entropy

MDPI

Article

Modelling Spirals of Silence and Echo Chambers by Learning from the Feedback of Others

Sven Banisch [1,2,*], Felix Gaisbauer [2] and Eckehard Olbrich [2]

[1] Institute of Technology Futures, Karlsruhe Institute of Technology, 76133 Karlsruhe, Germany
[2] Max Planck Institute for Mathematics in the Sciences, 04103 Leipzig, Germany
* Correspondence: sven.banisch@universecity.de

Abstract: What are the mechanisms by which groups with certain opinions gain public voice and force others holding a different view into silence? Furthermore, how does social media play into this? Drawing on neuroscientific insights into the processing of social feedback, we develop a theoretical model that allows us to address these questions. In repeated interactions, individuals learn whether their opinion meets public approval and refrain from expressing their standpoint if it is socially sanctioned. In a social network sorted around opinions, an agent forms a distorted impression of public opinion enforced by the communicative activity of the different camps. Even strong majorities can be forced into silence if a minority acts as a cohesive whole. On the other hand, the strong social organisation around opinions enabled by digital platforms favours collective regimes in which opposing voices are expressed and compete for primacy in public. This paper highlights the role that the basic mechanisms of social information processing play in massive computer-mediated interactions on opinions.

Keywords: social dynamics; group dynamics; spiral of silence; echo chambers; silent majorities; reinforcement learning; social feedback; social neuroscience; opinion dynamics

MSC: 91D30; 91F10; 00A69

Citation: Banisch, S.; Gaisbauer, F.; Olbrich, E. Modelling Spirals of Silence and Echo Chambers by Learning from the Feedback of Others. *Entropy* **2022**, *24*, 1484. https://doi.org/10.3390/e24101484

Academic Editor: Federico Vazquez

Received: 26 July 2022
Accepted: 12 October 2022
Published: 18 October 2022

1. Introduction

A better understanding of the collective processes underlying public opinion expression is crucial for a better understanding of modern society. Sociological models drawing on network science [1–3] and basic principles of human interaction behaviour [4,5] have already provided useful insights into collective phenomena related to mass mobilisation [6–8], societal-level change of behaviour [9–11] and beliefs [12]. However, for a change to happen and for a movement to gain pace, the alternative must be voiced by a sufficiently large group [13]. Furthermore, to be voiced, it must be perceived as something that can be said without "fear of isolation" [14].

The spiral of silence theory [15] is based on the old "law of opinion" ([14] John Locke). It focuses on the collective perception of what can be publicly voiced and hence impact the further perception of public opinion. The theory assumes that humans possess a "quasi-statistical organ" [16] to perceive what can be said without being socially sanctioned and explains public opinion dynamics as a spiralling process in which silence may lead to more silence. In this paper, we propose a mathematical model for this process based on reinforcement learning (RL) by social feedback [17]. In repeated games played over a network, agents receive signals of approval or disapproval for expressing their opinion to peers. Agents evolve an expectation about the social reward obtained when expressing their opinion and remain silent if they expect punishment (negative reward). In this way, social feedback dynamics naturally capture the assumed "quasi-statistical" perception of the opinion landscape surrounding an agent.

Our paper develops a computational model that captures the basic assumptions of the spiral of silence theory and grounds it in neuroscientific research on social information processing. While the spiral of silence theory frequently refers to the "social nature of man" [18], no attempt of grounding this assumption in social psychology and neuroscience has been made. On the other hand, the potential for explaining collective behaviour based on mechanisms identified in cognitive and social neuroscience is frequently emphasized [19–21], but its integration with sociological theories of collective opinion expression [6,7,15] is lacking. The social feedback theory (SFT) bridges this gap by formulating collective processes of opinion expression as a multiagent problem in which individual agents adapt according to a reward- and value-based learning scheme identified in neuroscientific research [22–27]. With these repeated opinion expression games, the SFT provides a coherent framework for modelling collective opinion processes that integrates basic neuroscientific findings, adaptive decision-making [28] and a political theory of public opinion [15,16].

Given that the processing of and learning by social feedback is so deeply rooted in the human brain, it is of the uttermost importance to better understand the collective consequences of these processes. Especially in social media environments, a tremendous number of quick feedback decisions is made day by day by billions of users. "Like buttons" and quantitative markers of collective endorsement can be associated with low cognitive costs, which suggests that a dominant role is played by the fast value-processing mechanisms accounted for by RL. Recent studies have provided evidence for that [29,30]. While it is reasonable to assume that the social reward circuit has evolved to facilitate cohesion and cooperation in small groups [20] with intensive pair-bonding [31], this reasoning may not apply for societies of increased complexity [32,33]. In complex social networks, the human ability to coordinate with in-groups may come at the expense of an increasing alienation to out-groups and therefore drive polarization dynamics [17]. Here, we show that social feedback dynamics provide a neurobiologically grounded explanation of collective processes involved in "spirals of silence" [15] and analyse how structural transformations enabled by social media give voice to groups that previously went unheard.

Providing a mechanism-based approach [5] to model the phenomena of collective opinion expression or silence enables a more general application of the assumptions underlying the spiral of silence theory. Most importantly, our model allows us to relate structural variations across different opinion groups to different regimes of collective opinion expression. In this regard, we show that social feedback mechanisms may lead to spirals of silence in unstructured random networks, but that the same mechanism generates highly active echo chambers if social networks become more assortative and homophilous with respect to opinions.

2. Model

2.1. Social Feedback Processing in the Brain

Our modelling choices are well-grounded in neuroscientific insights into human social nature. Social neuroscience aims to identify neural mechanisms involved into the processing of social cues. fMRI studies have shed light on the interaction and interconnectedness of different brain regions and their functional role in social cognition. While it has long been controversial whether human nature evolved a neural circuity specifically for handling social information or not [20,24,31,34], it is now relatively settled that a basic "reinforcement circuit" [23,35] is strongly involved into value-based decisions and learning from social feedback [29,30,36–39]. Other brain processes interfere with this circuity [20,23,40], especially when social situations and tasks involve higher cognitive functions such as trust [35], morality [41] or representations of self and the other [42,43].

Temporal difference reinforcement learning (TDRL) [44,45] has provided a useful computational account of the brain mechanisms underlying social reward processing and learning [24–27]. In TDRL, a new estimate of the expected value Q^{t+1} associated with an action is a function of the current estimate Q^t and the temporal difference (TD) error δ^t between this estimate and the reward that is actually obtained: $Q^{t+1} = Q^t + \alpha\delta^t$. With

a rate governed by α (referred to as learning rate), this scheme converges to a stable equilibrium in which the TD error δ^t approaches zero such that the expectations and actual reception of rewards are aligned [45]. The usefulness of TDRL in computational neuroscience derives from the finding that the activity of dopaminergic neurons in the midbrain regions is quantitatively related to the "reward–prediction error" [20] between the experienced reward and its expected value [22,46,47], that is, to δ_t. Social neuroscience has provided ample evidence that such a basic reward processing circuit is also highly involved into peer influence processes [29,38], social conformity [37] and approval [39].

2.2. Opinion Expression Games

We consider the situation that two groups with different standpoints on a controversial issue have evolved and engage in public discourse. In contrast to most existing opinion dynamics models, we consider that the opinions of agents are fixed, because we want to understand the conditions under which agents with a given opinion become silent. Individuals within both opinion groups have two available actions: they can decide to express (E) their standpoint or to be silent (S). They receive supportive feedback from their respective in-group and negative feedback from agents in the out-group when expressing their opinion. Individual interaction is hence formulated as repeated opinion expression games with a reward system that captures approval and disapproval by peers:

$$r_i^t = \begin{cases} -c & \text{silent neighbour} \\ -c+1 & \text{agreement} \\ -c-1 & \text{disagreement} \end{cases} \tag{1}$$

The parameter c corresponds to a fixed cost of opinion expression and i refers to the individual agent. Having received a social reward during an interaction, agents update the expected value $Q_i(A)$ of their current action by TDRL

$$Q_i(A)^{t+1} = Q_i(A)^t + \alpha \underbrace{(r_i^t - Q_i(A)^t)}_{\text{TD error}} \tag{2}$$

with learning rate α. As the reward of silence (S) is zero in the game, we only have to keep track of the value for the opinion expression and skip action indices in the sequel ($Q_i(E) = Q_i$). Q_i hence accounts for the subjective reward that agent i expects when expressing their opinion, and the agent will remain silent if this value is negative. Given the current value of opinion expression Q_i an agent has learned in previous interactions, the probability of opinion expression follows a softmax choice model of the form

$$p_i = \frac{1}{1+e^{-\beta Q_i}} \tag{3}$$

in which β governs the rate of exploration. Taken together, the action selection (3) and the TDRL scheme (2) naturally account for the effect that agents become more (less) willing to speak out after receiving positive (negative) feedback.

2.3. Group Setting

Assume that we can characterise the two opinion groups G_1 and G_2 in terms of their sizes (N_1 and N_2), their in-group cohesion and intergroup connectivity. The probability of in-group influence is q_{11} for group 1 and q_{22} for group 2. The interaction probability across groups is denoted by q_{12}, q_{21}, respectively. We assume that these interaction probabilities are equal for all agents within the same opinion group. Following a mean-field approach, we derive a dynamical system governing the average behaviour of agents in G_1 and G_2. That is, we are interested in the average values of opinion expression $Q_1 = \frac{1}{N_1}\sum_{i\in G_1} Q_i$ and $Q_2 = \frac{1}{N_2}\sum_{i\in G_2} Q_i$ and their evolution. For further details and a mathematical justification of this group-level description the reader is referred to [48].

Given the group sizes N_1 and N_2 and the homogeneous interaction probabilities q_{11}, q_{22}, q_{12} and q_{21}, we define the *structural strength* of G_1 and G_2 (denoted as γ and δ) as

$$\gamma = \frac{(N_1 - 1)}{N_2} \frac{q_{11}}{q_{12}} \text{ and } \delta = \frac{(N_2 - 1)}{N_1} \frac{q_{22}}{q_{21}}. \tag{4}$$

The structural strength of a group is determined by the relative size of the group and the relative in-group connectivity or *cohesion* [49,50]. As γ and δ determine the probability of in-group versus out-group interaction ($\gamma/(\gamma+1)$ versus $1/(\gamma+1)$ for group 1), they also govern the expected rewards for opinion expression for the two groups with

$$\mathbb{E}(r_1) = p_1 \frac{\gamma}{\gamma+1} - p_2 \frac{1}{\gamma+1} - c, \tag{5}$$

for opinion group G_1 and

$$\mathbb{E}(r_2) = p_2 \frac{\delta}{\delta+1} - p_1 \frac{1}{\delta+1} - c, \tag{6}$$

for G_2. Note that the probabilities for opinion expression p_1 and p_2 are given by (3) substituting the agent index i by the respective group index. As an example, consider an agent in G_1 when expressing its opinion (Equation (5)). With a probability of $\frac{\gamma}{\gamma+1}$, the agent's neighbour will be in G_1 as well and provide positive feedback with probability p_1. With probability $\frac{1}{\gamma+1}$, the agent will meet a neighbour in the opposing opinion group G_2 and receive negative feedback when agents in G_2 are expressive (i.e., with p_2).

As visible in Equation (2), in TD learning the change of Q-values from one time step to the other is given by the TD error times the learning rate α. Similarly, at the group level the update of the Q-values Q_1 and Q_2 from one time step to the next can be written as

$$\begin{aligned} \Delta Q_1 &= Q_1^{t+1} - Q_1^t = \alpha(\mathbb{E}(r_1) - Q_1) \\ \Delta Q_2 &= Q_2^{t+1} - Q_2^t = \alpha(\mathbb{E}(r_2) - Q_2) \end{aligned} \tag{7}$$

where we introduce the expected group rewards $\mathbb{E}(r_1)$ and $\mathbb{E}(r_2)$ derived above. In the continuous time limit [51–53], we replace $t+1$ by $t+\delta t$ and α by $\alpha\delta t$ and take $\delta t \to 0$. This allows us to describe the model dynamics as a system of two differential equations

$$\begin{aligned} \dot{Q}_1 &= \mathbb{E}(r_1) - Q_1 \\ \dot{Q}_2 &= \mathbb{E}(r_2) - Q_2, \end{aligned} \tag{8}$$

where we can omit the prefactor α by rescaling time. (When performing the continuous time limit, we have to rescale the learning rate α' with $\alpha' = \frac{\alpha}{\delta t}$. Thus, the equations would have the form $\dot{Q}_g = \alpha'(\mathbb{E}(r_g) - Q_g)$. Without losing any generality, we can set $\alpha' = 1$ by rescaling time.) As the right hand side is zero when the Q-value estimate is equal to the expected reward, the fixed points of (8) are possible equilibria of the associated collective game. (A game-theoretic analysis of the model was presented in [48]).

3. Applications

3.1. Organized Minorities and Silent Majorities

We applied this model to a minority–majority setting in which one third of the population supports opinion 1 and the other two thirds hold the majority view opinion 2. The group size ratio $(N_2 - 1)/N_1$ approached two for a large number of agents. In the first scenario, the interaction probabilities were homogeneous over the entire population ($q_{11} = q_{22} = q_{12} = q_{21} = q$). This corresponded to the Erdős–Rényi random graph [54,55] with link probability q and represented a situation without any particular organisation of social relations within and in between both camps. The structural strength indicators (4) were then determined by the relative group sizes: $\gamma = 1/2$ and $\delta = 2$. For example,

consider that there are $N_1 = 100$ agents in the minority and $N_2 = 200$ in the majority group, and let $q = 0.05$ such that each agent has 15 links on average. In this unstructured case, agents from the minority are connected to 5 agents of the in-group and to 10 agents of the out-group, whereas an agent in the majority has an expected number of 10 neighbours in its own majority group and only 5 out-group connections. In this random graph setting, the dynamical system (8) has only one stable fixed point at $Q_1 \approx -0.66$ and $Q_2 \approx 0.66$. This situation is shown on the left-hand side in Figure 1, where the respective fixed point is marked by the blue circle. The associated expression rates are $p_1 \approx 0.067$ for the minority and $p_2 \approx 0.995$ for the majority. That is, the majority is expressive and the minority silent. The phase plot shows that even if expressive in the beginning (i.e., $Q_1 > 0$), agents in the minority find less and less support for their opinion and increasingly avoid expressing their opinion in public.

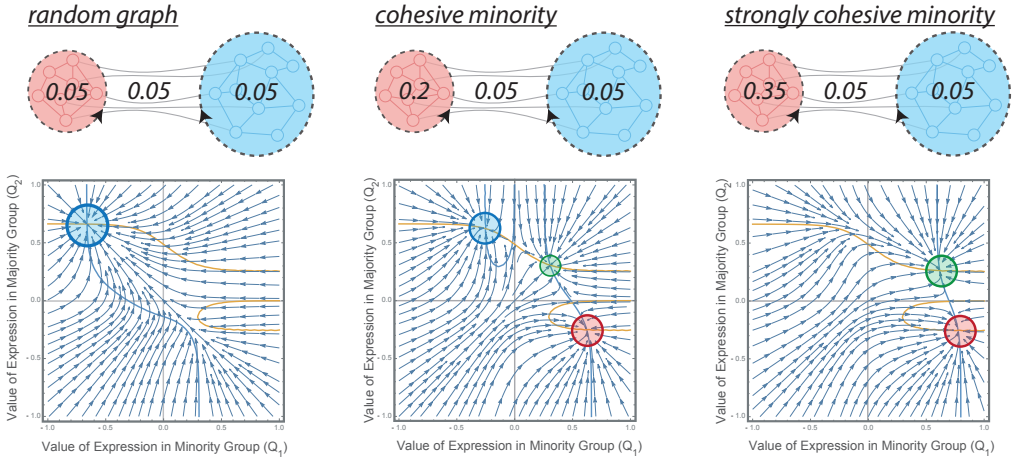

Figure 1. Two groups supporting two different opinions struggle for public expression. The majority group (blue) is twice as big as the minority group (red). The three different situations represent subsequent increases of internal cohesion of the minority group and their effect on the respective phase dynamics of the system. The phase portraits show the evolution of Q_1 and Q_2 towards the fixed points of Equation (8). The isoclines of the dynamical system are shown and the stable fixed points at their intersection are coloured according to whether the majority (blue circle), the minority (red circle), or both (green circle) are expressive. While minority expression is unstable in an unstructured random graph, the minority can compensate their quantitative inferiority by a stronger internal organisation. Results for an exploration rate $\beta = 8$ and $c = 0$. Random graph (**left**): ER graph with link probability $q_{11} = q_{22} = q_{12} = q_{21} = 0.05$. For group sizes $N_1 = 100$ and $N_2 = 200$, each agent has 15 links on average. The minority is connected to 5 agents of the in-group and to 10 agents of the out-group and vice versa for the majority leading to structural strength indicators $\gamma = 0.5$ and $\delta = 2$. The resulting system has only one stable fixed point at $Q_1 \approx -0.66$ and $Q_2 \approx 0.66$ with associated expression rates of $p_1 \approx 0.067$ and $p_2 \approx 0.995$. That is, only the majority group is expressive and the minority silent. Cohesive minority (**centre**): Increasing internal organisation of the minority group by raising the connection probability within the minority to $q_{11} = 0.2$. This increased group cohesion is reflected in an increased structural strength $\gamma = 2$. δ is not affected. The system becomes symmetric and minority (red circle) or majority group expression (blue circle) are solutions reached depending on the initial values of expression. An additional fixed point (green circle) emerges in which two groups are loud. Strongly cohesive minority (**right**): The in-group cohesion of the minority further increases ($q_{11} = 0.35$) leading to $\gamma = 4$ and $\delta = 2$. The case that only the majority is in expression mode is no longer stable and the minority will always express its opinion. Coexistence is still possible.

However, the minority can gain public impact if the internal organization of the group becomes more cohesive. The effect of this structural transition towards a stronger minority's organisation is shown in Figures 1 and 2. Figure 1 shows the phase portraits of Equation (8) for three different values of $\gamma = 1/2, 2, 4$ that result from an increasing connectivity in the

minority group (q_{11}). It also shows the respective isoclines of the dynamical system and the stable fixed points at their intersection. As the probability q_{11} of in-group connections increases, the system undergoes a series of saddle-node bifurcations. First, a small increase of q_{11} (and hence γ) gives rise to an additional stable fixed point in which only the minority is expressive (not shown in Figure 1, yellow regime of competition in Figure 2). The minority and majority compete for public voice. As the internal connectivity of the minority group increases to $q_{11} = 4q$, the situation becomes symmetric with $\gamma = \delta = 2$. In other words, the minority can compensate its quantitative inferiority by a more cohesive internal organisation. Both groups can readily express their opinion if the other group is silent (competition, Figure 1, centre). However, an additional stable fixed point in which both opinions coexist also appears through another saddle-node bifurcation (coexistence). Finally, if the internal cohesion of the minority group becomes very large ($q_{11} = 7q$), the fixed point associated to a loud majority and silent minority disappears. That is, the minority always voices their view in public while the majority may become silent (see Figure 1, r.h.s.).

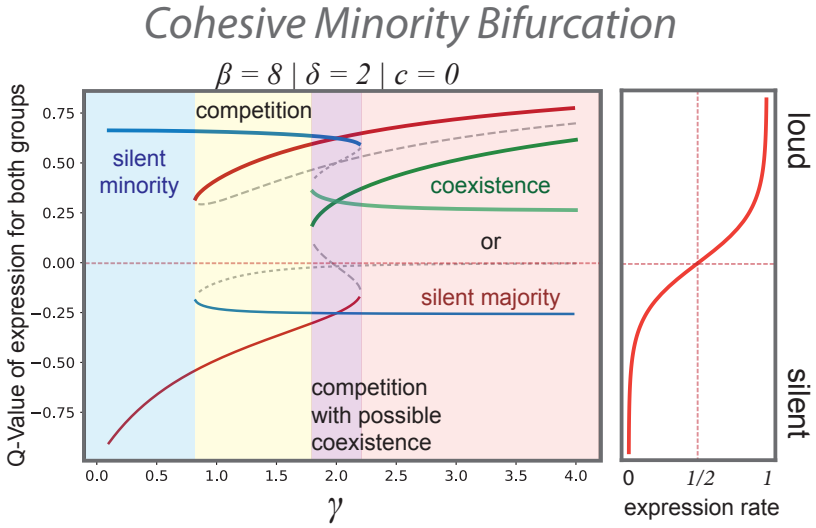

Figure 2. Bifurcation plot of the scenario in which a minority ($N_1/N_2 = 1/2$) gains public voice through stronger internal organisation (see also Figure 1). On the right-hand side, the expression rate (3) is shown as a function of Q for $\beta = 8$. The strength δ of the second group is kept constant ($\delta = 2$) and the costs of expression are zero ($c = 0$). As the internal cohesion of the minority group increases, for instance, due to strategic linking or tying group symbols, the system undergoes a series of saddle-node bifurcations. The minority's expression becomes more and more likely. Solid lines show the Q-values at the stable fixed points. For equilibria in which only one group is loud, the blue lines represent the majority group Q_2, the red lines the minority Q_1. The green lines correspond to the coexistence equilibrium in which both groups are expressive. While an unorganised minority is forced into silence (blue regime), a slight increase in group cohesion makes the minority's expression a stable outcome if the majority is silent (competition, yellow). At a certain point ($q_{11} = 4q$ and $\gamma = \delta = 2$), the minority's organisation can compensate numerical inferiority and a stable coexistence of two expressive groups is possible. By further increasing the minority group's strength, it is always visible in public while the majority may enter a spiral of silence.

3.2. "Spirals of Silence" as a Particular Regime of a More General Process

In the model, agents observe and react to their social environment in a way that is strongly reminiscent of Noelle-Neumann's theory of the spiral of silence [15,16,18]. In repeated interaction within their local neighbourhood, agents form a "quasi-statistical" impression of the current opinion climate in terms of an internalized expectation (Q-values) of which opinion is prevalent in their public spheres and whether their opinion can be articulated without being sanctioned. If their opinion corresponds to the perceived majority

view, they become more willing to speak out. If they perceive themselves to hold the minority view, they become less willing to do so. If all agents adapt to the current opinion landscape in this way, minorities are forced into a spiralling process in which silence leads to more silence. However, this happens only if the minority is perceived as minority in both groups. A bifurcation analysis of our model shows (see Figure 2) that majorities can also be forced into silence if a minority acts as a cohesive whole. Even a slight increase of homophily with respect to minority interaction can lead to a situation where a loud minority dominates public discourse because the majority is silent. Individuals with the actual majority opinion learn that voicing their view in public is rarely answered by support and is more often challenged by an expressive minority. The silence of the majority group is then collectively reinforced because each individual member is worse off by expressing their opinion.

The spiral of silence theory emerged as an attempt to explain a series of "last-minute swings" during German elections in the sixties and seventies [16]. (Termed "bandwagon effect", this phenomenon had already been observed by Lazarsfeld and colleagues in the 1940 US presidential elections [56]). While surveyed voting intentions where head-to-head between the two major parties until the very last days of the campaigning period, the evolution of expectations about who would win the election showed a clear trend towards the final winner during the month before the election day. Developing a series of refined survey instruments, Noelle-Neumann showed that differences in the willingness to publicly support a party were one source of these trends. Our model captures this dynamical feedback between the internalized expectations of the majority and the willingness to actively speak out for one's party and suggests that the situation of election campaigns at that time is characterised by the competitive regime in Figures 2 and 3.

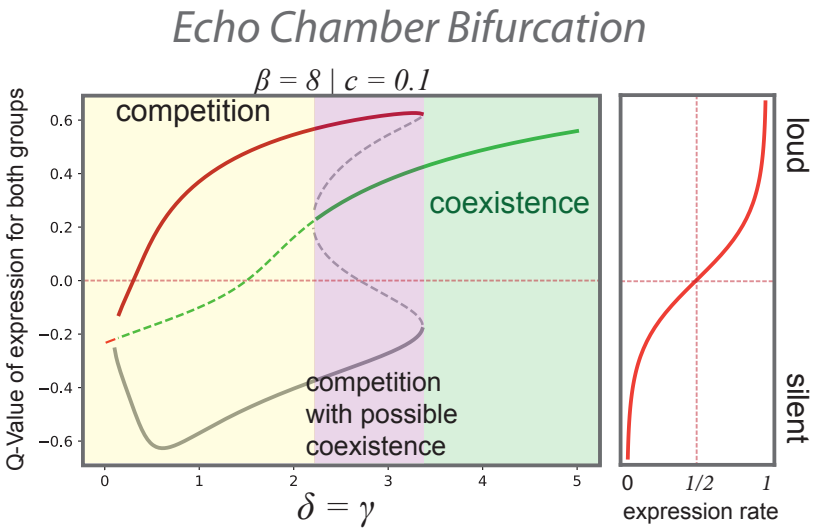

Figure 3. Bifurcation plot of the scenario in which two groups of equal size become more structured around the opinion they support. An increase of homophily for both opinion groups is captured by increasing γ and δ at the same time. The situation is symmetric and only Q_1 is shown. After a phase of competition, if homophily is low (yellow), coexistence emerges as a fixed point (violet) and becomes the only solution after a further slight increase of homophily (green). Both groups express their opinion within their own niches. (Results for $\beta = 8$ and $c = 0.1$).

Our research shows that the assumptions underlying the spiral of silence theory are well-grounded in neuroscientific research on the processing of social feedback. However, our model shows that the collective process described by Noelle-Neumann—that is, the spiral of silence—is only one possible outcome of the individual-level assumptions on which the theory builds. The bifurcation analysis of the dynamical system (8) reveals

that the structural transformations of group interaction may lead to qualitatively different regimes of collective opinion expression, including echo chambers.

3.3. Social Feedback and Echo Chambers

Today, social media are rapidly transforming the landscape of public opinion expression providing niches for virtually every opinion. Social network services have flexibilized options to connect with like-minded others no matter the topic or political stance—may that be on social media under hashtags such as #MeToo, on Telegram channels [57] or on imageboards with extremist content [58]. These fragmented online spheres hence provide previously unseen opportunities to escape the "fear of isolation" and to learn that there are others who share a similar view.

In the model, group interaction that is more and more structured around shared opinions is captured by a simultaneous increase of γ and δ meaning that social interaction with like-minded agents becomes more probable for both opinion groups. The qualitative effects of this structural transition towards more assortative networks is shown in Figure 3. As in-group ties become more prevalent in both groups, the system undergoes two saddle-node bifurcations from a competitive regime where only one group is aloud to a regime where coexistence is the only stable outcome. Private or semipublic rooms for expressing opinion online act as "echo chambers" and enable opinions previously marginalized or placed under taboo to resist the spiral of silence and become salient in the more general public discourse.

Our model entails that the perception of public support for an opinion is biased not only by the local connectivity of individuals [59], but also by the willingness of the supporters of different opinions to engage in the media. If opinions are shaped in social circles sorted around opinion, opposing opinion groups find their own views backed with social support and, in turn, become convinced of their primacy. The effect of ideological asymmetries [60,61] in opinion expression might overcome local homophily biases in the social network structure, since neutral observers and lurkers also get confronted with this distorted impression of public opinion. Democratic societies currently struggle with this transformed multifaceted climate of opinions because the foundational idea of government built on the common ground of public opinion [62,63] is fundamentally challenged.

4. Discussion

The SFT aims to contribute to a better understanding of societal-level implications of human social nature in modern information society. It provides a link between recent research on the neurological basis of social behaviour and the sociological theory of public opinion formation and expression. The model presented in this paper involved abstractions and assumptions at three different analytical levels (see Figure 4) each being subject to intensive research from different disciplinary angles. Providing a coherent theoretical account that integrates sociological modes of structural explanation [3,5], adaptive decision theory [28,44] and its underlying neurological mechanisms [64,65], the SFT offers a unique framework for guiding future interdisciplinary research on how social and cognitive mechanisms involved in platform-mediated communication on opinions play out at the scale of larger collectives.

At the collective level (top row), the SFT relates structural transformations in how we interact with one another to different regimes of collective opinion expression. The main modelling assumption made at this level is to map complex networks of social interaction to the relations within and across groups. Network science has brought about a portfolio of graph models to more realistically capture social interaction patterns [1,49,59,66], to which the model but not necessarily its formulation as a 2D system of differential equations can be applied. On the other hand, our theory suggests that empirical networks inferred on the basis of digital trace data [67] may be inherently biased by the activity of users who learnt that interaction on the media is rewarding. In fact, our model suggests that retrieved interaction patterns such as retweet networks [68–70] may render a situation

more polarized than it actually is, because public expression is less rewarding for actors who maintain relations across different opinion camps. Research on Twitter has also shown that retweets and replies give rise to very different global patterns of group interaction [61] suggesting that they serve rather different communicative functions. By bridging from individual decisions to express opinions to emergent collective activity patterns, the SFT provides a useful theoretical framework to analyse how those different communicative functions play out at large.

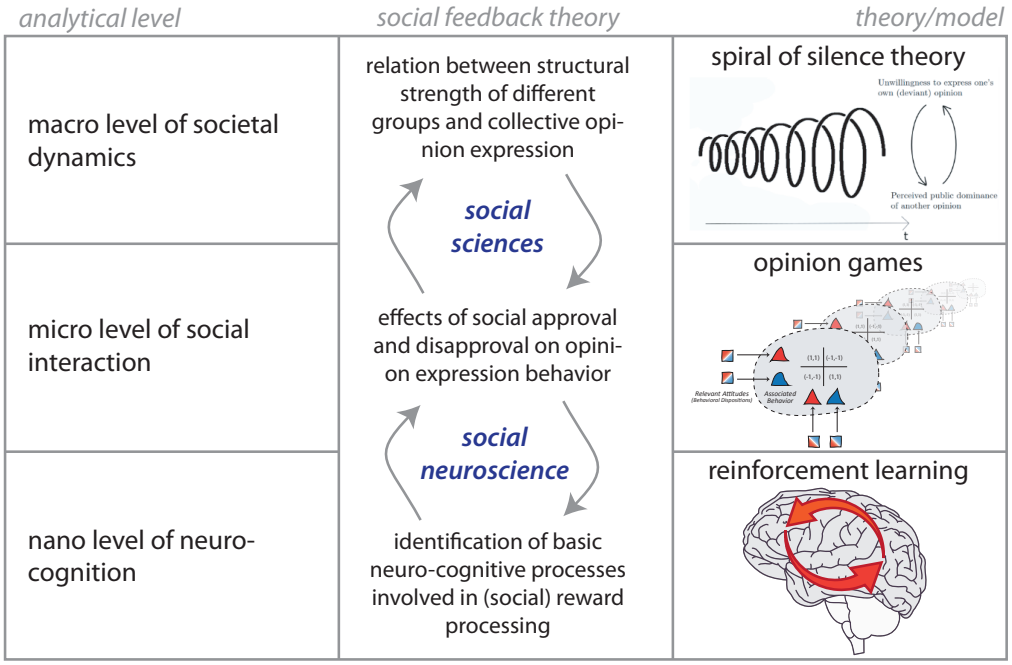

Figure 4. Schematic summary of social feedback theory of opinion expression. The theory involves three analytical levels from the level of neurocognitive processes to the level of social interaction, to the macro level of collective dynamics. The SFT bridges these levels through the notion of opinion games: first, by assuming that agent behaviour and the associated expected rewards adapt according to a reinforcement learning scheme that accurately models the reward processing system in the brain; and second, by bridging from individual decisions to express opinions to emergent patterns of collective opinion expression.

In the model, the micro level of social interaction (second row) is conceived as repeated opinion expression games in which agents respond to one another with signals of approval or disapproval. This entails simplifications such as dyadic interaction and a reward system that is homogeneous across individuals and groups. However, by drawing on games for modelling individual interaction, the SFT is well-equipped to take into account individual differences in reward perception as well as characteristics of the incentive structure of different social media platforms. In contrast to most previous models of social learning and opinion dynamics, the SFT takes into account that users have to express their opinions within the technical constraints of a platform in question. Conceiving social interaction as communication games that account for the incentives to engage online shifts the explanatory focus from forms of social influence to the rewards and incentives of opinion expression in different online settings. Of note, opinion games are also flexible enough to include cognitive costs associated to, for instance, preference falsification [7,71] and other sources of cognitive dissonance [72].

The social feedback framework draws on a neurocognitive foundation of TDRL (bottom row). In order to demonstrate that biologically rooted mechanisms of reward and value processing capture collective processes described in, for instance, the spiral of silence theory, we relied on the most simple TDRL scheme in the model. Social neuroscience is

quickly advancing towards a better understanding of how brain areas related to cognitive control interfere with this basic reward circuit. Recent work has revealed, for instance, that neural responses to social feedback are influenced by the social relation with the interaction partner [73] and that the reward valuation circuit is highly involved in shaping these relations [74]. Experimental designs that mimic interaction on social media [29,30] could clarify the role of different incentive systems for online engagement on opinion. This would contribute to a more systematic understanding of the types of games that are played in social media environments.

5. Conclusions

Massive social interaction in modern information society favours fast and largely unconscious modes of information processing. The model developed in this paper showed that the basal brain processes governing our reactions to social approval and disapproval could have a tremendous impact on collective processes of opinion expression. Simple feedback mechanisms may be at the root of phenomena such as silent majorities and enable well-organized minority groups to gain public voice or even dominate a discourse. Social media that facilitate massive and strategic social organisation around opinions can fundamentally alter the perception of public opinion in a society.

Social feedback theory has been proposed as a modelling framework that more explicitly takes into account the decision processes involved in expressing opinions online [17]. It extends previous work in opinion dynamics by allowing agents to refrain from participating in opinion exchange processes under certain circumstances and in certain environments. This has practical implications for computational social science methods aiming to measure opinions on online data where only users that actively engage in opinion exchanges become visible. However, for the sake of mathematical tractability, the current model was limited in terms of social interaction networks and did not account for different means of communication that online platforms may provide. Future models have to more realistically map the distinctive features and affordances of real social media platforms to become a practical tool for exploring digital communication devices that better serve deliberative modes of online opinion exchange.

Author Contributions: The manuscript was mainly written by S.B. All authors contributed to model development and the formal analyses. All authors have read and agreed to the published version of the manuscript.

Funding: This research received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 732942 (www.Odycceus.eu (accessed on 20 March 2022)).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 732942 (www.Odycceus.eu (accessed on 20 March 2022)). The paper has benefited from many fruitful discussions within the Odycceus group seminar in Leipzig. We especially acknowledge fruitful interactions with Roger Berger, Marcel Sarkoezi and Armin Pournaki. We would like to thank Wolfram Barfuss, Marc Keuschnigg and Stefan Westermann for their valuable feedback on earlier versions of this paper. Thanks also to Thomas Endler for their sketch of the human brain in Figure 4. We acknowledge support by the KIT-Publication Fund of the Karlsruhe Institute of Technology.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Newman, M.E.; Watts, D.J.; Strogatz, S.H. Random graph models of social networks. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 2566–2572. [CrossRef] [PubMed]
2. Newman, M.E.; Barabási, A.L.E.; Watts, D.J. *The Structure and Dynamics of Networks*; Princeton University Press: Princeton, NJ, USA, 2006.

3. Borgatti, S.P.; Mehra, A.; Brass, D.J.; Labianca, G. Network analysis in the social sciences. *Science* **2009**, *323*, 892–895. [CrossRef] [PubMed]
4. Bechtel, W.; Abrahamsen, A. Explanation: A mechanist alternative. *Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci.* **2005**, *36*, 421–441. [CrossRef] [PubMed]
5. Hedström, P.; Ylikoski, P. Causal mechanisms in the social sciences. *Annu. Rev. Sociol.* **2010**, *36*, 49–67. [CrossRef]
6. Granovetter, M.; Soong, R. Threshold models of diversity: Chinese restaurants, residential segregation, and the spiral of silence. *Sociol. Methodol.* **1988**, *18*, 69–104. [CrossRef]
7. Kuran, T. Sparks and prairie fires: A theory of unanticipated political revolution. *Public Choice* **1989**, *61*, 41–74. [CrossRef]
8. Lohmann, S. Collective Action Cascades: An Informational Rationale for the Power in Numbers. *J. Econ. Surv.* **2000**, *14*, 655–684. doi: 10.1111/1467-6419.00128. [CrossRef]
9. Centola, D. An experimental study of homophily in the adoption of health behavior. *Science* **2011**, *334*, 1269–1272. [CrossRef]
10. Bond, R.M.; Fariss, C.J.; Jones, J.J.; Kramer, A.D.; Marlow, C.; Settle, J.E.; Fowler, J.H. A 61-million-person experiment in social influence and political mobilization. *Nature* **2012**, *489*, 295–298. [CrossRef]
11. Christakis, N.A.; Fowler, J.H. Social contagion theory: Examining dynamic social networks and human behavior. *Stat. Med.* **2013**, *32*, 556–577. [CrossRef]
12. Friedkin, N.E.; Proskurnikov, A.V.; Tempo, R.; Parsegov, S.E. Network science on belief system dynamics under logic constraints. *Science* **2016**, *354*, 321–326. [CrossRef] [PubMed]
13. Centola, D.; Becker, J.; Brackbill, D.; Baronchelli, A. Experimental evidence for tipping points in social convention. *Science* **2018**, *360*, 1116–1119. [CrossRef] [PubMed]
14. Locke, J. *An Essay Concerning Human Understanding: And a Treatise on the Conduct of the Understanding. Complete in One Volume with the Author's Last Additions and Corrections*; Hayes & Zell: Philadelphia, PA, USA, 1860.
15. Noelle-Neumann, E. The spiral of silence a theory of public opinion. *J. Commun.* **1974**, *24*, 43–51. [CrossRef]
16. Noelle-Neumann, E. *Öffentliche Meinung: Die Entdeckung der Schweigespirale*; Ullstein: Berlin, Germany, 1996.
17. Banisch, S.; Olbrich, E. Opinion polarization by learning from social feedback. *J. Math. Sociol.* **2019**, *43*, 76–103. [CrossRef]
18. Noelle-Neumann, E.; Petersen, T. The spiral of silence and the social nature of man. In *Handbook of Political Communication Research*; Routledge: London, UK, 2004; pp. 357–374.
19. Fareri, D.S.; Delgado, M.R. Social rewards and social networks in the human brain. *Neuroscientist* **2014**, *20*, 387–402. [CrossRef]
20. Ruff, C.C.; Fehr, E. The neurobiology of rewards and values in social decision making. *Nat. Rev. Neurosci.* **2014**, *15*, 549–562. [CrossRef]
21. Orr, M.G.; Lebiere, C.; Stocco, A.; Pirolli, P.; Pires, B.; Kennedy, W.G. Multi-scale resolution of neural, cognitive and social systems. *Comput. Math. Organ. Theory* **2019**, *25*, 4–23. [CrossRef]
22. Schultz, W.; Dayan, P.; Montague, P.R. A neural substrate of prediction and reward. *Science* **1997**, *275*, 1593–1599. [CrossRef]
23. Haber, S.N.; Knutson, B. The reward circuit: Linking primate anatomy and human imaging. *Neuropsychopharmacology* **2010**, *35*, 4. [CrossRef]
24. Behrens, T.E.; Hunt, L.T.; Rushworth, M.F. The computation of social behavior. *Science* **2009**, *324*, 1160–1164. [CrossRef]
25. Maia, T.V. Reinforcement learning, conditioning, and the brain: Successes and challenges. *Cogn. Affect. Behav. Neurosci.* **2009**, *9*, 343–364. [CrossRef] [PubMed]
26. O'Doherty, J.P.; Dayan, P.; Friston, K.; Critchley, H.; Dolan, R.J. Temporal difference models and reward-related learning in the human brain. *Neuron* **2003**, *38*, 329–337. [CrossRef]
27. Averbeck, B.B.; Costa, V.D. Motivational neural circuits underlying reinforcement learning. *Nat. Neurosci.* **2017**, *20*, 505. [CrossRef] [PubMed]
28. Simon, H.A. Rationality as process and as product of thought. *Am. Econ. Rev.* **1978**, *68*, 1–16.
29. Sherman, L.E.; Payton, A.A.; Hernandez, L.M.; Greenfield, P.M.; Dapretto, M. The power of the like in adolescence: Effects of peer influence on neural and behavioral responses to social media. *Psychol. Sci.* **2016**, *27*, 1027–1035. [CrossRef]
30. Sherman, L.E.; Hernandez, L.M.; Greenfield, P.M.; Dapretto, M. What the brain 'Likes': Neural correlates of providing feedback on social media. *Soc. Cogn. Affect. Neurosci.* **2018**, *13*, 699–707. [CrossRef]
31. Dunbar, R.I.; Shultz, S. Evolution in the social brain. *Science* **2007**, *317*, 1344–1347. [CrossRef]
32. Dunbar, R.I. The social brain hypothesis. *Evol. Anthropol. Issues News Rev. Issues News Rev.* **1998**, *6*, 178–190. [CrossRef]
33. Andersson, C.; Törnberg, P. Toward a Macroevolutionary Theory of Human Evolution: The Social Protocell. *Biol. Theory* **2019**, *14*, 86–102. [CrossRef]
34. Poldrack, R.A. Can cognitive processes be inferred from neuroimaging data? *Trends Cogn. Sci.* **2006**, *10*, 59–63. [CrossRef]
35. Fareri, D.S.; Chang, L.J.; Delgado, M.R. Effects of direct social experience on trust decisions and neural reward circuitry. *Front. Neurosci.* **2012**, *6*, 148. [CrossRef] [PubMed]
36. Izuma, K.; Saito, D.N.; Sadato, N. Processing of social and monetary rewards in the human striatum. *Neuron* **2008**, *58*, 284–294. [CrossRef] [PubMed]
37. Klucharev, V.; Hytönen, K.; Rijpkema, M.; Smidts, A.; Fernández, G. Reinforcement learning signal predicts social conformity. *Neuron* **2009**, *61*, 140–151. [CrossRef] [PubMed]
38. Campbell-Meiklejohn, D.K.; Bach, D.R.; Roepstorff, A.; Dolan, R.J.; Frith, C.D. How the opinion of others affects our valuation of objects. *Curr. Biol.* **2010**, *20*, 1165–1170. [CrossRef] [PubMed]

39. Izuma, K.; Saito, D.N.; Sadato, N. Processing of the incentive for social approval in the ventral striatum during charitable donation. *J. Cogn. Neurosci.* **2010**, *22*, 621–631. [CrossRef] [PubMed]
40. Rilling, J.K.; Sanfey, A.G. The neuroscience of social decision-making. *Annu. Rev. Psychol.* **2011**, *62*, 23–48. [CrossRef] [PubMed]
41. Cushman, F. Action, outcome, and value: A dual-system framework for morality. *Personal. Soc. Psychol. Rev.* **2013**, *17*, 273–292. [CrossRef]
42. Amodio, D.M.; Frith, C.D. Meeting of minds: The medial frontal cortex and social cognition. *Nat. Rev. Neurosci.* **2006**, *7*, 268. [CrossRef]
43. Izuma, K. The social neuroscience of reputation. *Neurosci. Res.* **2012**, *72*, 283–288. [CrossRef]
44. Sutton, R.S.; Barto, A.G.; Williams, R.J. Reinforcement learning is direct adaptive optimal control. *IEEE Control Syst. Mag.* **1992**, *12*, 19–22.
45. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.
46. Hollerman, J.R.; Schultz, W. Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat. Neurosci.* **1998**, *1*, 304. [CrossRef] [PubMed]
47. O'Doherty, J.P. Reward representations and reward-related learning in the human brain: Insights from neuroimaging. *Curr. Opin. Neurobiol.* **2004**, *14*, 769–776. [CrossRef] [PubMed]
48. Gaisbauer, F.; Olbrich, E.; Banisch, S. The dynamics of opinion expression. *arXiv* **2019**, arXiv:1912.12631.
49. Wasserman, S.; Faust, K. *Social Network Analysis: Methods and Applications*; Cambridge University Press: Cambridge, UK, 1994; Volume 8.
50. Morris, S. Contagion. *Rev. Econ. Stud.* **2000**, *67*, 57–78. [CrossRef]
51. Tuyls, K.; Verbeeck, K.; Lenaerts, T. A selection-mutation model for q-learning in multi-agent systems. In Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems, Melbourne, Australia, 14–18 July 2003; pp. 693–700.
52. Sato, Y.; Crutchfield, J.P. Coupled replicator equations for the dynamics of learning in multiagent systems. *Phys. Rev. E* **2003**, *67*, 015206. [CrossRef]
53. Kianercy, A.; Galstyan, A. Dynamics of Boltzmann Q learning in two-player two-action games. *Phys. Rev. E* **2012**, *85*, 041145. [CrossRef]
54. Erdős, P.; Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **1960**, *5*, 17–60.
55. Gilbert, E.N. Random graphs. *Ann. Math. Stat.* **1959**, *30*, 1141–1144. [CrossRef]
56. Lazarsfeld, P.F.; Berelson, B.; Gaudet, H. *The People's Choice. How the Voter Makes Up Their Mind in a Presidential Campaign*; Columbia University Press: New York City, NY, USA, 1944.
57. Walther, S.; McCoy, A. US extremism on Telegram. *Perspect. Terror.* **2021**, *15*, 100–124.
58. Tuters, M.; Hagen, S. (((They))) rule: Memetic antagonism and nebulous othering on 4chan. *New Media Soc.* **2020**, *22*, 2218–2237. [CrossRef]
59. Lee, E.; Karimi, F.; Wagner, C.; Jo, H.H.; Strohmaier, M.; Galesic, M. Homophily and minority-group size explain perception biases in social networks. *Nat. Hum. Behav.* **2019**, *3*, 1078–1087. [CrossRef] [PubMed]
60. Jost, J.T. Ideological asymmetries and the essence of political psychology. *Political Psychol.* **2017**, *38*, 167–208. [CrossRef]
61. Gaisbauer, F.; Pournaki, A.; Banisch, S.; Olbrich, E. How Twitter affects the perception of public opinion: Two case studies. *arXiv* **2020**, arXiv:2009.01666.
62. Hume, D. *Essays: Moral, Political and Literary*; Oxford University Press: Oxford, UK, 1963; pp. 1741–1742.
63. The Gettysburg Address Delivered by Abraham Lincoln Nov. 19 1863 at the Dedication Services on the Battle Field. Available online: https://www.loc.gov/item/2004671506/ (accessed on 20 March 2022).
64. Dayan, P.; Daw, N.D. Decision theory, reinforcement learning, and the brain. *Cogn. Affect. Behav. Neurosci.* **2008**, *8*, 429–453. [CrossRef] [PubMed]
65. Glimcher, P.W.; Fehr, E. *Neuroeconomics: Decision Making and the Brain*; Academic Press: Cambridge, MA, USA, 2013.
66. Scott, J. Social network analysis. *Sociology* **1988**, *22*, 109–127. [CrossRef]
67. Lazer, D.; Pentland, A.; Adamic, L.; Aral, S.; Barabási, A.L.; Brewer, D.; Christakis, N.; Contractor, N.; Fowler, J.; Gutmann, M.; et al. Computational social science. *Science* **2009**, *323*, 721–723. [CrossRef] [PubMed]
68. Conover, M.D.; Ratkiewicz, J.; Francisco, M.; Gonçalves, B.; Menczer, F.; Flammini, A. Political polarization on twitter. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, 17–21 July 2011.
69. Garimella, K.; Morales, G.D.F.; Gionis, A.; Mathioudakis, M. Quantifying controversy on social media. *ACM Trans. Soc. Comput.* **2018**, *1*, 3. [CrossRef]
70. Gaumont, N.; Panahi, M.; Chavalarias, D. Reconstruction of the socio-semantic dynamics of political activist Twitter networks—Method and application to the 2017 French presidential election. *PLoS ONE* **2018**, *13*, e0201879. [CrossRef]
71. Kuran, T. Now out of never: The element of surprise in the East European revolution of 1989. *World Politics* **1991**, *44*, 7–48. [CrossRef]
72. Festinger, L. *A Theory of Cognitive Dissonance*; Stanford University Press: Redwood City, CA, USA, 1957; Volume 2.
73. Hughes, B.L.; Leong, J.K.; Shiv, B.; Zaki, J. Wanting to like: Motivation influences behavioral and neural responses to social feedback. *bioRxiv* **2018**, 300657. [CrossRef]
74. Zerubavel, N.; Hoffman, M.A.; Reich, A.; Ochsner, K.N.; Bearman, P. Neural precursors of future liking and affective reciprocity. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 4375–4380. [CrossRef] [PubMed]