Generating consensus and dissent on massive discussion platforms with a semantic-vector model

Alfredo Ferrer,[1,2] David Muñoz-Jordán,[1,*] Alejandro Rivero,[1,2] Alfonso Tarancón,[3] Carlos Tarancón,[2] and David Yllanes[1,3,4,5]

[1]*Instituto de Biocomputación y Física de Sistemas Complejos (BIFI), Universidad de Zaragoza, 50018 Zaragoza, Spain*
[2]*Kampal Data Solutions, WTCZ, Avda. Maria Zambrano 31, 50018 Zaragoza, Spain*
[3]*Departamento de Física Teórica, Universidad de Zaragoza, 50009 Zaragoza, Spain*
[4]*Zaragoza Scientific Center for Advanced Modeling (ZCAM), 50018 Zaragoza, Spain*
[5]*Fundación ARAID, Diputación General de Aragón, 50018 Zaragoza, Spain*

(Received 21 January 2026; revised 2 June 2026; accepted 29 June 2026; published 21 July 2026)

Reaching consensus on massive discussion networks is critical for reducing noise and achieving optimal collective outcomes. However, the natural tendency of humans to preserve their initial ideas constrains the emergence of global solutions. To address this, collective intelligence (CI) platforms facilitate the discovery of globally superior solutions. We introduce a dynamical system based on the standard $O(N)$ model to drive the aggregation of semantically similar ideas. The system consists of users represented as nodes in a $d = 2$ lattice with nearest-neighbor interactions, where their ideas are represented by semantic vectors computed with a pretrained embedding model. We analyze the system's equilibrium states as a function of the coupling parameter β. Our results show that $\beta > 0$ drives the system toward a ferromagneticlike phase (global consensus), while $\beta < 0$ induces an antiferromagneticlike state (maximum dissent), where users maximize semantic distance from their neighbors. This framework offers a controllable method for managing the tradeoff between cohesion and diversity in CI platforms.

DOI: 10.1103/kmg5-zhmr

I. INTRODUCTION

The dynamics of opinion formation in large-scale multiagent systems is a central theme in sociophysics. In social networks and collaborative environments, ideas compete for dominance, exhibiting evolutionary patterns analogous to physical processes such as percolation [1–3]. When bot or hybrid populations are considered, consensus in multiagent systems emerges as a particularly challenging problem in the field of artificial intelligence [4].

The dynamics governing consensus generation have been extensively characterized in statistical physics, where the study of active matter [5,6] has been particularly fruitful in models and experiments. Indeed, the study of flocks is the foundation of this field, beginning with the celebrated Vicsek model [7] and developing with studies of birds [8,9], bacterial colonies [10], or sheep [11], to give just a few examples. A vast literature has hence developed on the topics of mutual and self-alignment in active matter [12–16]. Beyond alignment interactions, the ideas of "quorum sensing" [17,18] and the motility-induced phase transition [19] have been important to understand collective phenomena in microbiological systems [20–22]. The active-matter approach has also found application to human systems, from the interactions of pedestrians [23] to the dynamics of mosh pits in heavy-metal concerts [24].

A usual characteristic of active and other nonequilibrium systems is the lack of reciprocity [25–28], which becomes particularly important when considering social groups. As a first toy—but seminal—example, the Kuramoto model of coupled oscillators can be adapted to represent a combination of "conformists" and "contrarians" [29]. More generally, heterogeneity in populations has been shown to have dramatic effects on collective motion. For instance, in collective migration "small proportions of individuals actively acquir[e] directional information from their environment, whereas the majorities use a socially facilitated movement behavior" [30]. The effect of such "opinion leaders" can be enhanced—and consensus promoted—by the existence of uninformed individuals [31], while on the other hand, a small number of active dissenters can disrupt alignment interactions and destroy flocks [32].

In social physics opinion dynamics have increasingly focused on nonequilibrium frameworks to explain the emergence of polarization and echo chambers [33,34]. These models typically characterize an agent's opinion as a unidimensional scalar value where the sign denotes a binary stance—positive or negative—on a specific controversial issue, and the magnitude represents the strength of that conviction. Within these systems, nonequilibrium phase transitions between consensus and polarization are driven by mechanisms such as reinforcement-based dynamics or

*Contact author: dmunoz@bifi.es

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

coevolving networks, where agents preferentially interact with those holding similar opinions.

In the contexts discussed thus far, "consensus" has traditionally been defined as the alignment of these scalar stances or the convergence of low-dimensional collective behaviors. Recent research has generalized these concepts to multidimensional topic spaces [35,36]. In these generalized frameworks, an agent's opinion is represented as a vector where each component corresponds to a distinct, though possibly overlapping theme or topic. These models investigate how feedback-driven polarization or "information cocoons" arise from the dynamic adaptation of user preferences across multiple themes.

Our approach, however, interprets consensus more literally as the adoption of complex, multifaceted ideas within a social group, rather than a position on a single scale. Unlike models centered on extreme polarization (e.g., pro-choice vs pro-life), our framework utilizes high-dimensional semantic vectors to represent open-ended responses to a single question. In this semantic space, lack of similarity between two responses does not inherently imply ideological opposition or bipolarization but simply a lack of semantical agreement.

Early experiments in small face-to-face groups demonstrated the emergence of collective intelligence (CI) [37], where group interactions lead to ideas superior to those of isolated individuals. The digital era has scaled these dynamics, enabling massive online platforms designed to foster constructive collaboration. Recent experiments involving up to a thousand participants have evinced the emergence of solutions superior to individual proposals. These benefits span from sensitive social issues, such as cyberbullying and gender equality, to deterministic academic problems in mathematics or physics [38,39].

In these experimental setups, users are typically arranged on a virtual two-dimensional square lattice with periodic boundary conditions. Each user is presented with a problem and asked to propose a solution in the form of a typed string of text. Periodically, participants are allowed to interact with their four nearest neighbors. Convergence toward a unique global solution relies on an imitation mechanism, where superior solutions ideally propagate through the lattice, increasing their domain size over time via successive copying events. At variance with the directions of motion in a flock, however, it is not immediately obvious how to quantify the similarity of two competing ideas or how to propose a dynamical law that would drive the system toward consensus (or dissent).

Humans and bots exhibit significant inertia. They frequently retain their initial opinions rather than adopting their neighbors' views, even when presented with marginally superior alternatives. To overcome this resistance and facilitate consensus, artificial dynamics are often introduced, such as forced copying events, topological rewiring or the extinction of low-frequency responses [40]. All these rules inherently require a selection bias favoring certain solutions over others. Defining an objective fitness function is problematic for social or philosophical questions, where multiple valid viewpoints coexist and no objective "ground truth" exists. Imposing an external value judgment compromises the system's independence as there are no "sacred" solutions.

In an effort to avoid such value judgments, Ref. [40] proposed using the number of times a response appears in the system—that is, its frequency—as a simple goodness criterion. Although this metric proved effective, it presents two major limitations. First, it depends on exact string matching, treating semantically equivalent but textually distinct responses as unrelated states. Second, it introduces nonlocal effects. A change in the state of a single node instantly alters the global frequency distribution, thereby influencing the transition probabilities of all nodes in the system simultaneously.

In this work, building on the tools of statistical mechanics and on recent advances in large language models (LLMs), we propose an approach to overcome these limitations. Language models based on neural networks have undergone a sequence of revolutions, first by increasing the hidden size of each layer and the number of layers, and then by expanding the number of simultaneous tokens that are attended to in each step, which resulted in the transformers architecture [41]. Generative models such as GPT specialize this architecture to predict the next token, a simple process that allows training hundreds of billions of parameters. BERT models [42] specialize in predicting intermediate deleted tokens, requiring more complex training, limited to hundreds of millions of parameters, but allowing each token to be aware of the meaning of the entire sentence. It is then possible to use a large BERT model to extract the semantic meaning by averaging the hidden sizes across all the tokens of the sentence. Current literature calls the vector produced in this way an *embedding*, but we prefer to use the classical denomination and say that each idea will be mapped to a (high-dimensional) *semantic vector*. The BERT model is fine tuned so that the similarity between two responses can be quantified by the scalar product of their vectors.

This strategy allows us to transition from a frequency-based global coupling to a semantics-based local interaction. From this local interaction we shall define a total energy for the system—complete consensus corresponding to total alignment of the semantic vectors and, hence, to the energy ground state. This energy function will allow us to define a (stochastic) dynamical law, where a fictitious temperature will regulate how strongly the system is driven to consensus—or to maximum dissent, for the case of a negative temperature. We shall demonstrate this model by conducting several numerical experiments where a system with initially N different ideas will be subjected to temperature cycles, taking inspiration from the simulated-annealing algorithm, and the approach to global solutions will be characterized.

The rest of the paper is organized as follows. In Sec. II we describe the principles of the CI experiment and the motivation for the proposed model. In Sec. III we introduce the model and a Markov process based on it that will define our dynamics. Numerical simulations are presented in Sec. IV, and our concluding remarks are summarized in Sec. V.

II. THE COLLECTIVE INTELLIGENCE PARADIGM AND SEMANTIC VECTORS

A CI experiment [40] consists of a setup where participants are presented with a context describing a specific situation and then asked a series of questions that they are required

to answer. The participants are arranged on a square lattice with periodic boundary conditions, allowing them to view the responses of their four nearest neighbors. Consequently, they can either copy a neighbor's response or formulate a new one.

To facilitate convergence, three dynamical mechanisms are implemented. The first is a copying mechanism, where a participant adopts a neighbor's response. The second is a permutation mechanism, in which a participant's position within the lattice is changed, exposing them to new neighbors and responses. The third is an extinction mechanism, where the least frequent responses are eliminated, leaving the response empty for those participants.

Experiments are organized into successive stages with evolving behaviors. In the first stage participants operate in isolation to generate original responses. Starting from the second stage, the lattice structure becomes active, enabling participants to view their neighbors' answers. Subsequent stages incorporate the dynamical mechanisms designed to enhance collective interaction.

Regarding the copying dynamics, the transition probabilities can be determined by response frequencies. Let $f(n,t)$ be the frequency of the response held by node n at time t. If a neighbor $n+\mu$ holds a response with frequency $f(n+\mu,t)$, the probability that user n overwrites their response with that of the neighbor is given by

$$P(n,n+\mu,t)=\frac{f(n+\mu,t)}{f(n+\mu,t)+f(n,t)}. \tag{1}$$

While this algorithm successfully generated consensus, it suffers from a fundamental physical limitation: nonlocality. A response change at a single node n instantaneously alters the frequency count $f(n,t)$, thereby modifying the probabilities of every other node holding that response, regardless of their distance in the lattice.

To implement a dynamical system based on ideas, we must first define a quantitative metric for "semantic similarity." A direct approach would involve querying a generative LLM to rate the similarity between two texts on a continuous scale [0,1]. While this method offers high semantic accuracy, it presents an insurmountable computational bottleneck. In a typical CI experiment involving 1000 users, 5–10 new ideas are generated per second. If every new idea requires a call to the application programming interface (API), the latency induced by thousands of API calls would render this solution unfeasible and seriously expensive.

A computationally efficient alternative is to map each textual response to a semantic vector $\boldsymbol{\phi}$ using a pretrained embedding model. These models typically generate high-dimensional vectors, with a number N of components from $N=256$ to 1024. Consequently, the semantic similarity between two responses i and j can be quantified by the scalar product between their semantic vectors $\boldsymbol{\phi}(i)\cdot\boldsymbol{\phi}(j)$. Unlike calling an LLM API, calculating the scalar product is negligible on modern GPUs and even the production of the vector can be done locally, as large BERT models really involve only around 0.5 billions of parameters, easy to fit in the GPU vRAM.

As we demonstrate in this work, it is possible to define a local interaction based on semantic similarity that can similarly drive the system toward consensus while strictly adhering to locality principles. Furthermore, we will discuss how the additional dynamics defined in previous studies naturally emerge as implicit features of this model.

III. THE MODEL

While it is possible to define a heuristic copying mechanism similar to the dynamics in Eq. (1) without an underlying statistical-mechanics framework, establishing a rigorous physical model offers significant advantages. It provides a solid foundation for understanding the system's behavior, analyzing its limiting cases, and systematically controlling parameters. Consequently, we propose a Hamiltonian-based model that not only satisfies the requirements of the previous section but also facilitates a simpler and more robust analysis of the experiments.

Let us consider a system where each participant n holds a single response i_n. Each response is mapped to a normalized semantic vector $\boldsymbol{\phi}(i_n)$ such that $\|\boldsymbol{\phi}\|=1$. Participants are placed on a square lattice of size $V=L\times L$ with periodic boundary conditions and nearest-neighbor interactions, so each node only *sees* its four adjacent neighbors $n+\mu$ and therefore interacts only with them.

We can define an energy for this system, which will be lowest when the alignment between semantic vectors is maximum (consensus) and increase when dissenting responses coexist. The associated model can be defined as

$$H=-\sum_{n,\mu}\boldsymbol{\phi}(i_n)\cdot\boldsymbol{\phi}(i_{n+\mu}). \tag{2}$$

This formulation is inspired by the standard $O(N)$ model in $d=2$ dimensions with nearest-neighbor interactions; see, e.g., [43,44]. Here, N represents the dimensionality of the semantic space, which typically ranges from 256 to 1024 in our case of use. Now, according to the Hohenberg-Mermin-Wagner theorem, continuous symmetries cannot be spontaneously broken with nearest-neighbors interactions in $d=2$ [45,46]. This would suggest that consensus cannot be reached in such a model at any finite temperature. The same situation arises in flocking systems [6], where the problem is skirted due to the nonequilibrium nature of the models. Similarly, our dynamical law will not lead to the equilibrium canonical distribution of the $O(N)$ model.[1]

The first difference is that our set of available vectors does not span the entire continuous $O(N)$ hypersphere. Instead, the available state space, which we denote as $R(N)$, is restricted to the subset of active responses currently held by at least one user. Furthermore, this set is dynamic and nonincreasing under our Markov process: if a response is held by a single user who then copies a neighbor, that unique vector is permanently erased from the system, causing $R(N)$ to shrink. In a

[1]Long-range interactions are the other way around the Hohenberg-Mermin-Wagner theorem [49]. This is the case, for instance, of thermalized elastic membranes, where a flat phase is possible because in-plane phonons couple with height fluctuations, causing the bending rigidity to be renormalized [50]. In this work, however, we are interested in local interactions only.

real experiment, $R(N)$ changes dynamically due to both this copying mechanism and the users' active modifications.

We propose the following Markov process:

(1) We choose a user n in the network. The choice can be sequential or random.

(2) For a given node n, the set of new allowed states is restricted to the current responses possessed by its four nearest neighbors $\{i_{n+\mu}\}$, plus its current state. If the node retains its response, the change in energy is $\Delta H = 0$. If the node copies the response $i_{n+\nu}$ from a specific neighbor ν, the change in energy is given by

$$\Delta H = -\left[\sum_{\mu} \boldsymbol{\phi}(i_{n+\nu}) \cdot \boldsymbol{\phi}(i_{n+\mu}) - \sum_{\mu} \boldsymbol{\phi}(i_n) \cdot \boldsymbol{\phi}(i_{n+\mu})\right]. \quad (3)$$

(3) Use ΔH to decide whether to accept the new state. Several choices for the transition matrix are possible, such as the standard Metropolis rule. We employ the heat-bath algorithm due to its efficiency in handling small discrete state spaces. The probability of choosing response $i_{n+\nu}$ is, then,

$$P(i_n \to i_{n+\nu}) = \frac{\exp\left(-\beta \Delta H(i_n \to i_{n+\nu})\right)}{\sum_{\mu} \exp\left(-\beta \Delta H(i_n \to i_{n+\mu})\right)}. \quad (4)$$

In this equation, the sum includes the case $\mu = 0$, corresponding to no change. Our dynamical law includes a variable parameter β, which, by analogy with equilibrium statistical mechanics, we shall refer to as an (inverse) temperature.

It is important to distinguish this process from the canonical dynamics of the $O(N)$ model, where a tentative change can be to any $O(N)$ vector. First, such a move would, in almost all cases, explore regions of the semantic-vector space that do not correspond to any sensible human response. Second, because vectors in $R(N)$ can permanently disappear under our dynamics, the system naturally tends toward absorbing states. Therefore, unlike an ergodic transition matrix in the standard $O(N)$ model, the dynamics proposed above are capable of breaking the continuous symmetry and, hence, achieving consensus. Hence, our β will not represent true thermal equilibrium but, as we shall explain later, it will still control energy fluctuations and allow us to define an "annealing" procedure to drive the system toward ordered states.

To study the evolution of this model, we define two metrics. The first is the *living-response* count l_{v}, defined as the number of different responses on the lattice; therefore $l_{\mathrm{v}} = |R(N)|$. Note that under our proposed dynamic, this metric can only decrease. If the last user holding response i_n copies a neighbor, i_n is permanently extinguished from the system. The second is the *semantic energy* e_{s} defined as

$$e_{\mathrm{s}} = \frac{H}{4V} = -\frac{1}{4V} \sum_{n,\mu} \boldsymbol{\phi}(n) \cdot \boldsymbol{\phi}(n+\mu). \quad (5)$$

Note that this is an intensive metric, $e_{\mathrm{s}} \in [-1, 1]$, where -1 corresponds to perfect global alignment. Furthermore, we can decompose e_{s} into the *local semantic energy* $e_{\mathrm{ls}}(n)$ of a node n as

$$e_{\mathrm{ls}}(n) = -\frac{1}{4} \sum_{\mu} \boldsymbol{\phi}(n) \cdot \boldsymbol{\phi}(n+\mu), \quad (6)$$

ensuring that $e_{\mathrm{s}} = \frac{1}{V} \sum_n e_{\mathrm{ls}}(n)$.

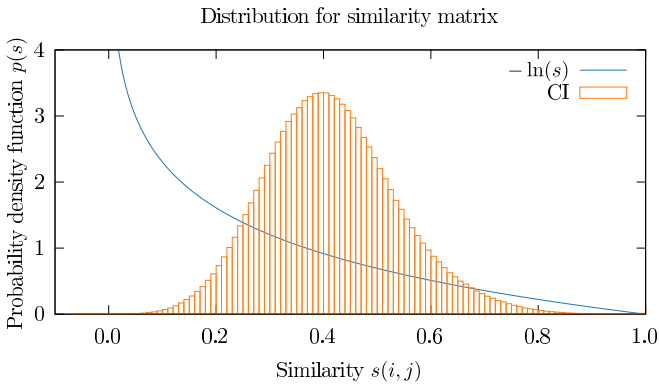


FIG. 1. Distributions used in the experiments. The graph of the function $-\ln(s)$ represents the probability density function of $P(X \cdot Y)$, where X and Y are independent random variables with a uniform distribution on [0,1]. The CI is the histogram of the distribution of the similarity matrix between the embeddings of the responses from a real CI experiment with more than 5000 different responses.

IV. SIMULATIONS

To study the model we perform Monte Carlo simulations on lattices with periodic boundary conditions. Let $V = L \times L$ be the size of the lattice, which we initialize with V different responses, one for each node. Subsequently, we construct the similarity matrix $s(i, j)$ where each element (i, j) represents the scalar product between responses i and j, $s(i, j) = \boldsymbol{\phi}(i) \cdot \boldsymbol{\phi}(j)$.

We have studied the behavior of the model for two different starting conditions. We first consider a synthetic dataset, where the similarity matrix $s(i, j)$ is generated from the expected distribution $P(X \cdot Y)$ of two independent random variables X and Y with a uniform distribution on [0,1]. The density probability function for a given similarity s is $p(s) = -\ln(s)$. We initialize the system such that every user starts with a unique response.

Alternatively, we use data from a real CI experiment where the users generated more than 5000 different responses (see Appendix A). In this case we computed the semantic vectors using the Arctic Embed 2.0 model, specifically, `snowflake-arctic-embed-l-v2.0` [47] with an output dimension of 256. The model weights can be obtained from the Hugging Face repository in [48]. For each simulation we initialize the lattice by randomly sampling a unique subset of V responses from the total pool, ensuring that every user starts with a distinct response.

In all simulations the lattice size is set to $V = 32 \times 32 = 1024$.

As can be observed in Fig. 1, the scalar products $s(i, j)$ for both the synthetic dataset and the real CI experiment take only positive values. For the real CI data, this strict positivity is a direct consequence of the embedding model's training. Because the similarity between any pair of interacting users is always greater than zero, the semantic energy e_s of our system is naturally restricted to negative values ($e_s < 0$). Thus, while the theoretical bounds defined in Eq. (5) allow for $e_s \in [-1, 1]$, the dynamics in our specific simulations will exclusively take place in the negative-energy domain.

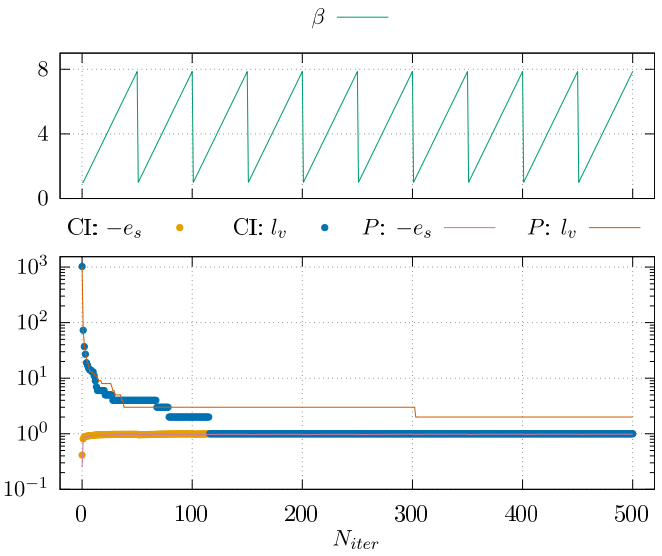


FIG. 2. Example of evolution for the Standard annealing process with $\beta \in [1, 8]$. Results for the CI distribution are represented with points, and results for the $P(X \cdot Y)$ distribution are represented with lines.

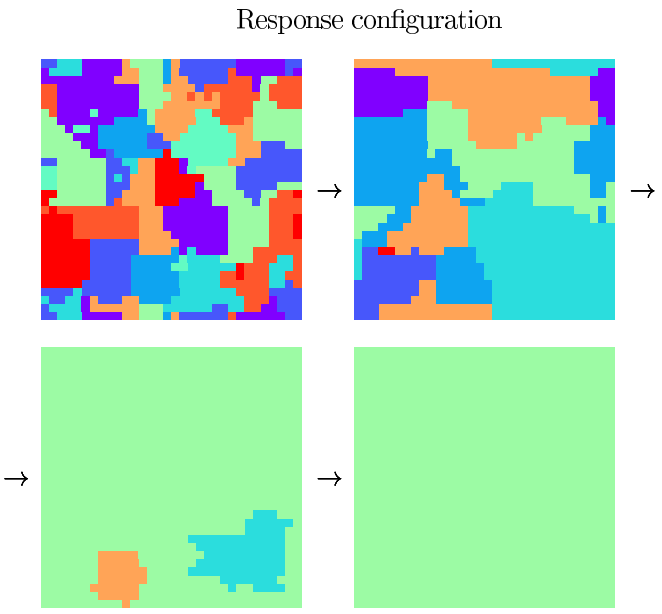


FIG. 3. Snapshots of the response configuration in the ferromagnetic regime. Nodes are colored according to the response (idea) they represent. Arrows indicate the direction of time evolution. The corresponding local energy distribution for each frame is shown in Fig. 4.

To study the evolution of the system, we will measure the semantic energy e_s, Eq. (5), and the number of living responses l_v.

Taking inspiration from the simulated annealing algorithm, we consider several possibilities for temperature-cycling protocols, aiming to study consensus and dissent phenomena and to accelerate the system's convergence. These are the three options:

(1) Standard: We start with β close to 0 and increase it during the Monte Carlo process to positive values. In our case, we start with 1 and end with 8.

(2) Negative Standard: We start with a negative β close to 0 and decrease during the Monte Carlo process to negative values. In our case, we start with -1 and end with -8.

(3) Alternating positive and negative β: We start with negative values of β and increase them to positive values of β. In our case we start with -8 and end with 8. In this way we have a consensus-dissent procedure, where users are surrounded alternatively with similar or differing ideas.

For each of these options, the annealing schedule is structured as a series of thermal cycles. The range of β is divided into 50 discrete steps to form a single cycle. This cycle is repeated 10 times, resulting in a total of $N_{\text{iter}} = 500$ steps for β following a sawtooth pattern. Furthermore, at each discrete step of β, the system is relaxed by performing ten full Monte Carlo sweeps. Measurements are taken only after this relaxation of the system.

Figure 2 shows a typical evolution with the Standard annealing process. This option leads the system to a configuration where all response vectors are aligned, which we call ferromagnetic, using the standard language for the $O(N)$ model. The system starts with $l_v = V$ and e_s in a maximum. As the system evolves, l_v decreases and the energy e_s tends to -1. Eventually, the system reaches equilibrium where only one response is left, $l_v = 1$ and $e_s = -1$, as clearly seen for the CI distribution. We can say that this is a state of consensus, because all nodes in the lattice are surrounded by neighbors with the same response, that is, people that share the same idea.

We illustrate the evolution of the system on the lattice in Fig. 3, which displays four snapshots of the response configuration. Initially, the system exhibits a disordered multidomain structure characterized by numerous small clusters. As time progresses, a coarsening process takes place where smaller clusters vanish while larger domains expand, eventually leading to a state where a single response dominates (consensus). In Fig. 4 we depict the corresponding evolution of the local energy. In the bulk of the domains, the local energy assumes its minimum value, $e_s = -1$. Higher energy values are strictly confined to the cluster boundaries, highlighting the energy cost associated with lack of local consensus.

Using the Negative Standard annealing process, we can see a different behavior that leads to an antiferromagnetic configuration. Again, the system starts with $l_v = V$, but Fig. 5 shows a different behavior for the energy e_s. As β decreases, the energy also decreases. Note that e_s never approaches -1, unlike in the previous case, even though l_v decreases. The equilibrium in that case is $l_v = 2$, where both responses are arranged in the lattice in a checkerboard pattern. This represents a state of dissent because all nodes in the lattice are surrounded by neighbors with a different response, that is, people that do not share the same idea.

The evolution of the system in the antiferromagnetic regime is illustrated in Fig. 6. Initially, the system displays a highly disordered configuration with a diverse set of responses. However, in contrast to the ferromagnetic phase, the evolution does not lead to the formation of monochromatic clusters. Instead, the responses arrange themselves into

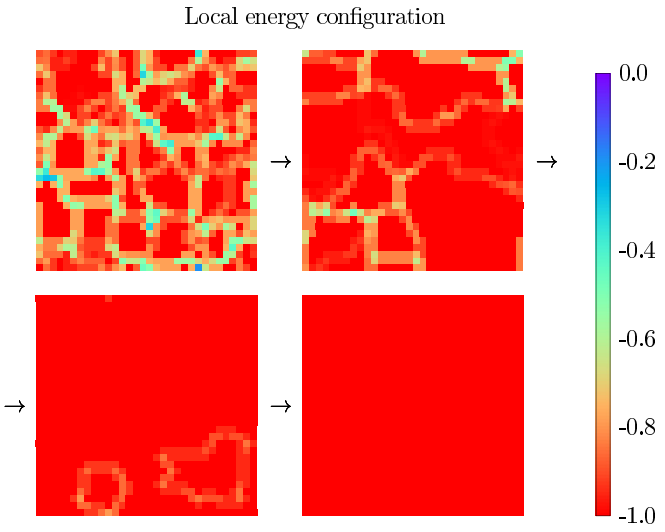


FIG. 4. Evolution of the local energy configuration in the ferromagnetic state. Nodes are colored according to their local energy, as defined in (6). Each panel corresponds to the respective time step shown in Fig. 3.

a checkerboardlike structure, where each node maximizes its semantic difference relative to its nearest neighbors.

This behavior is quantitatively captured by the local energy configuration shown in Fig. 7. As the system evolves, the local energy in the bulk of the domains stabilizes at a relatively high value close to zero. This high-energy state indicates that neighboring vectors are nearly orthogonal to each other. Interestingly, the energy only drops to lower values at the boundaries where the perfect checkerboard pattern is disrupted. This confirms that even though the checkerboard pattern comprises distinct responses, the system successfully

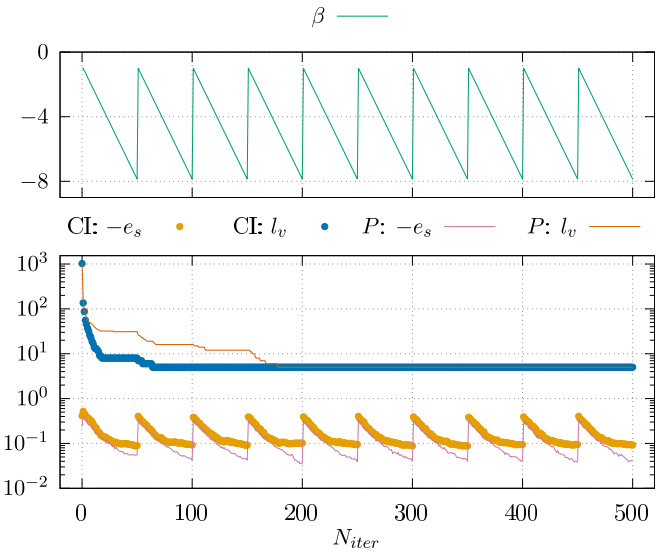


FIG. 5. Example of evolution for the Negative Standard annealing process with $\beta \in [-8, -1]$. Results for the CI distribution are represented with points, and results for the $P(X \cdot Y)$ distribution are represented with lines.

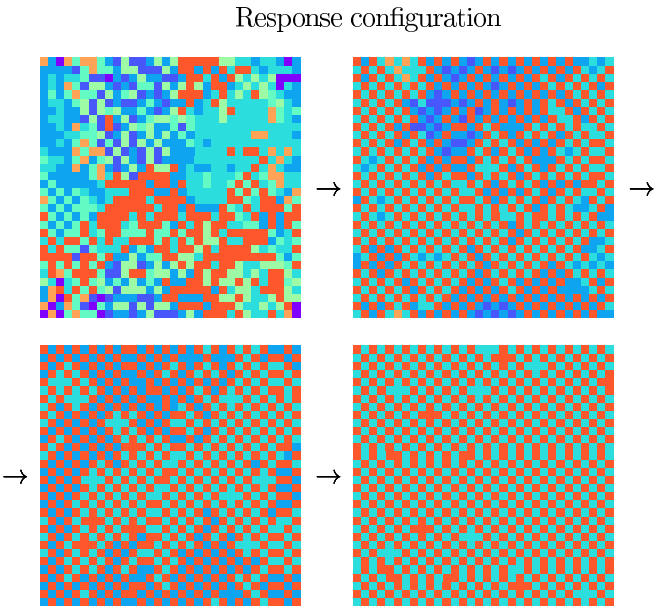


FIG. 6. Snapshots of the spatial configuration in the antiferromagnetic regime. Nodes are colored according to the response they represent. Unlike the ferromagnetic case, the system evolves into a checkerboardlike pattern. Arrows indicate the direction of time evolution. The corresponding local energy distribution is shown in Fig. 7.

maintains a state of global diversity by minimizing local semantic overlap (dissent).

When using the Consensus-Dissent strategy, we find that convergence is achieved sooner. Figure 8 shows a typical example of the evolution for this Consensus-Dissent annealing process. Here β starts with negative values, where the energy

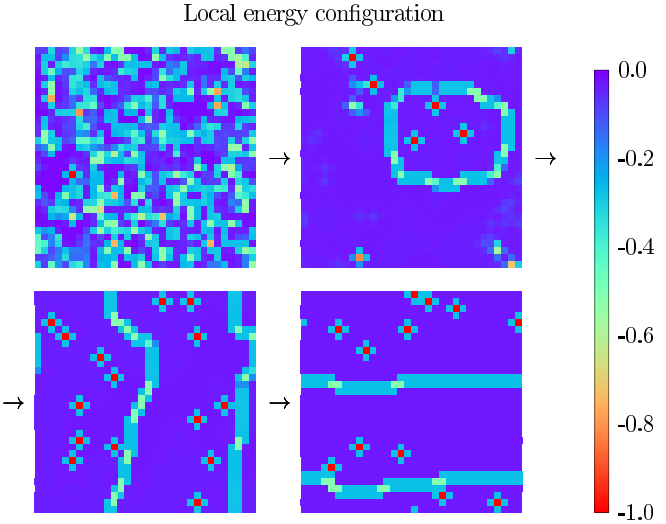


FIG. 7. Evolution of the local energy landscape in the antiferromagnetic state. Nodes are colored according to their local energy defined in (6). In this regime, the system stabilizes at higher energy values, with lower energy excitations confined to domain boundaries. Each panel corresponds to the frames in Fig. 6.

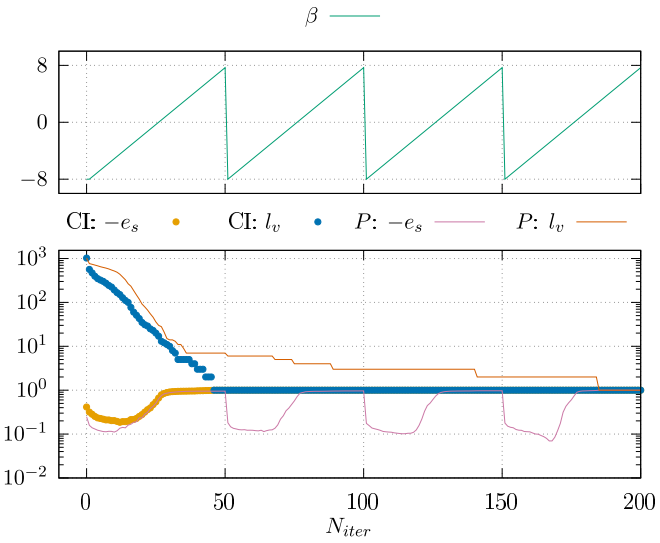


FIG. 8. Example of evolution for the Consensus-Dissent annealing process with $\beta \in [-8, 8]$. Results for the CI distribution are represented with points, and results for the $P(X \cdot Y)$ distribution are represented with lines.

is in a local minimum, and when β increases to positive values, the energy rapidly approaches -1.

In all the experiments, the initial values of l_v and e_s are what we expect. $l_v = V$ is satisfied because we start with a unique response for each node. For e_s, at the beginning, the expected value is

$$e_s^{t=0} = -\frac{1}{4V} \sum_{n,\mu} E[\boldsymbol{\phi}(n) \cdot \boldsymbol{\phi}(n+\mu)]. \tag{7}$$

At $t = 0, \boldsymbol{\phi}(n) \cdot \boldsymbol{\phi}(n+\mu)$ have the distribution shown in Fig. 1; therefore we can write $E[\boldsymbol{\phi}(n) \cdot \boldsymbol{\phi}(n+\mu)] = E[s]$.

In the case of $P(X \cdot Y)$ where $p_S(s) = -\ln(s)$, the expected value is

$$E[s] = -\int_0^1 s \ln(s) \mathrm{d}s = \frac{1}{4}. \tag{8}$$

Therefore

$$e_s^{t=0} = -\frac{1}{4V} \sum_{n,\mu} \frac{1}{4} = -\frac{1}{4}. \tag{9}$$

In the case of the CI experiment, $E[s]$ is given by the mean of the histogram in Fig. 1, which is $E[s] \approx 0.4$. Therefore

$$e_s^{t=0} \approx -\frac{1}{4V} \sum_{n,\mu} 0.4 = -0.4. \tag{10}$$

In all cases the evolution starts at the expected initial values.

Regarding the initial distribution, the only change in the system behavior is that convergence is achieved faster with the CI distribution compared to $P(X \cdot Y)$. On the other hand, the distributions used have the property that, due to the post-training of the model, the similarity is always positive; therefore in our system $e_s < 0$.

V. CONCLUSIONS

We presented a semantic-vector model for CI experiments. Textual responses from participants are converted into N-component semantic vectors ($N = 256-1024$) with the use of an embedding model. The Hamiltonian of the standard $d = 2$ $O(N)$ model, based on the scalar product between nearest-neighbor vectors on a square lattice, is then employed to define an energy that quantifies the level of consensus in the system. This energy function can then be used to define a dynamical process based only on local interactions. This approach both simplifies the method used in previous work [40], which relied on nonlocal counting of response frequency, and provides a more faithful description of the state of the system, since it can account for semantically similar but not textually identical responses.

With this model we can overwrite responses using only local information and we can extinguish responses. Furthermore, we can adjust the value of a coupling constant β so each node naturally becomes surrounded by similar ideas ($\beta > 0$) or distinct ideas ($\beta < 0$). Previously, dissent in the system could only be explored by exchanging the lattice sites of participants in the CI experiment [40].

We show that with this model we can generate consensus for $\beta > 0$, where all nodes in the lattice share the same response, and we can generate maximum dissent for $\beta < 0$, where responses are placed in a checkerboard pattern.

ACKNOWLEDGMENTS

This work was partially supported by Ministerio de Ciencia, Innovación y Universidades (Spain), and by the European Regional Development Fund (MCIU/AEI/10.13039/501100011033/FEDER, UE) through Grants No. PID2024-158623NB-C22 and No. PID2022-136374NB-C22. We were also partly funded by Gobierno de Aragón (Research Group E30_23R).

DATA AVAILABILITY

There are no publicly available research data or software supporting this manuscript. Requests for further information or data should be sent to the authors.

APPENDIX: COLLECTIVE INTELLIGENCE EXPERIMENT

We used responses generated from a collective intelligence experiment involving 656 adolescent participants [38]. The scenario presented involved two youths who experienced an incident of grooming. Six questions were posed regarding this context. In total, more than 5000 unique responses were generated in the experiment.

For each simulation we take $V = L \times L$ different responses and compute their respective embeddings with the Arctic Embed 2.0 model [47]. Each response and embedding is randomly placed on the lattice to start the simulation.

[1] M. Perc, J. J. Jordan, D. G. Rand, Z. Wang, S. Boccaletti, and A. Szolnoki, Statistical physics of human cooperation, Phys. Rep. **687**, 1 (2017).

[2] M. Jusup, P. Holme, K. Kanazawa, M. Takayasu, I. Romić, Z. Wang, S. Geček, T. Lipić, B. Podobnik, L. Wang, W. Luo, T. Klanjšček, J. Fan, S. Boccaletti, and M. Perc, Social physics, Phys. Rep. **948**, 1 (2022).

[3] B. Bollobás and O. Riordan, *Percolation* (Cambridge University Press, New York, 2006).

[4] A. Amirkhani and A. Barshooi, Consensus in multi-agent systems: A review, Artif. Intell. Rev. **55**, 3897 (2022).

[5] M. C. Marchetti, J. F. Joanny, S. Ramaswamy, T. B. Liverpool, J. Prost, M. Rao, and R. A. Simha, Hydrodynamics of soft active matter, Rev. Mod. Phys. **85**, 1143 (2013).

[6] W. van Saarlos, V. Vitelli, and Z. Zeravcic, *Soft Matter* (Princeton University Press, New Jersey, 2024).

[7] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet, Novel type of phase transition in a system of self-driven particles, Phys. Rev. Lett. **75**, 1226 (1995).

[8] M. Ballerini, N. Cabibbo, R. Candelier, A. Cavagna, E. Cisbani, I. Giardina, V. Lecomte, A. Orlandi, G. Parisi, A. Procaccini, M. Viale, and V. Zdravkovic, Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study, Proc. Natl. Acad. Sci. USA **105**, 1232 (2008).

[9] A. Cavagna and I. Giardina, Bird flocks as condensed matter, Annu. Rev. Condens. Matter Phys. **5**, 183 (2014).

[10] H. P. Zhang, A. Be'er, E.-L. Florin, and H. L. Swinney, Collective motion and density fluctuations in bacterial colonies, Proc. Natl. Acad. Sci. USA **107**, 13626 (2010).

[11] F. Ginelli, F. Peruani, M.-H. Pillot, H. Chaté, G. Theraulaz, and R. Bon, Intermittent collective dynamics emerge from conflicting imperatives in sheep herds, Proc. Natl. Acad. Sci. USA **112**, 12729 (2015).

[12] J. Toner, Y. Tu, and S. Ramaswamy, Hydrodynamics and phases of flocks, Ann. Phys. **318**, 170 (2005).

[13] F. Julicher, K. Kruse, J. Prost, and J.-F. Joanny, Active behavior of the cytoskeleton, Phys. Rep. **449**, 3 (2007).

[14] T. Vicsek and A. Zafeiris, Collective motion, Phys. Rep. **517**, 71 (2012).

[15] A. Doostmohammadi, J. Ignés-Mullol, J. Yeomans, and F. Sagués, Active nematics, Nat. Commun. **9**, 3246 (2018).

[16] P. Baconnier, O. Dauchot, V. Démery, G. Düring, S. Henkes, C. Huepe, and A. Shee, Self-aligning polar active matter, Rev. Mod. Phys. **97**, 015007 (2025).

[17] Y. Duan, J. Agudo-Canalejo, R. Golestanian, and B. Mahault, Dynamical pattern formation without self-attraction in quorum-sensing active matter: The interplay between nonreciprocity and motility, Phys. Rev. Lett. **131**, 148301 (2023).

[18] T. Lefranc, A. Dinelli, C. Fernández-Rico, R. P. A. Dullens, J. Tailleur, and D. Bartolo, Synthetic quorum sensing and absorbing phase transitions in colloidal active matter, Phys. Rev. X **15**, 031050 (2025).

[19] M. E. Cates and J. Tailleur, Motility-induced phase separation, Annu. Rev. Condens. Matter Phys. **6**, 219 (2015).

[20] M. B. Miller and B. L. Bassler, Quorum sensing in bacteria, Annu. Rev. Microbiol. **55**, 165 (2001).

[21] G. Liu, A. Patch, F. Bahar, D. Yllanes, R. D. Welch, M. C. Marchetti, S. Thutupalli, and J. W. Shaevitz, Self-driven phase transitions drive *Myxococcus xanthus* fruiting body formation, Phys. Rev. Lett. **122**, 248102 (2019).

[22] H. Zhao, A. Košmrlj, and S. S. Datta, Chemotactic motility-induced phase separation, Phys. Rev. Lett. **131**, 118301 (2023).

[23] I. Karamouzas, B. Skinner, and S. J. Guy, Universal power law governing pedestrian interactions, Phys. Rev. Lett. **113**, 238701 (2014).

[24] J. L. Silverberg, M. Bierbaum, J. P. Sethna, and I. Cohen, Collective motion of humans in mosh and circle pits at heavy metal concerts, Phys. Rev. Lett. **110**, 228701 (2013).

[25] M. Fruchart, R. Hanai, P. Littlewood, and V. Vitelli, Nonreciprocal phase transitions, Nature (London) **592**, 363 (2021).

[26] Z. You, A. Baskaran, and M. C. Marchetti, Nonreciprocity as a generic route to traveling states, Proc. Natl. Acad. Sci. USA **117**, 19767 (2020).

[27] A. V. Ivlev, J. Bartnick, M. Heinen, C.-R. Du, V. Nosenko, and H. Löwen, Statistical mechanics where Newton's third law is broken, Phys. Rev. X **5**, 011035 (2015).

[28] D. Martin, D. Seara, Y. Avni, M. Fruchart, and V. Vitelli, Transition to collective motion in nonreciprocal active matter: Coarse graining agent-based models into fluctuating hydrodynamics, Phys. Rev. X **15**, 041015 (2025).

[29] H. Hong and S. H. Strogatz, Kuramoto model of coupled oscillators with positive and negative coupling parameters: An example of conformist and contrarian oscillators, Phys. Rev. Lett. **106**, 054102 (2011).

[30] V. Guttal and I. D. Couzin, Social interactions, information use, and the evolution of collective migration, Proc. Natl. Acad. Sci. USA **107**, 16172 (2010).

[31] I. D. Couzin, C. C. Ioannou, G. Demirel, T. Gross, C. J. Torney, A. Hartnett, L. Conradt, S. A. Levin, and N. E. Leonard, Uninformed individuals promote democratic consensus in animal groups, Science **334**, 1578 (2011).

[32] D. Yllanes, M. Leoni, and M. C. Marchetti, How many dissenters does it take to disorder a flock? New J. Phys. **19**, 103026 (2017).

[33] F. Baumann, P. Lorenz-Spreen, I. M. Sokolov, and M. Starnini, Modeling echo chambers and polarization dynamics in social networks, Phys. Rev. Lett. **124**, 048301 (2020).

[34] J. Liu, S. Huang, N. M. Aden, N. F. Johnson, and C. Song, Emergence of polarization in coevolving networks, Phys. Rev. Lett. **130**, 037401 (2023).

[35] F. Baumann, P. Lorenz-Spreen, I. M. Sokolov, and M. Starnini, Emergence of polarized ideological opinions in multidimensional topic spaces, Phys. Rev. X **11**, 011012 (2021).

[36] J. Piao, J. Liu, F. Zhang, J. Su, and Y. Li, Human–AI adaptive dynamics drives the emergence of information cocoons, Nat. Mach. Intell. **5**, 1214 (2023).

[37] A. W. Woolley and I. Aggarwal, Collective intelligence and group learning, in *The Oxford Handbook of Group and Organizational Learning* (Oxford University Press, Oxford, 2020).

[38] S. Orejudo, J. Cano-Escoriaza, A. B. Cebollero-Salinas, P. Bautista, J. Clemente-Gallardo, A. Rivero, P. Rivero, and A. Tarancón, Evolutionary emergence of collective intelligence in large groups of students, Front. Psychol. **13**, 848048 (2022).

[39] A. Cebollero-Salinas, J. Cano-Escoriaza, and S. Orejudo, Social networks, emotions, and education: Design and validation of

e-COM, a scale of socio-emotional interaction competencies among adolescents, Sustainability **14**, 2566 (2022).

[40] T. García-Egea, A. Rivero, A. Tarancón, and C. Tarancón, The physics of collective human intelligence and opinion propagation on the lattice, Phys. Lett. A **522**, 129767 (2024).

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017), Vol. 30.

[42] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault (Association for Computational Linguistics, 2020), pp. 8440–8451.

[43] J. Cardy, *Scaling and Renormalization in Statistical Field Theory* (Cambridge University Press, Cambridge, 1996).

[44] J. Zinn-Justin, *Quantum Field Theory and Critical Phenomena*, 4th ed. (Clarendon Press, New York, 2005).

[45] N. D. Mermin and H. Wagner, Absence of ferromagnetism or antiferromagnetism in one- or two-dimensional isotropic Heisenberg models, Phys. Rev. Lett. **17**, 1133 (1966).

[46] P. C. Hohenberg, Existence of long-range order in one and two dimensions, Phys. Rev. **158**, 383 (1967).

[47] P. Yu, L. Merrick, G. Nuti, and D. Campos, Arctic-embed 2.0: Multilingual retrieval without compromise, arXiv:2412.04506.

[48] https://huggingface.co/Snowflake/snowflake-arctic-embed-l-v2.0.

[49] B. Halperin, On the Hohenberg–Mermin–Wagner theorem and its limitations, J. Stat. Phys. **175**, 521 (2019).

[50] D. R. Nelson and L. Peliti, Fluctuations in membranes with crystalline and hexatic order, J. Phys. France **48**, 1085 (1987).