

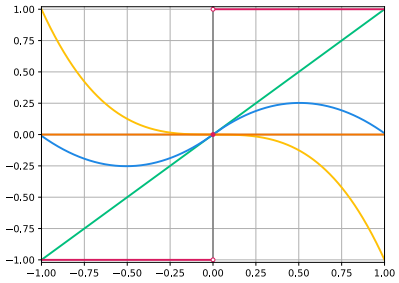


Fig. 1: Bias functions from Ex. 1 in the region $[-1,1]^2$: `degroot` (in green), `conf` (in blue), `backf` (in yellow), `fan` (in red), `ins` (in orange).

update function can be non-linear in the agents' opinions (see e.g., functions `backf` and `conf`). Thus, the analysis of opinion convergence using Markov chain theory for linear-system evolution as done for the DeGroot model is no longer applicable. In Section 3 we study opinion convergence using methods from real analysis.

2.5 Vaccine Example

Let us suppose that the proposition of interest is *"vaccines are unsafe"* and $G = (A, E, I)$ is as in Fig. 2. Suppose that initally the agents $1, 2, 3$ are *anti-vaxers* with opinion values $1.0, 0.9, 0.8$ about the proposition. In contrast, agents $4, 5, 6$ are initially *pro-vaxers*, with opinion values $0.2, 0.1, 0.0$ about the proposition. Thus, the initial state of opinion is $B^0 = (1.0, 0.9, 0.8, 0.2, 0.1, 0.0)$.

Notice that although agent 1 is the most extreme anti-vaxer, agent 6, the most extreme vaxer, has the highest possible influence over them. As we shall illustrate below, depending on the bias of 1 towards 6, this may have a strong impact on the evolution of the opinion of agent 1.

We now consider the evolution of their opinion under different update functions obtained by combining biases from Ex. 1. In Fig. 3 we show the evolution of opinions of vaxxers and anti-vaxxers using combinations of the bias functions from Fig. 1. Consider Fig. 3a. Agent 2 reaches the extreme opinion value 1.0 rather quickly because of their die-hard fanaticism towards the opinion of 1 (i.e., $\beta_{2,1} = \texttt{fan}$). As the influence of 6 on agent 1 backfires ($\beta_{1,6} = \texttt{back}$), agent 1 stays with belief value 1.0. Eventually, all the other biases contribute to changing the belief value of the influenced agents towards 1.0. Indeed, the agents converge to a consensus that vaccines are unsafe.

In Fig. 3b, the influence of 3 on agent 5 backfires, since $\beta_{5,3} = \texttt{backf}$. This makes their disagreement increase, moving agent 5's opinion closer to 0. On the other hand, the opinion of agent 6 is influenced at the same time by the belief values of 5 and 4 as in the DeGroot model ($\beta_{6,5} = \beta_{6,4} = \texttt{degroot}$) so her opinion stays between theirs.

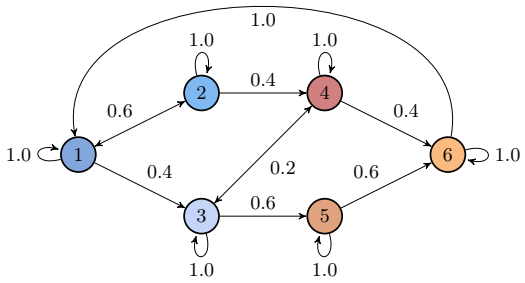


Fig. 2: Influence graph for vaccine example. The weight on edge (i, j) is the value $I_{i,j}$.

Notice that in Fig. 3c agent 5 reacts to 3 with die-hard fanaticism ($\beta_{5,3} = \texttt{fan}$) while 3's belief value does not converge to 0.0 or 1.0. Thus we obtain the looping behaviour of agent 5. The fanaticism of agent 5 propagates also to agent 6 since he is influenced by agent 5 by `degroot` bias.

Finally, notice the behaviours in Fig. 3 when all the agents have the same bias. In particular, Fig. 3g suggests convergence to consensus when all the agents are under confirmation bias. In fact convergence to consensus is indeed guaranteed for this example as we shall later see in this paper. Also, Fig. 3f is an example of why the clamp function might be necessary to guarantee that the belief values are always in $[0, 1]$.

The above illustrates that different types of biases can have strong impact on opinion evolution for a given influence graph. In the next section, we will identify meaningful families of bias as functions in the region $[-1, 1]^2$.

3 Bias Region and Consensus

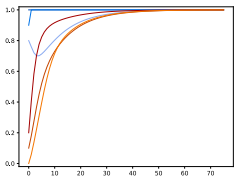
Consensus is a property of central interest in social learning models. Indeed, failure to converge to a consensus is often an indicator of polarization in a society.

Definition 4 (Consensus). *Let (G, B^0, μ_G) be an opinion model with $G = (A, E, I)$. We say that the subset of agents $A' \subseteq A$* converges to an opinion value *$v \in [0, 1]$ iff for every $i \in A'$, $\lim_{t\to\infty} B_i^t = v$. We say $A' \subseteq A$* converges to consensus *iff A' converges to an opinion value v for some v.*

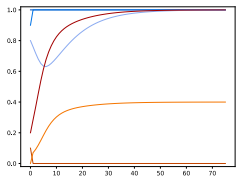
In this section we identify a broad and meaningful region of $[-1, 1]^2$ where all the *continuous* disagreement bias functions guarantee that agents converge to consensus under certain topological conditions on the influence graph.

3.1 Bias Regions

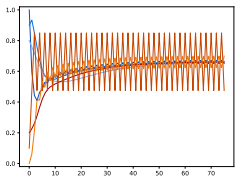
In what follows we say that a bias $\beta_{i,j}$ *is in a region* $R \subseteq [-1, 1]^2$ if its function graph is included in R, i.e., if $\{(x, \beta_{i,j}(x)) \mid x \in [-1, 1]\} \subseteq R$. We now identify regions of $[-1, 1]^2$ that capture several notions of cognitive bias.



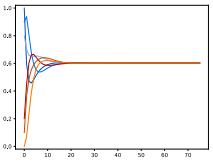
(a) $\beta_{2,1} = \texttt{fan}$, $\beta_{1,6} = \texttt{backf}$, $\beta_{4,2} = \beta_{1,2} = \texttt{degroot}$, otherwise $\beta_{i,j} = \texttt{conf}$.



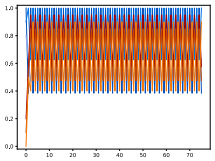
(b) $\beta_{2,1} = \texttt{fan}$, $\beta_{1,6} = \beta_{5,3} = \texttt{backf}$, $\beta_{6,5} = \beta_{6,4} = \texttt{degroot}$, otherwise $\beta_{i,j} = \texttt{conf}$.



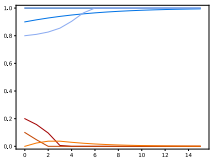
(c) $\beta_{5,3} = \texttt{fan}$, $\beta_{4,3} = \texttt{backf}$, $\beta_{6,5} = \beta_{1,6} = \texttt{degroot}$, otherwise $\beta_{i,j} = \texttt{conf}$.



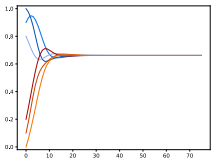
(d) Each $\beta_{i,j} = \texttt{degroot}$.



(e) Each $\beta_{i,j} = \texttt{fan}$.



(f) Each $\beta_{i,j} = \texttt{backf}$.



(g) Each $\beta_{i,j} = \texttt{conf}$.

Fig. 3: Simulations for G in Fig. 2 with $B^0 = (1.0, 0.9, 0.8, 0.2, 0.1, 0.0)$ using different biases. Each plot represents the evolution in time of the opinion of the agent in Fig. 2 with the same color.

Definition 5 (Bias Regions). *Let* $\mathbf{S}$ *be the square region* $[-1, 1]^2$. *Let the (sub)regions* $\mathbf{M}, \mathbf{R}, \mathbf{B}, \mathbf{I} \subseteq \mathbf{S}$, *named* Malleability, Receptive-Resistant, Backfire *and* Insular, *be defined as follows:*

$$\begin{aligned}
\mathbf{M} &= \{(x, y) \in \mathbf{S} \mid (x < 0 \text{ and } y \leq x) \text{ or } (x > 0 \text{ and } y \geq x) \text{ or } x = 0\} \\
\mathbf{R} &= \{(x, y) \in \mathbf{S} \mid (x < 0 \text{ and } x < y < 0) \text{ or } (x > 0 \text{ and } 0 < y < x) \text{ or } x = y = 0\} \\
\mathbf{B} &= \{(x, y) \in \mathbf{S} \mid (x < 0 \text{ and } 0 < y) \text{ or } (x > 0 \text{ and } y < 0) \text{ or } x = y = 0\} \\
\mathbf{I} &= \{(x, y) \in \mathbf{S} \mid y = 0\}.
\end{aligned}$$

The regions are depicted in Fig. 4. Notice that if a point (x, y) of a bias $\beta_{i,j}$ is in the *Malleability* region $\mathbf{M}$ (i.e., $y = \beta_{i,j}(x)$ and $(x, y) \in \mathbf{M}$) it means that for a disagreement $x = B_j - B_i$ between j and i, the bias will contribute with a magnitude $|y| \geq |x|$ (multiplied by $\overline{I_{j,i}}$) to changing the opinion of i in the direction of j's opinion. Since $|y| \geq |x|$, depending on the value of $\overline{I_{j,i}}$, the

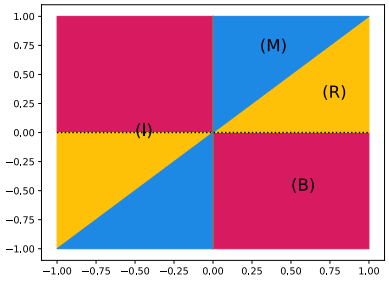


Fig. 4: Bias Regions: Malleability (**M**, in blue), Receptive-Resistant (**R**, in yellow), Backfire (**B**, in red), Insular (**I**, the dotted line $y = 0$).

opinion of i may move to match j's opinion or even further (which can make i's new opinion even more extreme than that of j). Individuals that blindly follow authoritative or influential figures, easily swayed agents, fanaticism, among others, can be modelled by bias functions in this region. Indeed the function `fan` from Ex. 1 is in **M** (see Fig. 1). The identity bias function `degroot` is also in **M**.

Like in the case above, if a point (x, y) of a bias $\beta_{i,j}$ is in the *Receptive-Resistant* region **R**, it also means that for a disagreement $x = B_j - B_i \neq 0$ between j and i, the bias contributes to changing the opinion of i in the direction of j's opinion. Nevertheless, the magnitude of contribution is not as high as the previous case, namely it is $|y|$ with $|x| > |y| > 0$. Individuals that are receptive to other opinions but, unlike malleable ones, may demonstrate some resistance, reluctance, or skepticism to fully accept them, can be modelled in this region. The confirmation bias function $\texttt{conf}(x) = x(1 + \delta - |x|)/(1 + \delta)$ from Ex. 1, where $\delta > 0$ is a very small constant, is in **R** (see Fig. 1).

In fact, it is worth noticing that for any constant $\delta > 0$, the resulting bias function $\beta_{i,j}(x) = x(1 + \delta - |x|)/(1 + \delta)$ is in **R**. In the limit, however, we have $\lim_{\delta \to \infty} x(1 + \delta - |x|)/(1 + \delta) = x = \texttt{degroot}(x)$ which is not in **R** but in **M**. Therefore, δ could be viewed as a *parameter of receptiveness*; the higher the value of δ, the more receptive and less resistant agent i is toward j's opinion. In the limit, agent i is not resistant and behaves as a malleable agent towards j.

Contrary to the previous two cases, if a point (x, y) of a bias $\beta_{i,j}$ is in the *Back-Fire* region **B**, it means that for a disagreement $x = B_j - B_i \neq 0$ between j and i, the bias contributes to changing the opinion of i *but in the opposite direction* of j's opinion. This bias can then cause the disagreement between i and j to grow. Individuals that become more extreme when confronted with a different opinion can be modelled by bias functions in this region. Indeed, the function `backf` from Ex. 1 is in **B** (see Fig. 1).

Finally, if a point (x, y) of a bias $\beta_{i,j}$ is in the Insular region **I**, it means that $y = 0$, thus for a disagreement $x = B_j - B_i \neq 0$ between j and i, the bias causes i to completely ignore the opinion of j. Individuals that are stubborn

or closed-minded can be modelled with the function in this region. In fact, the function `ins` from Ex. 1 is the only function in **I** (see Fig. 1).

We conclude this section with a proposition stating that we can dispense with the clamp function whenever all the bias functions are in the **R** region.

Proposition 1 (Update with Bias in R). *Given a Bias Opinion Model* (G, B^0, μ_G) *with* $G = (A, E, I)$*, if for all* $(a, b) \in E$ *we have* $\beta_{b,a} \in \mathbf{R}$*, then for all* $B \in [0,1]^{|A|}$ *and* $i \in A$*:* $\mu_G(B)_i = B_i + \sum_{j \in A_i} \overline{I_{j,i}} \beta_{i,j}(B_j - B_i)$.

The proof of this proposition can be found in the technical report [2].

3.2 Consensus under Receptiveness in Strongly Connected Graphs

Our first main result states the convergence to consensus for strongly connected societies when all bias functions are continuous and in the Receptive-Resistant Region defined in 3.1. We need some standard notions from graph theory.

Recall that a *path* from i to j in $G = (A, E, I)$ is a sequence $i_0 i_1 \dots i_m$ such that $i = i_0$, $j = i_m$ and $(i_0, i_1), (i_1, i_2), \dots (i_{m-1}, i_m)$ are edges in E. The graph G is *strongly connected* iff there is path from any agent to any other. We can now state our first consensus result.

Theorem 1 (Consensus I). *Let* (G, B^0, μ) *be a bias opinion model with a strongly connected graph* $G = (A, E, I)$*. Suppose that for every* $(j, i) \in E$*,* $\beta_{i,j}$ *is a continuous function in* **R***. Then the set of agents* A *converges to consensus.*

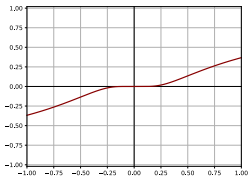
Hence, the continuous bias functions in **R** guarantee consensus in strongly connected graphs, regardless of initial beliefs. Intuitively, the theorem says that a strongly connected community/society will converge towards consensus if its members are receptive but resistant to the opinions of others.

Notice that the Vaccine Example in Sec. 2.5 with all agents under confirmation bias satisfy the conditions of Th. 1, so their convergence to consensus is guaranteed. In fact, the opinion difference between any two agents grows smaller rather rapidly (Fig. 3g illustrates this). In contrast, Fig. 5 illustrates an example, with a different bias also in **R**, where the opinion difference grows smaller much slowly. But since such an example also satisfies the conditions of Th. 1, convergence to consensus is guaranteed.

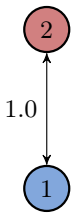
Before outlining the proof of this theorem, we elaborate on its conditions.

Discontinuous Bias. Requiring continuity for the bias functions in Th. 1 seems reasonable; small changes in an opinion disagreement value $x = B_j - B_i$ should result in small changes in i's biased reaction to x. Nevertheless, if we relaxed the continuity requirement, we would have the following counter-example.

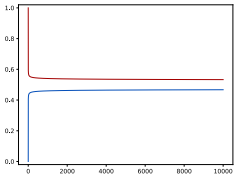
Consider a strongly connected graph with two agents with $I_{1,2} = I_{2,1} = 1$, agent 1 influences agent 2 with the bias functions $\beta_{1,2} = \beta_{2,1} = f$, satisfying $f(x) = \frac{x}{8}$ if $x \in [-\frac{1}{2}, \frac{1}{2}]$, $f(x) = \frac{x-0.5}{8}$ if $x \in (\frac{1}{2}, 1]$ and $f(x) = \frac{x+0.5}{8}$ if $x \in [-1, -\frac{1}{2})$. If one agent starts with belief value 1.0 and the other 0.0, then they will not converge to consensus (their belief values will approach $\frac{3}{4}$ and $\frac{1}{4}$, but will never reach those values). Figure 6 illustrates this example.



(a) Bias function $\beta_{1,2}(x) = \beta_{2,1}(x) = \begin{cases} 0 & \text{if } x = 0 \\ \frac{x}{|x|} \cdot e^{-\frac{1}{|x|}} & \text{if } x \neq 0 \end{cases}$



(b) Influence Graph ($I_{1,2} = I_{2,1} = 1.0$).



(c) Each plot represents the evolution in time of the opinion of the agent in Fig. 5b with the same color.

Fig. 5: Simulations with $B^0 = (0.0, 1.0)$ using a bias function in region **R**, with very slow convergence.

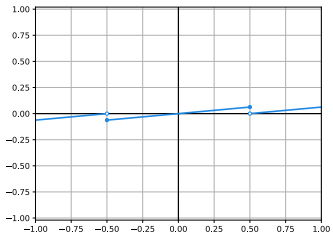
Bias Outside **R**. Notice that Th. 1 requires bias functions to be in the responsive-resistant region **R**. We consider counter-examples where we allow bias functions outside this region in Th. 1. If we allowed continuous bias functions outside **R** with points in the backfire region **B**, then the scenario in Fig. 3f provides a counter-example to consensus. If we allow continuous bias functions outside **R** with points in region **M**, then the scenario in Fig. 7 is a counter-example to consensus: notice how the absolute value of their disagreement begins at 0.001 and increases until it reaches 1. Finally, it is clear that if we allowed the only function in **I**, the insular bias, with the graph in Fig. 5b and initial beliefs $B^0 = (0, 1)$, consensus will never be reached since the agents will ignore each other.

3.3 Proof Outline of Th. 1

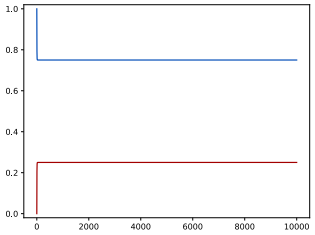
In this Section we outline the proof of Th. 1. In the process we single out the central properties of the behaviour of agents that are receptive and yet resistant to disagreement. The complete proof can can be found in the technical report [2].

Let (G, B^0, μ) be as in the statement of Th. 1. Suppose $B = \mu^t(B^0)$ is the state of opinion at some time $t \geq 0$ where consensus has not yet been reached: i.e., assume $\min(B) \neq \max(B)$ where $\min(B)$ and $\max(B)$ are the minimum and maximum opinion values in B. By assumption, all the biases $\beta_{i,j}$ are in **R**. Thus $\beta_{i,j}(x) = y$, where $x = B_j - B_i$, contributes to update the opinion of i in the direction of the opinion of j but with a magnitude $|y| > 0$ strictly smaller than $|x|$ if $|x| > 0$ (or equal to 0 if $|x| = 0$). Using this and Prop. 1[7], we show the new (updated) opinion of each i, $\mu(B)_i$, is bounded as follows:

[7] This follows from the known property that weighted averages of any set of values are always between the minimum and the maximum of those values.



(a) Bias function $\beta_{1,2}(x) = \beta_{2,1}(x) =$
$$\begin{cases} \frac{x+0.5}{8} & \text{if } -1 \leq x < -\frac{1}{2} \\ \frac{x}{8} & \text{if } -\frac{1}{2} \leq x \leq \frac{1}{2} \\ \frac{x-0.5}{8} & \text{if } \frac{1}{2} < x \leq 1 \end{cases}$$



(b) Each plot represents the evolution in time of the opinion of the agent in Fig. 5b with the same color.

Fig. 6: Counter-example to consensus for two agents with non-continuous bias functions in $\mathbf{R}$, for G in Fig. 5b and $B^0 = (1.0, 0.0)$.

Lemma 1 (Update Bounds). *For each* $i \in A$, $\min(B) \leq \mu(B)_i \leq \max(B)$.

We use the above lemma to prove that the bounded sequences of minimum and maximum opinion values at each time, $\{\min(B^t)\}_{t\geq 0}$ and $\{\max(B^t)\}_{t\geq 0}$, are monotonically non-decreasing and monotonically non-increasing. Thus by the Monotone convergence theorem [11], they both converge. Therefore, by the Squeeze theorem [11], to prove Th. 1, it suffices to show that $\{\min(B^t)\}_{t\geq 0}$ and $\{\max(B^t)\}_{t\geq 0}$ converge to the same value.

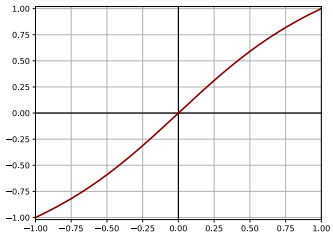
We first prove the following lemma which intuitively states that the number of extreme agents decrease with time.

Lemma 2 (Extreme Agents Reduction). *Suppose that* $\min(B) \neq \max(B)$ *and let* $M = \max(B)$. *If* G *has a path* $i_1 \dots i_n$ *such that* $B_{i_n} = M$ *and* $B_{i_1} < M$, *then* $|\{j \in A : B_j \geq M\}| > |\{j \in A : \mu(B)_j \geq M\}|$. *A symmetric property applies to the minimum.*

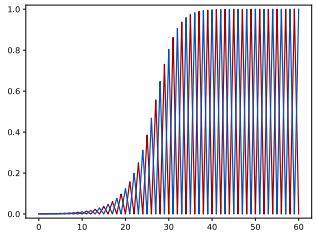
To see the lemma's intuition, notice that since G is strongly connected and $\min(B) < \max(B)$, G indeed has a path $i_1 \dots i_n$ such that $B_{i_n} = M = \max(B)$ and $B_{i_1} < M$. In the path some agent i_k whose belief value is equal to M will be influenced by some agent with smaller belief value. Thus, since the bias functions are in $\mathbf{R}$, the opinion of i_k will change in the direction of the smaller value, and thus will strictly decrease. Also, no agent that had a smaller belief value will reach the current maximum, as the bias functions are in the region $\mathbf{R}$.

Thus, because of Lem. 2 and G being strongly connected, we conclude that the maximum (minimum) belief value will eventually decrease (increase). I.e.,

Corollary 1. *Suppose that* $\min(B) \neq \max(B)$. *Then there exist* $s, t > 0$ *such that* $\max(\mu^s(B)) < \max(B)$ *and* $\min(\mu^t(B)) > \min(B)$.



(a) Bias function $\beta_{1,2}(x) = \beta_{2,1}(x) = \frac{\arctan x}{\arctan 1}$



(b) Each plot represents the evolution in time of the opinion of the agent in Fig. 5b with the same color.

Fig. 7: Counter-example for consensus when all bias function are continuous but allowed to have points in $\mathbf{M}$, with initial belief vector $B^0 = (0.0, 0.001)$ and the influence graph of figure 5b.

We now apply Bolzano-Weierstrass theorem [11][8] to find a sub-sequence $\{B^t\}_{t\in\Delta}$ of $\{B^t\}_{t\in\mathbb{N}}$ that converges to some B^∞. Notice that $\{\max(B^t)\}_{t\in\Delta}$ converges to $\max(B^\infty)$ and it is a sub-sequence of the convergent sequence $\{\max(B^t)\}_{t\in\mathbb{N}}$, so $\{\max(B^t)\}_{t\in\mathbb{N}}$ should also converge to $\max(B^\infty)$. Since each bias function $\beta_{i,j}$ is continuous, the update function μ is continuous. Therefore, $\{\mu(B^t)\}_{t\in\Delta}$ converges to $\mu(B^\infty)$, and thus $\{\max(\mu(B^t))\}_{t\in\Delta}$ converges to $\max(\mu(B^\infty))$. But since the sequence $\{\max(\mu(B^t))\}_{t\in\Delta} = \{\max(B^{t+1})\}_{t\in\Delta}$ is a sub-sequence of the convergent sequence $\{\max(B^t)\}_{t\in\mathbb{N}}$, both must converge to the same value, hence $\max(B^\infty) = \max(\mu(B^\infty))$. Similarly, we can show that $\min(B^\infty) = \min(\mu(B^\infty))$. It can thus be shown that if we repeatedly apply μ to B^∞, the maximum should not change, and the same applies to the minimum. More precisely, we conclude the following.

Corollary 2. $\max(B^\infty) = \max(\mu^t(B^\infty))$ *and* $\min(B^\infty) = \min(\mu^t(B^\infty))$ *for each* $t \geq 0$.

Consequently, if $\min(B^\infty) \neq \max(B^\infty)$ then Cor. 1 and Cor. 2 lead us to a contradiction. Therefore, $\min(B^\infty) = \max(B^\infty)$ and thus, $\{\min(B^t)\}_{t\geq 0}$ and $\{\max(B^t)\}_{t\geq 0}$ converge to the same value $\max(B^\infty)$ as wanted. □

3.4 Consensus Under Receptiveness in Arbitrary Graphs

Recall that Th. 1 applies to strongly connected influence graphs. Our second main result applies to arbitrary influence graphs. First we need to recall the notion of strongly-connected components of a graph.

[8] Every infinite bounded sequence in $\mathbb{R}^n$ has a convergent sub-sequence.

A *strongly-connected component* of G is a *maximal subset* $S \subseteq A$ such that for each two $i, j \in S$, there is path from i to j. A strongly-connected component S is said to be a *source component* iff there is no edge $(i, j) \in E$ such that $i \in A \setminus S$ and $j \in S$. We use $\mathcal{S}(G)$ to denote the set of source components of G.

Intuitively, a source component of a graph can be thought of as a closed group that is not externally influenced but may influence individuals outside the group. The following theorem gives a characterization of consensus with biases in $\mathbf{R}$ for arbitrary graphs in terms of source components.

Theorem 2 (Consensus II). *Let (G, B^0, μ) be a bias opinion model with $G = (A, E, I)$. Suppose that for every $(j, i) \in E$, $\beta_{i,j}$ is a continuous function in $\mathbf{R}$. Then the set of agents A converges to consensus iff there exists $v \in [0, 1]$ such that every source component $S \in \mathcal{S}(G)$ converges to opinion v.*

The above theorem, whose proof can be found in the technical report [2], provides the following intuitive yet insightful remark. Namely, upon agreeing on an opinion, the closed and potentially influential groups, can make all individuals converge to that opinion in a society whose members are receptive but resistant.

4 Concluding Remarks and Related Work

We introduced a generalization of the DeGroot Model where agents interact under different biases. We identified the notion of bias on disagreement and made it the focus our model. This allowed us to identify families of biases that capture a broader range of social dynamics. We also provided theoretical results characterizing the notion of consensus for a broad family of cognitive biases.

The relevance of biased reasoning in human interactions has been studied extensively in [13], [10], [12], and others.

There is a great deal of work on formal models for belief change in social networks; we focus on the work on *biased* belief update, which is the focus of this paper. Some models were previously proposed to generalize the DeGroot model and introduce bias, for instance [5], [4] and [15] analyse the effects of incorporating a bias factor for each agent to represent biased assimilation: how much of the external opinions the agent will take into consideration. [14] extends the model [5] to include the effect of backfire-effect as well. The main difference of these models to our model is that biases are not incorporated in those models in terms of the disagreement level between agents, but either as an exponential factor that reduces the impact of the opinion of neighbours or by dynamically changing the weights of the DeGroot model. Thus, our model brings a new point of view to how distinct types of biases can be represented and identified.

In [8], it is proved that "constricting" update functions, roughly, functions where the extreme agents move closer to each other, lead to convergence in strongly connected social networks. This is similar to our theorem, indeed, the functions in our $\mathbf{R}$ region are easily shown to be constricting. However, their social network model is more abstract than ours and further from real social networks, and they do not directly analyse biases as a function of disagreement.

References

1. Alvim, M.S., Amorim, B., Knight, S., Quintero, S., Valencia, F.: A multi-agent model for polarization under confirmation bias in social networks. In: International Conference on Formal Techniques for Distributed Objects, Components, and Systems. pp. 22–41. Springer (2021)
2. Alvim, M.S., da Silva, A.G., Knight, S., Valencia, F.: A multi-agent model for opinion evolution under cognitive biases (2024)
3. Aronson, E., Wilson, T., Akert, R.: Social Psychology. Upper Saddle River, NJ : Prentice Hall, 7 edn. (2010)
4. Chen, Z., Qin, J., Li, B., Qi, H., Buchhorn, P., Shi, G.: Dynamics of opinions with social biases. Automatica **106**, 374–383 (2019). https://doi.org/https://doi.org/10.1016/j.automatica.2019.04.035, https://www.sciencedirect.com/science/article/pii/S0005109819301955
5. Dandekar, P., Goel, A., Lee, D.: Biased assimilation, homophily and the dynamics of polarization. Proceedings of the National Academy of Sciences of the United States of America **110** (03 2013). https://doi.org/10.1073/pnas.1217220110
6. DeGroot, M.H.: Reaching a consensus. Journal of the American Statistical Association **69**(345), 118–121 (1974)
7. Golub, B., Sadler, E.: Learning in social networks. Available at SSRN 2919146 (2017)
8. Mueller-Frank, M.: Reaching Consensus in Social Networks. IESE Research Papers D/1116, IESE Business School (Feb 2015)
9. Nyhan, B., Reifler, J.: When corrections fail: The persistence of political misperceptions. Political Behavior **32**(2), 303–330 (2010). https://doi.org/10.1007/s11109-010-9112-2
10. Ramos, V.J.: Analyzing the role of cognitive biases in the decision-making process. Advances in Psychology, Mental Health, and Behavioral Studies (2018), https://api.semanticscholar.org/CorpusID:150306265
11. Sohrab, H.H.: Basic Real Analysis. Birkhauser Basel, 2nd ed edn. (2014)
12. Tappin, B.M., Gadsby, S.: Biased belief in the bayesian brain: A deeper look at the evidence. Consciousness and Cognition **68**, 107–114 (2019). https://doi.org/https://doi.org/10.1016/j.concog.2019.01.006, https://www.sciencedirect.com/science/article/pii/S1053810018305075
13. Williams, D.: Hierarchical bayesian models of delusion. Consciousness and Cognition **61**, 129–147 (2018). https://doi.org/https://doi.org/10.1016/j.concog.2018.03.003, https://www.sciencedirect.com/science/article/pii/S1053810017306219
14. X, C., P, T., J, L., T, D.B.: Opinion dynamics with backfire effect and biased assimilation. PLoS ONE **16**(9) (2021). https://doi.org/https://doi.org/10.1371/journal.pone.0256922, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0256922
15. Xia, W., Ye, M., Liu, J., Cao, M., Sun, X.M.: Analysis of a nonlinear opinion dynamics model with biased assimilation. Automatica **120**, 109113 (2020). https://doi.org/https://doi.org/10.1016/j.automatica.2020.109113, https://www.sciencedirect.com/science/article/pii/S0005109820303113