Research Article

An Agent-Based Model of Opinion Polarization Driven by Emotions

Frank Schweitzer,1,2 Tamas Krivachy,1 and David Garcia2,3

1Chair of Systems Design, ETH Zurich, Weinbergstrasse 58, 8092 Zurich, Switzerland
2Complexity Science Hub Vienna, Josefstädter Strasse 39, 1080 Vienna, Austria
3Section for Science of Complex Systems, CeMSIIS, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria

Correspondence should be addressed to Frank Schweitzer; fschweitzer@ethz.ch

Received 30 August 2019; Revised 28 January 2020; Accepted 18 March 2020; Published 10 April 2020

Academic Editor: Giulio Cimini

Copyright © 2020 Frank Schweitzer et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We provide an agent-based model to explain the emergence of collective opinions not based on feedback between different opinions, but based on emotional interactions between agents. The driving variable is the emotional state of agents, characterized by their valence, quantifying the emotion from unpleasant to pleasant, and their arousal, quantifying the degree of activity associated with the emotion. Both determine their emotional expression, from which collective emotional information is generated. This information feeds back on the dynamics of emotional states and individual opinions in a nonlinear manner. We derive the critical conditions for emotional interactions to obtain either consensus or polarization of opinions. Stochastic agent-based simulations and formal analyses of the model explain our results. Possible ways to validate the model are discussed.

1. Introduction

In the past decades, the significance of emotions in opinion formation and decision making has been recognized by the scientific community, and its study is mainly pioneered by the field of behavioral economics and empirical psychology. Experimental research on individual behavior shows that emotions fuel information sharing [1], that emotional arousal drives reactions to become more extreme [2], and that emotional states frame the way we process information [3]. Recent results on observational data analysis suggest that collective opinions and decisions, such as election outcomes or pricing dynamics, can be influenced by emotional states like collective mood. A notable example has been the analysis of mood in Twitter and the stock market [4, 5], which led to further studies of Twitter emotions in pricing of other assets like Bitcoin [6].

The modelling of emotions and opinions has mainly focused on their interaction at the individual level, modelling how affective and cognitive mechanisms influence each other [7]. This has left aside the modelling of the interplay between emotions and collective opinions, mainly due to the absence of data to test those models and the technical challenge to formulate and test hypotheses on them. Our aim in this article is to fill that gap, proposing a computational model of collective opinions and emotions. Such a model, when designed based on plausible and testable assumptions, can be used as a hypothesis generator to guide future empirical research [8].

We apply the principles of a modelling framework of collective emotions in which the collective state arises through interactions via a common information field and not through one-on-one interactions [9]. The individual dynamics of this model have been calibrated against results from self-reports [10] and empirical data on the dynamics of emotional interaction [11]. We model how the opinions of agents are influenced by a common information field accessible to all agents, capturing this way how opinions evolve as a function of shared emotions. We present analytical and numerical results on this model that constitute stylized facts that can be tested in future empirical studies.
2. Background

Emotions are psychological states of high relevance for the individual that imply cognitive and physiological effects. They are closely related to our behavior and how we interpret our own actions [12]. Current research in affective science conceives emotions through their measurable components or dimensions, proposing models to quantify and measure them. Among these models, the component process model of emotions [13] conceptualizes emotions as composed of sequential appraisal processes, such as evaluating if an event is positive or adverse, and parallel processes that include physiological dynamics, action tendencies, and subjective feelings.

In terms of measurement of emotions, the circumplex model of core affect [14] focuses on emotions on a short timescale in which they do not have a clear target. Emotions in the circumplex model consume energy in the organism and strongly affect cognition and action tendencies, like verbal expression. The circumplex model provides the quantitative basis for our approach: emotions are quantified by a dimension of valence (pleasure associated with the emotion) and of arousal (degree of activity induced by the emotion). These two dimensions allow the mapping of emotions in a circle that captures a large amount of the variance of emotional experience [15]. While further dimensions are informative of the experience of emotions, like potency and unpredictability [16], valence and arousal have prevailed as the two major factors to measure short-lived emotional states. Recent research has modelled emotions as dynamical systems. For example, Sander et al. [17] modelled emotional fight-or-flight reactions as cusp catastrophes, and empirical studies have calibrated the internal relaxation and variation dynamics of emotions against empirical data [10, 18, 19].

The emotions of humans do not exist in isolation and often collective emotional states are triggered or emerge in a crowd. Collective emotions are defined as emotional states shared by large amounts of people at the same time [20]. Research on collective emotions, while learning from established works from social psychology and sociology, is still a growing field [21, 22]. The hyperlens model of social regulation of emotion is an adaptation of previous models of social factors of emotions to capture collective aspects of emotional life [23]. This model calls research to “get out of the lab” and investigate collective emotions in the naturalistic scenarios where they appear. Studies on shared emotions in collective gatherings have shown the long-term effects of these collective emotions for the feelings of social cohesion and identity of those involved [24]. Similarly collective emotions and group-based emotions play key roles in intergroup conflicts [25].

The availability of data produced by the digital society motivated the study of collective emotions in online communities and social media. Collective emotions have been analyzed through sentiment analysis of real-time group chats [26], product reviews [27], and of forum discussions [28]. To understand their emergence and dynamics, the Cyberemotions modelling framework was designed to simulate collective emotions in a vast variety of online media [9]. The Cyberemotions framework was designed to be testable in both controlled experiments and observational analyses of digital traces, following a wider trend of calibrating agent-based models against large-scale datasets [29]. Models in the Cyberemotions framework have been used to explain polarization of emotions in product reviews [27], collective emotions in chatrooms [26, 30] and in blogs [31], and emotion spreading in the MySpace social network [32]. Further applications of this framework have focused on modelling how bots and dialog systems could drive collective mood in a discussion [33–35] and to drive the emotion dynamics of 3D virtual humans in real-time interaction [36]. We follow the Cyberemotions framework in our model design to produce generalizable and testable formulations that can be put in reference to previous empirical and analytical results.

While such results from data-driven modelling of emotions were quite convincing, recent proposals to formally relate emotion dynamics to opinion dynamics have received much less evidence and support. Opinion dynamics itself is an established field of research and one of the areas, where methods from statistical physics have been successfully applied to model social phenomena [37, 38]. A highly relevant question tackles the emergence of consensus and polarization, motivating bounded confidence models [39–41], information accumulation systems in modular networks [42], and studies of the role of biased assimilation and assortativity in the polarization of opinions [43, 44].

To link emotions and opinion polarization, recent computational models simply rephrased the dimension of valence as opinion, to then study cusp catastrophes of state changes depending on arousal [45, 46] and on tolerance parameters inherited from bounded confidence models [47]. To date, what is missing is a model that includes both the fast dynamics of emotional states and the slower dynamics of opinions in an integrated approach that can explain the role of emotions in opinion polarization. This needs to be done based on principles testable in psychological studies and observational analyses, rather than recasting previous models by simply replacing the terminology of opinions for the one of emotions. We aim to fill this gap by formulating such model and providing an analysis of its dynamics, opening new research questions for future empirical research.

3. Agent-Based Model

3.1. Modeling Emotional Dynamics. The main focus of our model is to explain the evolution of opinions based on emotions. Precisely, we do not assume that agents respond to the opinions of other agents, directly. Instead, we assume that the expression of opinions is tightly coupled to the expression of emotions, and these emotions in fact influence other agents. This allows us to build on our previously developed agent-based framework of emotional influence [9]. It has proven to describe collective emotional states in different social systems [26, 27, 31, 35].
3.1.1. Emotional Information. Here, we only recap the core dynamics of our agent-based framework of emotional influence, schematically shown in Figure 1. The horizontal layer represents the agent. Its emotion is characterized by the two dimensions valence $v$ and arousal $a$. Both determine the content of the written expression $s$, which contributes, e.g., to an online discussion, as follows:

$$s_i(t) = s \cdot \text{sign}(v_i(t)) \cdot \Theta[a_i(t) - \tau_i].$$  \hspace{1cm} (1)$$

Agent $i$ makes an individual contribution $s_i$ if its arousal $a_i(t)$ exceeds a given individual threshold $\tau_i$. $\Theta[x]$ is the Heaviside function: $\Theta[x] = 1$ only if $x \geq 0$. The threshold $\tau_i$ is a random value selected uniformly from an interval $[\tau_{\text{min}}, \tau_{\text{max}}]$. This assumption is aligned with previous models in this framework [9] and matches empirical results in which the probability to participate in empirical discussions grows steadily with arousal after a minimal value [11]. If an agent expresses itself, the emotional content of that contribution results from its valence $v_i$ at that time, precisely from the sign of its valence, i.e., positive or negative, but the amount of the contribution, $s$, is a parameter equal for all agents.

The emotional information generated this way by all agents participating is contained in the information field $h(t)$. It consists of two components, $h_+(t)$ and $h_-(t)$, which contain the emotional expressions with positive or negative valence, respectively. The dynamics for each component is given by

$$\frac{dh_k}{dt} = -\gamma_v h_k(t) + sN_k(t) + I_k(t),$$  \hspace{1cm} (2)$$

where $h_+$ represents either $h_+$ or $h_-$. It is increased by all agents, $N_k(t)$, that make a positive/negative contribution ($s$) at time $t$, but can also decay over time at a rate $\gamma_v$, which reflects the decay of attention for older information. For simplicity, we assume that $s$ is the same across agents, but future models can include heterogenous impact following the patterns captured by Social Impact Theory [48, 49]. $I_k(t)$ captures emotional information resulting from external influences, e.g., influences from the media, but is neglected in the following.

We further define

$$h(t) = h_+(t) + h_-(t);$$

$$\Delta h(t) = h_+(t) - h_-(t).$$  \hspace{1cm} (3)$$

This way, the total information field $h(t)$ accumulates the emotional expressions of all agents in a weighted manner, i.e., with some memory because of the exponential decay. $h(t)$ can be seen as a measure of the activity of the agents, as it directly builds on their contributions. On the other hand, $\Delta h(t)$ is a measure of the average emotional charge, i.e., the average valence of the information field.

Dynamics of valence and arousal: As indicated in Figure 1, the information field further influences the agent, affecting the individual valence and arousal. Hence, the vertical layer represents the indirect emotional communication between agents, i.e., their coupling by means of the emotional information field. This requires us to determine how the individual valence $v_i(t)$ and arousal $a_i(t)$ are affected by the emotional information $h$. For the dynamics of both variables, we have proposed stochastic equations that follow the modelling framework of Brownian agents [50], i.e., we have a superposition of deterministic and stochastic influences:

$$\frac{dv_i}{dt} = -\gamma_v v_i(t) + g_v + A_v \xi_v(t),$$

$$\frac{da_i}{dt} = -\gamma_a a_i(t) + g_a + A_a \xi_a(t).$$  \hspace{1cm} (4)$$

Here, $A_v$ and $A_a$ denote the strength of the stochastic influences, whereas $\xi_v(t)$ and $\xi_a(t)$ are numbers randomly chosen from a standard normal distribution. The damping constants $\gamma_v$, $\gamma_a$ ensure that in the absence of any influences, both valence and arousal of an agent relax in the course of time (silent mode). The terms $g_v$ and $g_a$ are nonlinear functions to capture the influence of valence and arousal, and the available emotional information $h$ as given in equation (3).

To specify them, we use the general form of a power series:

$$3g_v[h(t), v_i(t)] = h(t) \sum_{k=0}^{n} b_k v_i^k(t) = h(t)[b_0 + b_1 v_i(t) + b_2 v_i^2(t) + b_3 v_i^3(t)],$$

$$g_a[h(t), a_i(t)] = h(t) \sum_{k=0}^{n} d_k a_i^k(t) = h(t)[d_0 + d_1 a_i(t) + d_2 a_i^2(t)].$$  \hspace{1cm} (5)$$

This approach matches the observed dynamics in experiments of emotion interaction in online media [11]. Higher exponents in the above polynomials were not found to be informative in regression analysis of emotion changes during online interaction, and the value of parameters in the above equation can be chosen with respect to the observed empirical dynamics.

Regarding the sign and value of the coefficients $b_k$, $d_k$, we just summarize the detailed discussion in [9]. For the dynamics of valence, we have considered contributions up to
3rd order. If we want to ensure a “silent mode,” \( b_0 = 0 \) has to be chosen. Further, \( b_2 = 0 \) if we want to avoid a built-in preference for either positive or negative valence values. With this, we have for the dynamics of valence \([9]\)

\[
\frac{dv_i(t)}{dt} = [b_i h_k - \gamma_v] v_i(t) + b_i h_v \xi_i(t) + A_v \xi_v(t) \tag{6}
\]

To prevent a valence explosion, in this dynamics \( b_1 < 0 \), i.e., a saturation behavior for large valence values, has to be chosen. With this, we further see that nontrivial solutions \( v \neq 0 \) can be only obtained if \( b_1 > 0 \). Specifically, \( b_1 > \gamma_v/h_v \) has to be reached. This condition is likely met for large values of the emotional information \( h_v \), i.e., already from this condition, we can expect a transition toward a strong emotional regime (characterized by large absolute values of valence) if the emotional information is sufficiently large.

For the dynamics of arousal, we have considered contributions up to 2nd order, i.e.,

\[
\frac{da_i(t)}{dt} = [d_1 h(t) - \gamma_a] a_i(t) + h(t)\{d_0 + d_2 a_i^2(t)\} + A_a \xi_a(t) \tag{7}
\]

Our model requires a small initial positive bias, \( d_0 > 0 \), in order to start the communication process in the absence of previous interactions. To allow for solutions characterized by at least two different activity levels (low and high arousal), we further have to choose \( d_1 \neq 0 \) and \( d_2 \neq 0 \). The expected collective dynamics then very much depends on the sign of \( d_2 \). For \( d_2 < 0 \), the arousal dynamics becomes saturated. If this saturation level is above the individual threshold \( r_i \), the agent will generate an emotional expression and \( a_i \) is set back to zero afterwards. If, at that point, fluctuations push the agent’s arousal to negative values, it will not return to positive arousal again. Hence, we obtain a scenario where agents express their emotions most likely only once. This may lead to collective emotions, but not repeatedly.

For \( d_2 > 0 \), however, we may obtain two different stationary solutions with negative arousal. At low levels of emotional information \( h(t) \), e.g., after some silent periods, fluctuations are able to push the agent’s arousal to positive values, from where a new communication cycle starts. Hence, we obtain a scenario in which waves of collective emotions over time can be expected.

We note that in both cases, fluctuations play an important role in establishing an active regime. They first push agents to a positive arousal which is then amplified by the positive feedback, until it reaches the threshold. This then generates emotional expressions that establish a communication field which in turn feeds back on the agent’s valence and arousal. In our model, valence decides about the “content,” determining the sign of the emotional information generated. Arousal, on the other hand, decides about the activity pattern.

3.2. Modeling Opinion Dynamics. We now have to specify how the emotional interaction described above influences the dynamics of opinions. We will proceed in two steps: first we introduce the dynamics of opinions and subsequently the coupling between opinions and emotions.

Specifically, our model shall explain the polarization of opinions. In politics, such a polarized state can be illustrated by the opposition between democrats (D) and republicans (R), which each represent particular opinions towards a given set of subjects. Obviously, in this context, the opinion \( \theta(t) \) of a particular agent \( i \) at time \( t \) cannot be represented as a binary variable, e.g., \( \theta_i \in [-1, +1] \), as this would imply that a disagreement with \( R \) automatically translates into an agreement with \( D \). Also, to extend the opinion space to include a neutral opinion, e.g., \( \theta_i \in [-1, 0, +1] \), does not really solve the problem, as it neglects the heterogeneity of agents with respect to their opinions. Therefore, a continuous variable \( \theta_i \in (-1, +1) \) would be the most appropriate approach.

Given that opinions are continuous variables, consensus has to be defined as a certain (narrow) range of opinions or, more precisely, as a certain distribution of opinions, \( \mathcal{P}(\theta, \sigma^2) \), with a mean value \( \bar{\theta} \) and a (small) variance \( \sigma^2 \). Polarization, on the other hand, should be defined by a bimodal distribution of opinions \( \mathcal{P}(\theta) \), with a large variance \( \sigma^2 \). The two distant peaks represent the polarized opinions, while modest opinions in the middle range are less frequent.

To specify the dynamics of opinions, we start with the general ansatz of a power series already employed in equation (5),

\[
\frac{d\theta_i(t)}{dt} = \sum_{k=0}^{n} a_{\theta} \xi_k^i(t) + A_{\theta} \xi_\theta(t), \tag{8}
\]

which postulates a nonlinear feedback of the current opinion of an agent on the change of this opinion. The power series accounts for the fact that we need additional assumptions that later can be encoded in the coefficients \( a_{\theta} \), \( A_{\theta} \xi_\theta(t) \) is a stochastic term, to consider random influences on the opinion.

For our discussion, we will use terms up to 3rd order from the power series. Neglecting individual differences and stochastic influences for the moment, we can express the opinion dynamics as

\[
\frac{d\theta(t)}{dt} = a_0 + a_1 \theta(t) + a_2 \theta^2(t) + a_3 \theta^3(t). \tag{9}
\]

To discuss the possible stationary solutions, \( d\theta/dt \) of equation (9), for the moment we assume that \( a_0 \longrightarrow 0 \) and \( a_2 \longrightarrow 0 \). Then, we find

\[
\bar{\theta}(1) = 0; \quad \bar{\theta}(2,3) = \pm \sqrt{\frac{a_1}{a_3}}. \tag{10}
\]

I.e., to obtain nontrivial values for the opinion, \( a_1 \) and \( a_3 \) have to be of opposite sign. The two possible cases are illustrated in Figure 2, where \( d\theta/dt \) is plotted against \( \theta \) according to equation (9). We note that for \( a_3 > 0, a_1 < 0 \) solutions with extreme opinions are always instable. In particular, there is a force toward the neutral opinion \( \theta = 0 \),
it would be difficult to explain polarization of opinions based on such a parameter constellation.

For \( \alpha_1 < 0, \alpha_1 > 0 \), on the other hand, we find that there can be a stable coexistence of extreme opinions. Hence, in the following, we choose \( \alpha_1 < 0 \). This also accounts for a saturation dynamics in the case of extreme opinions. With that, it is obvious that we will have only one possible solution for the opinion, i.e., \( \theta = 0 \), if \( \alpha_1 < 0 \). However, if \( \alpha_1 > 0 \), we will find bimodality, i.e., always two opposing opinions, whereas the neutral opinion \( \theta = 0 \) is instable. This is the case of polarization, we are interested in. The corresponding bifurcation diagram is shown in Figure 3(a), where we vary \( \alpha_1 \) as the control parameter.

To understand the impact of \( \alpha_0 \neq 0 \) on the opinion dynamics, let us now fix \( \alpha_0 > 0 \), i.e., we allow for polarized opinions in principle. For simplicity, we still neglect \( \alpha_2 \rightarrow 0 \), but we will include a nonnegligible influence \( \alpha_2 \neq 0 \) later in the discussion. The corresponding bifurcation diagram is shown in Figure 3(b), where we keep \( \alpha_1 \) fixed, but vary \( \alpha_0 \) as the second control parameter. If \( \alpha_0 > 0 \) and \( |\alpha_0| \) is large, then we find only one stationary solution for opinion, which is negative, i.e., \( \theta^{(3)} = -\bar{\theta} \). If \( \alpha_0 < 0 \) and \( |\alpha_0| \) is large, then we find only one stationary solution for opinion, which is positive, i.e., \( \theta^{(3)} = +\bar{\theta} \).

For intermediate values of \( |\alpha_0| \), we find a regime where there is bimodality of the opinions with different weights of the positive or negative solution and instable solutions to separate these. Hence, dependent on the concrete values of \( \alpha_k \), we can expect regimes in which we find a coexistence of polarized opinions, but as well regimes where only one opinion emerges.

3.3. Nonlinear Coupling between Emotions and Opinions.

In order to couple the opinion dynamics to the emotional interactions of the agents, we need to determine how the coefficients \( a_k \) may depend on emotions. We set \( a_1(t) \propto h(t) - h_{base}, \) i.e., make it proportional to the overall activity in emotional interactions, expressed by value of the emotional field. This activity has to overcome some threshold value expressed by \( h_{base} \), in order to allow for a sufficient coupling. Considering a small baseline value \( h_{base} \), without activity, i.e., without emotional emotional information, \( a_1 \) is negative and there is only a neutral opinion, according to the above discussion. But high activity, i.e., large emotional information, will drive \( a_1 \) to positive values, to allow for bimodality.

In our model, \( \alpha_2 \) and \( \alpha_3 \) do not depend on the emotional dynamics. \( \alpha_2 \) may be close to zero. It can be positive or negative, this way generating a global bias toward left/right opinions. \( \alpha_3 < 0 \) determines the saturation level of the opinions. In agreement with equation (10), the smaller \( |\alpha_3| \), the larger the possible polarization of opinions.

The important parameter for the coupling between emotions and opinions is \( \alpha_0 \), for which we assume \( a_0(t) \propto -h(t)\bar{\nu}(t) \). If the emotional activity, expressed by \( h(t) \) is very high and the average emotion \( \bar{\nu}(t) \) is either very negative or very positive, then there is likely only one opinion, which is also very likely very negative or very positive. Precisely, if \( \bar{\nu} < 0 \) and \( h \) is high, then \( \alpha_3 > 0 \) and \( |\alpha_3| \) is large and we find \( \bar{\theta}^{(3)} = -\bar{\theta} \). If \( \bar{\nu} > 0 \) and \( h \) is high, then \( \alpha_3 < 0 \) and \( |\alpha_3| \) is large and we find \( \bar{\theta}^{(3)} = \bar{\theta} \). In accordance with Figure 3(b), for moderate values of \( \bar{\nu} \) and \( h \) we can expect bimodality, i.e., a polarization of opinions.

In conclusion, the dynamics of the opinions read now

\[
\frac{d\theta(t)}{dt} = -c_0 h(t)\bar{\nu}(t) + c_1 [h(t) - h_{base}]\theta(t) + \alpha_2 \theta^2(t) + \alpha_3 \theta^3(t) + A_\mu \tilde{\epsilon}(t),
\]

where \( c_0, c_1 \) are some proportionality constants. They are chosen between 0 and 1 to mitigate the influence of the corresponding terms.

Relations to the bounded confidence model: To further understand this dynamics, let us neglect all small or higher-order terms and focus on the core dynamics, which reads for an individual agent \( i \):

\[
\frac{d\theta_i(t)}{dt} = h(t)\{c_1 \theta_i(t) - c_0 \bar{\nu}(t)\}.
\]

For the mean opinion in the agent population, we find

\[
\frac{d\bar{\theta}(t)}{dt} = -\frac{1}{N} \sum_i \frac{d\theta_i(t)}{dt} = h(t)\left[-\frac{1}{N} \sum_i c_1 \theta_i(t) - c_0 \bar{\nu}(t)\right].
\]

This leads, in equilibrium, \( \frac{d\bar{\theta}}{dt} = 0 \), to \( \bar{\theta} = (c_0/c_1)\bar{\nu} \), and the opinion dynamics, equation (12), simplifies to

\[
\frac{d\theta_i(t)}{dt} = \mu [\theta_i(t) - \bar{\theta}(t)],
\]

which is precisely the bounded confidence dynamics with \( \mu = c_1 h(t) \). That means, the higher the activity as measured by \( h(t) \), the faster the convergence toward the mean. A noticeable difference: in the bounded confidence model, agents only interact if \( \theta_i - \theta_j < \epsilon \). In our case, all agents interact through the communication field \( h(t) \), more precisely via the mean valence \( \bar{\nu}(t) \). In the following, we make use of the fact that the average valence \( \bar{\nu} \) can be approximated by the difference in the two field components, \( h_i, h_j \), equation (3) and choose

\[
\bar{\nu}(t) = c_0 \Delta h(t),
\]

which reduces the number of variables and parameters.

Very similar to the bounded confidence model, we should expect scenarios that can lead to consensus characterized by a unimodal opinion distribution. However, as we already know from the above discussion, the inclusion of the terms \( \alpha_2, \alpha_3 \) neglected in this derivation may lead to other scenarios of coexistence characterized by a bimodal opinion distribution. Hence, while \( \alpha_0 \) and \( \alpha_1 \) are essential
to couple the opinion dynamics to the emotion dynamics, \(a_2\) and \(a_3\) may decide about the level of polarization reached.

4. Results

4.1. Agent-Based Simulations. We first present the results of agent-based computer simulations, to verify that the model works as expected. The parameter values not explicitly mentioned in the following are chosen, in accordance with [11], as follows:

\[
\begin{align*}
\gamma_v &= 0.5, \\
b_1 &= 1, \\
b_3 &= -1, \\
A_v &= 0.3, \\
\xi_v &\sim \mathcal{N}(0, 0.5) \\
\gamma_a &= 0.9, \\
d_0 &= 0.05, \\
d_1 &= 0.5, \\
d_2 &= 0.1, \\
A_a &= 0.3, \\
\xi_a &\sim \mathcal{N}(0, 0.6) \\
\gamma_h &= 0.7, \\
s &= 0.6, \\
h_{\text{base}} &= 0.1, \\
\tau_{\text{min}} &= 0.1, \\
\tau_{\text{max}} &= 1.1, \\
N &= 100, \\
\delta t &= 0.2, \\
c_0 &= 0.1, \\
c_1 &= 1, \\
A_\theta &= 0.05.
\end{align*}
\]

Figure 4 shows, in the left panel, the opinion trajectories of individual agents over time and, in the right panel, the distribution of opinions after a sufficiently long time. We note that, because the model is intrinsically stochastic, we do not reach a stationary distribution in the strict sense.

As the first observation, we find indeed the polarization of opinions that were initially close, i.e., drawn from a normal distribution \(\mathcal{N}(\mu, \sigma^2) = \mathcal{N}(0, 0.3)\). This was expected because the range of parameters was adjusted such that a polarization regime emerges. However, we emphasize that some of the \(a_k\), namely, \(a_0\) and \(a_1\), are in fact not constants, but functions of the emotional field components \(h_+, h_-\) which in turn depend on the agents’ individual valence and arousal. Hence, these “parameters” were not chosen, but their value emerged from the emotional interactions between agents. Moreover, these values are not fixed but fluctuate over time according to the emotional dynamics. This leads to the nonstationary opinion dynamics observed.

As the second observation, we see that the number of agents with positive and negative opinions can differ significantly dependent on the parameter \(a_2\). Although it is small, it generates a global preference for either left or right opinions. Consequently, we find also the emergence of a minority/majority in the agent population. This is shown in different heights of the peaks of the bimodal distributions.

Eventually, we can also obtain scenarios in which consensus is reached, i.e., instead of a bimodal opinion distribution, we find a unimodal distribution. This is illustrated in Figure 5 for a consensus around the neutral opinion and a biased opinion.

4.2. Critical Transitions toward Polarized Opinions. To fully understand the role of the coefficients \(a_k\) in equation (9), we now focus on the so-called phase portrait. Different from Figure 2, which plots \(d\theta/dt\) against \(\theta\), the phase portrait investigates \(d^2\theta/d\tau^2\) against \(d\theta/d\tau\). With equation (9), the two variables of the phase portrait follow the dynamics:

\[
\frac{d\theta(t)}{d\tau} = \kappa(\theta) = a_0 + \alpha_1 \theta(t) + \alpha_2 \theta^2(t) + \alpha_3 \theta^3(t),
\]

\[
\frac{d\kappa(\theta)}{d\tau} = \kappa(\theta) \left[ \alpha_1 + 2\alpha_2 \theta(t) + 3\alpha_3 \theta^2(t) \right].
\]

These two coupled equations can be solved numerically using a 4th order Runge–Kutta method. The results are shown in Figure 6. Stationary solutions of the opinions are given by \(\kappa(\theta) = 0\). But, as the distribution of the colored squares along this horizontal line in Figure 6 verifies, the solutions concentrate on the left/right end of the horizontal line, indicating polarized opinions. The middle-ranged values of \(\theta\) resulting from \(\kappa(\theta) = 0\) are in fact unstable stationary solutions.

The range of these unstable solutions can be obtained by setting \([\alpha_1 + 2\alpha_2 \theta(t) + 3\alpha_3 \theta^2(t)] = 0\) in equation (17). As the result, we find

\[
\kappa = \begin{cases} 
\text{stable,} & \text{if } \theta < \theta_-, \\
\text{instable,} & \text{if } \theta_- < \theta < \theta_+, \\
\text{stable,} & \text{if } \theta_+ < \theta,
\end{cases}
\]

where \(\theta_+\) and \(\theta_-\) follow from the above quadratic equation:

\[
\theta_\pm = \frac{\alpha_2}{3\alpha_3} \pm \frac{1}{3\alpha_3} D;
\]

\[
D = \sqrt{\alpha_2^2 - 3\alpha_1\alpha_3}.
\]

These values are clearly indicated in Figure 6 by the two (empty) vertical regions around \(\theta_\pm \approx \pm 0.5\) in which the arrows of the vector field change their direction.

So where do agents end up with their opinions dependent on their initial conditions, if we only consider the deterministic dynamics? This is answered by the separatrix also depicted in Figure 6 as the thick-dashed line. Agents starting with initial conditions above the separatrix will tend
to obtain a positive \( \theta \) in the long run, while agents starting with initial conditions below the separatrix will reach a negative \( \theta \). The precise equation for the separatrix is given by

\[
S(\theta) = \begin{cases} 
\alpha_0 + \alpha_1 \theta + \alpha_2 \theta^2 + \alpha_3 \theta^3, & \text{if } \theta < \theta_-, \\
0, & \text{if } \theta_- < \theta < \theta_+, \\
\alpha_0^* + \alpha_1 \theta + \alpha_2 \theta^2 + \alpha_3 \theta^3, & \text{if } \theta_+ < \theta.
\end{cases}
\]

(20)

The only difference in the expression is in the values \( \alpha_0 \), \( \alpha_0^* \), which, when taken at first order in \( \alpha_2 \), can be reduced to

\[
\alpha_0^* = \frac{-2\alpha_1^2 + 9\alpha_1 \alpha_3 \Delta \pm 2\alpha_2 \Delta + 6\alpha_1 \alpha_3 D}{27\alpha_0^2} = \frac{3\alpha_1 \alpha_3 \pm 2\alpha_2 D}{9\alpha_3}
\]

(21)

It is important to notice that the dynamics captured by the phase portrait, shown in Figure 6, does not depend on \( \alpha_0 \). The coefficient \( \alpha_0 \) merely selects a solution curve and can be seen as a vertical shift in the diagram.

5. Discussion

In this paper, we have provided a model to formally link the dynamics of emotions to the dynamics of opinions. We follow an agent-based approach, that is, we focus on the emotions and opinions of individual agents with heterogeneous properties. Our research interest is to explain the emergence of collective opinions based on the emotional interaction between agents. More specifically, we want to understand under which conditions we obtain consensus, i.e., the emergence of one common opinion (reflected in a narrow opinion distribution), or polarization, i.e., the emergence of two opposing common opinions (reflected in a bimodal opinion distribution). As the interaction mechanism, we do not assume a feedback between different opinions, which would be the most simple way to obtain the two distinct opinion distributions. Instead, our main assumption is that the dynamics of opinions is driven by the dynamics of emotions.
Following established measures from social psychology, the dynamics of emotions is characterized by two agent variables, \textit{valence}, the pleasure associated with emotions and \textit{arousal}, the activity associated with emotions. Both variables determine the emotional expression of agents, from which collective emotional information is generated. This is quantified by two aggregated and time-dependent variables, the emotional field $h(t)$ and the average valence, $v(t)$. These two variables from the emotional interaction feedback on the individual opinion dynamics in a nonlinear manner. Our formal model makes transparent under which critical conditions for the emotional interaction, we can expect a polarization of opinions, without assuming a direct interaction on the level of opinions.

Our modelling approach fits a general framework to model active matter \cite{38}, a term to denote systems with the ability of self-organization, active motion, and structure formation, provided a critical supply of energy. In our case, the driving variables describe the emotional state, \{$a_i(t), v_i(t)$\} composed of valence and arousal, whereas the driven variable is the individual opinion, $\theta_i(t)$. Different from other approaches, our model respects the fact that emotions and opinions evolve on two different time scales. Emotions relax faster than opinions, i.e., they evolve on a slower timescale than opinions. 

**Figure 4**: (a, b) Opinion trajectories. (c, d) Opinion distribution at $t = 100$. (a, c) $\alpha_2 = 0$, (b, d) $\alpha_2 = 2$, other parameters: $\alpha_3 = -5$.

**Figure 5**: Opinion trajectories for the case of consensus. (a) $\alpha_2 = 0$, hence the neutral opinion $\theta = 0$ is obtained. (b) $\alpha_2 = 4$; hence, a global bias toward positive opinions exists.
analyzing techniques, we can estimate the polarization of a
discussion in terms of the simultaneous presence of
positive and negative expressions [58]. For very big and
longitudinal data, we can obtain the final distribution of
opinions \( P(\theta) \), to calculate the overall polarization.
Changes in the distribution and the subsequent po-
larization measure could then be related to changes in
the emotional information.

Polarization from the existence of network compo-
ents: The online communication between users can be
represented as a social network, on which we can
perform a community analysis, to detect communities
with different opinions [6]. This method makes sense if
links signal agreement or endorsement, like retweets or
follower links.

Extension to multi-dimensional opinion space: Our
model so far considers that opinion is a scalar, i.e. there is
only one opinion per agent, which is with respect to one
subject only. This one-dimensional description already
grafts the dichotomy in favor/against into the opinion
space. In most real situations, however, one could agree
with individual \( i \) on one particular subject and with \( j \)
on another one. Hence, more complex mixed opinions
should be possible in an extension of our model. For this we could
redefine the opinion of agent \( i \) as a vector \( \theta_i(t) = [\theta_{i1}(t), \theta_{i2}(t), \ldots] \), where \( \theta_{in}(t) \) expresses
the opinion of agent \( i \) towards subject \( n \) at time \( t \). It then
depends on the context how emotions drive opinions in
different dimensions.

We emphasize that a multi-dimensional opinion space
exacerbates the problem of consensus, i.e. the convergence
toward a common opinion. Convergence along one di-
ension does not necessarily implies agreement on other
subjects. On the contrary, agents which agree on given
subjects often choose to disagree on other subjects, to
distinguish themselves from other agents. Hence, we
expect that instead of consensus we frequently find co-
existence of (mixed) opinions. In developed democracies
such as Switzerland, this has lead to the emergence of a
political space with many parties coexisting, which opens
the possibility to form alliances regarding certain deci-
sions [6].

It is an open question how to define a multidimensional
polarization measure based on mixed opinions. This will be
addressed in a subsequent publication. Another open
question regards the link of such measures to available data.
For this, we could consider a topic model to reduce the
dimensionality of the opinion space. With a smaller number
of opinion dimensions, we can then test correlations to
identify additional polarization levels.

Data Availability

The manuscript does not have any data.

Conflicts of Interest

The authors declare that they have no conflicts of interest.
Acknowledgments
All authors acknowledge funding from the Swiss National Science Foundation (CR2111_146499). D.G. acknowledges funding from the Vienna Science and Technology Fund through the Vienna Research Group Grant "Emotional Well-Being in the Digital Society" (VRG16-005).

References


