nature machine intelligence

Article

https://doi.org/10.1038/s42256-023-00731-4

Human–AI adaptive dynamics drives the emergence of information cocoons

Received: 30 March 2023

Accepted: 6 September 2023

Published online: 09 October 2023

Check for updates

Jinghua Piao[1,3], Jiazhen Liu[1,3], Fang Zhang[2], Jun Su[2] & Yong Li[1]

Despite AI-driven recommendation algorithms being widely adopted to counter information overload, substantial evidence suggests that they are building cocoons of homogeneous contents and viewpoints, further aggravating social polarization and prejudice. Curbing these perils requires a deep insight into the origin of information cocoons. Here we investigate information cocoons in the real world using two large datasets and find that a large number of users are trapped in information cocoons. Further empirical analysis suggests that two ingredients, each corresponding to a fundamental mechanism in human–AI interaction systems, are correlated with the loss of information diversity. Grounded on the empirical findings, we derive a mechanistic model for the adaptive information dynamics in complex human–AI interaction systems governed by these fundamental mechanisms. It allows us to predict critical transitions between three states: diversification, partial information cocoons, and deep information cocoons. Our work not only empirically traces real-world information cocoons in two representative scenarios, but also theoretically unearths basic mechanisms governing the emergence of information cocoons. We provide a theoretical method for understanding major social issues resulting from adaptive information dynamics in complex human–AI interaction systems.

Artificial intelligence (AI) has permeated all kinds of human activity and catapulted algorithms into aspects of modern life[1]. As one of the best-known AI-driven technologies, recommendation algorithms are widely adopted to counter the problem of information overload[2,3]. Their widespread adoption ranges from the consumption of news[4,5] and videos[6,7] to friendship establishment[8]. However, recent years have witnessed that AI-driven recommendation algorithms are driving the formation of information cocoons (ICs)[9], in which individuals are being isolated from diverse information and eventually trapped in a single topic or viewpoint. Exposure to homogeneous information not only deprives humans of the diversity of information available for informed decision-making[10–12] but also exacerbates social polarization[13,14] and reinforces biases[4,15]. The perils of ICs are far-reaching, as they stifle creativity and innovation[16], impede progress toward a more inclusive world[17], and ultimately threaten the diversity of our society[8,13,15]. To curb these perils, understanding the origin of ICs is the crucial first step.

Current studies on the homogeneity process of online information focus primarily on either human behaviours or intelligent algorithms[4,10,11,15,18–24]. Though they empirically explore the potential drivers in the aggregation of homogeneous populations on social media[15,18,20,24] or the algorithmic filtering effects[4,20], they only provide correlational evidence. Recently, a few empirical studies[10,25,26] have conducted a causal analysis based on statistical methods. However, they do not offer mechanistic analysis or dynamic insights; hence, the fundamental mechanisms driving the system into ICs remain unexplored. The lack of insight into the mechanisms has a deep origin: current AI-driven recommendation algorithms are deeply rooted in deep learning methods[3,6]. Their black-box nature, originating from billions of parameters[27,28], further hinders an in-depth understanding of ICs from a purely empirical perspective.

[1]Beijing National Research Center for Information Science and Technology (BNRist), Department of Electronic Engineering, Tsinghua University, Beijing, People's Republic of China. [2]School of Public Policy and Management, Tsinghua University, Beijing, People's Republic of China. [3]These authors contributed equally: Jinghua Piao, Jiazhen Liu. e-mail: liyong07@tsinghua.edu.cn

On the other hand, a number of theoretical studies have pointed out that the adaptive dynamics, that is, features of individuals in complex systems co-evolving with the reformation of ties between individuals, is responsible for the homogeneity process in many social systems[13,14,22]. These studies hint that the adaptive dynamics in human–AI interaction systems is possibly the underlying driving force of the emergence of ICs. However, the existing studies lack insights into the mechanisms of co-evolution between humans and AI algorithms. This is because the interactions between humans and AI involve multiple entities and feedback, leading to intricate dynamical properties underpinning these interactions[11,21]. Hence, the origin of ICs remains unknown.

The purpose of this paper is to uncover the origin of ICs in the complex human–AI interaction system. As discussed above, existing studies only provide empirical evidence or qualitative models, lacking insights into the fundamental mechanisms driving the system towards the emergence of ICs. This requires a theoretical framework capable of accounting for the adaptive information dynamics underlying ICs. To explain the origin of the ICs, we point out that there are two mechanisms, (1) similarity-based matching and (2) positive/negative feedback, serving as the starting point of an adaptive information dynamics model for complex human–AI interactions. We analytically predict the transitions between information homogeneity and diversification states, validated by extensive simulations and empirical observations involving over 570 million records in two representative scenarios. We reveal that the imbalance between positive and negative feedback lures the system to move toward ICs, and then similarity-based matching further reinforces this tendency, eventually leading the system to deep ICs. We not only uncover the origin of ICs but also provide a theoretical method for understanding major social issues emerging from complex human–AI interactions.

Results

Empirical observations

To investigate real-world ICs, we adopt two large-scale real-world datasets. The first dataset is collected from one of the top three short-form video platforms in China. This dataset contains more than 111,000 (111K) new users, 9,000,000 (9M) videos and these users' 500M interaction records in the entire year of 2021. In the video dataset, there are 20 video topics. The second dataset is collected from Microsoft News[5], including 14 news topics with more than 90K users, 130K pieces of news and 36M interaction records for 6 weeks. We report detailed descriptive and temporal statistics of the two datasets in Supplementary Sections 1 and 2.

To empirically quantify the diversity of information accessible to users, we measure information entropy $s = -\sum_{j=1}^{N} f_l^{(j)} \log f_l^{(j)}$ for each user l[29,30] on two empirical datasets (Supplementary Section 1), where N is the number of unique topics and $f_l^{(j)} = \frac{n_l^{(j)}}{\sum_{j=1}^{N} n_l^{(j)}}$ is the relative frequency with which user l accesses items belonging to topic j ($n_l^{(j)}$ is the number of times user l accesses topic j). By measuring overall changes in entropy Δs over one year, we find that over 57% of the active users have witnessed a decline in information entropy (Fig. 1a), suggesting the prevalence of ICs across users in the real world. The users with a decline in entropy ($\Delta s < 0$) have all experienced a rapid slump in information diversity at an early stage of entering the platform (Fig. 1b). In particular, users in group 1, 11% of the total, have suffered a sharp decline for around a year, eventually witnessing a 24.8% drop in entropy. Indeed, we find that sharply falling diversity of topics has characterized the interactions between AI-driven recommendation algorithms and users. By randomly selecting a user from group 1, in which users have experienced the largest decline in information diversity (Fig. 1c), we observe a striking pattern that this user, entering the system with a wide variety of accessible topics, is eventually confined to a single topic after 1 year of interactions (see other cases in Supplementary Fig. 6). We notice that there is a subtle increment in the final state of group 5. This suggests that, in contrast to other groups, group 5 is not in a deep IC, and some fluctuations or interventions in the system could potentially bring more diverse information to them. However, the subtle increment does not change the apparent declining trend over the whole period (Supplementary Section 2.1). Overall, the above large-scale data analysis indicates that a large proportion of users suffer from a significant loss of information diversity, suggesting the existence of serious ICs, raising a critical but largely unanswered question: what are the key ingredients driving users toward ICs in the human–AI system?

On online platforms, AI-driven recommendation algorithms are deeply rooted in two common mechanisms: (1) similarity-based matching and (2) utilization of users' feedback. Similarity-based matching is the most fundamental mechanism in recommendation algorithms, designed to recommend items similar to those users liked in the past[2,3]. In response to the recommended items, users naturally give positive or negative feedback according to their preferences[2,3,31,32], where positive feedback reflects what users like while negative feedback reflects what users dislike. Then algorithms utilize the feedback as the prerequisite of the next-step similarity-based matching[2,3,6], forming a simple but fundamental feedback loop for the human–AI interaction system[2,3,33].

To empirically explore how these mechanisms have effects on the information homogeneity of users, we measure the normalized information entropy $\hat{s}$ and the similarity-based matching strength (Fig. 1d; see details in Supplementary Section 2), finding a negative correlation between them. Meanwhile, we observe that the ratio of positive feedback samples is negatively correlated with $\hat{s}$ (Fig. 1e), whereas the correlation between the ratio of negative feedback samples and $\hat{s}$ is positive (Fig. 1f). Though the statistical evidence shows correlations between the ICs and these key factors, they offer limited information about the basic mechanisms underlying the dynamics of the human–AI interaction system, that is, how the AI adaptively utilizes the users' feedback in the dynamic feedback loop and eventually drives users into ICs. This suggests that a theoretical model is required to uncover the mechanisms underlying the dynamics driving the emergence of ICs.

Adaptive information dynamics model

Below we focus on proposing an adaptive dynamics modelling framework for complex human–AI interaction systems to account for the emergence of ICs. In contrast to the deep-learning-based model incorporating billions of parameters, the proposed model only depends on four parameters originating from both the empirical observation and the working principle of current recommendation algorithms[2,3,6,33,34], integrating similarity-based matching, users' feedback and human exploration behaviours in the modelling framework (Fig. 1g).

(i) Similarity-based matching captures the similarity between a user's observed preference $\mathbf{u}_l = [u_l^{(1)}, u_l^{(2)}, \ldots, u_l^{(N)}]$ and item feature $\mathbf{i}_k = [i_k^{(1)}, i_k^{(2)}, \ldots, i_k^{(N)}]$, where $\sum_{j=1}^{N} u_l^{(j)} = 1, \sum_{j=1}^{N} i_k^{(j)} = 1$ and $0 \le u_l^{(j)} \le 1, 0 \le i_k^{(j)} \le 1$. Previous studies measured the similarity using metrics, such as the inner product, the cosine similarity and the Jensen–Shannon divergence[2]. Here we focus on the inner product in the main text and also include results of other metrics in the supplementary information (Supplementary Sections 2.2 and 4.3). On the basis of the similarity denoted by $\theta(\mathbf{u}_l, \mathbf{i}_k)$, the recommendation algorithm matches user l and item k with a probability proportional to $p_{lk} = \frac{e^{\beta\theta(\mathbf{u}_l, \mathbf{i}_k)}}{\sum_{\mathbf{i}_{k'} \in I} e^{\beta\theta(\mathbf{u}_l, \mathbf{i}_{k'})}}$, where I is the set of all candidate items and where β controls the strength of similarity-based matching (Description of adaptive information dynamics model). $\beta = 0$ represents a purely random recommendation strategy, whereas large β leads to the strategy that

https://doi.org/10.1038/s42256-023-00731-4

Fig. 1 | Empirical observations on ICs and our proposed adaptive information dynamics model. a, Δs, where the majority of active users (57%) have experienced increasingly homogeneous recommendation results. We divide these users evenly into five groups according to Δs. **b**, Temporal changes in s, where lines correspond to groups of users with different Δs ($n_1 = n_2 = n_3 = n_4 = n_5 = 7{,}359$), error bars represent 95% confidence intervals (CIs) and centres of error bars represent the average values. With the increase in interactions, 11% of the overall users (group 1) have witnessed a total drop in entropy from 2.32 to 1.75 (24.8%). **c**, Example of a randomly selected user in group 1, who has been strongly confined to homogeneous information by AI-driven recommendation algorithms. **d**–**f**, Correlation between $\hat{s}$ and the strength of similarity-based matching (**d**), the ratio of positive feedback samples (**e**) and the ratio of negative feedback samples (**f**). For illustration, we normalize information entropy over all users using the min–max method. Here shading represents the corresponding 95% prediction intervals and centres represent the estimated linear correlation functions. **g**, Overview of an adaptive information dynamics model, where humans and the AI-driven recommendation algorithm interact with each other, forming a feedback loop. The AI matches users and items on the basis of the estimated similarity (β). Users provide feedback. The AI learns from users' positive feedback (γ_+) and negative feedback (γ_-), as well as random self-exploration (σ), and then makes further recommendations.

items with high similarity scores are more likely to be recommended. Overall, the algorithm matches each user l with a set of similar items denoted by R_β. Since both datasets present most items only belonging to a single topic j, that is, $i_k^{(j)} = 1$ and $i_k^{(j' \neq j)} = 0$, we obtain s for each user l,

$$s = -\sum_{j=1}^{N} f_l^{(j)} \log f_l^{(j)} \approx -\sum_{j=1}^{N} \frac{1+\beta u_l^{(j)}}{N+\beta} \log \frac{1+\beta u_l^{(j)}}{N+\beta}. \quad (1)$$

Note that, in contrast to p_{lk} measuring the interaction probability between user l and item k, $f_l^{(j)}$ denotes the frequency of recommending each topic j to each user l, which is largely determined by the observed preference $\mathbf{u}_l$ (Description of adaptive information dynamics model).

(ii) The decision of a user to give positive feedback to the recommendation is largely determined by the similarity $\theta(\mathbf{x}_l, \mathbf{i}_k)$ between the user's intrinsic preference $\mathbf{x}_l = [x_l^{(1)}, x_l^{(2)}, \ldots, x_l^{(N)}]$ and item feature $\mathbf{i}_k$. Note that in contrast to $\mathbf{u}_l$, representing the preference speculated by the algorithm, the intrinsic preference

$\mathbf{x}_l$ encodes the inherent preference (Description of adaptive information dynamics model). Assuming random human decision-making, the probability of accepting the recommendation is denoted by $\pi(\mathbf{x}_l, \mathbf{i}_k) = \theta(\mathbf{x}_l, \mathbf{i}_k)$, whereas the probability of declining the recommended items is $1 - \pi(\mathbf{x}_l, \mathbf{i}_k)$. Items receiving positive feedback make up a set of positive feedback samples $R_\beta^+ \subseteq R_\beta$ and the others make up a set of negative feedback samples $R_\beta^- \subseteq R_\beta$.

(iii) Users also proactively explore through other resources, for example, search engines[34]. These exploration behaviours are also an important input of recommendation algorithms; hence we model users' proactive exploration behaviours as a stochastic process by introducing σ to capture the degree of random self-exploration.

Combining (i)–(iii), we obtain the following stochastic differential equation, capturing how the AI adaptively utilizes user feedback to update user l's observed preference $\mathbf{u}_l$:

$$\mathrm{d}\mathbf{u}_l = \gamma_+ \sum_{\mathbf{i}_k \in R_\beta^+} F(\mathbf{u}_l, \mathbf{i}_k)\,\mathrm{d}t + \gamma_- \sum_{\mathbf{i}_k \in R_\beta^-} F(\mathbf{u}_l, \mathbf{i}_k)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, \qquad (2)$$

where $\gamma_+ \geq 0$ and $\gamma_- \leq 0$ are the algorithmic utilization rates of positive feedback $\mathbf{i}_k \in R_\beta^+$ and negative feedback $\mathbf{i}_k \in R_\beta^-$. Physically, γ_+ and γ_- govern how likely the AI-driven algorithm is to recommend desirable and undesirable items to users. $F(\mathbf{u}_l, \mathbf{i}_k)$ represents the interaction function between a user l's observed preference $\mathbf{u}_l$ and a recommended item k's feature. W_t is the standard Wiener process[13,35,36] (Description of adaptive information dynamics model and Supplementary Section 3.1). Here we adopt $F(\mathbf{u}_l, \mathbf{i}_k) = \mathbf{i}_k - \mathbf{u}_l$, measuring the distance between $\mathbf{u}_l$ and $\mathbf{i}_k$ (Minimal model). Note that the different domains of $\gamma_+ \geq 0$ and $\gamma_- \leq 0$ allow $\mathbf{u}_l$ to be updated in opposite directions. The evolution of the observed preferences is characterized by three parts: the first term governed by γ_+ guides $\mathbf{u}_l$ to be more similar to positive feedback, the second term controlled by γ_- drives $\mathbf{u}_l$ to be more different from negative feedback and the third term determined by σ captures the algorithmic utilization degree of random self-exploration. Note that if the recommendation algorithm can perfectly infer user preferences, then the observed preference $\mathbf{u}_l$ should be equal to the intrinsic preference $\mathbf{x}_l$; however, such a perfect inference is still beyond the current AI's capabilities.

Emergence of ICs

Incorporating equation (1) with equation (2), we can derive the distribution of relative information entropy $P(\tilde{s} \equiv \{s/s^*\})$, where $s^* = -\sum_{j=1}^{N} x_l^{(j)} \ln x_l^{(j)}$ denotes the entropy of users' intrinsic preferences $x_l^{(j)}$ (Analytical solution). $P(\tilde{s})$ characterizes the distribution of information diversity relative to intrinsic preference diversity with parameters β, γ_+ and σ. As suggested by both theoretical analysis and simulation (Analytical solution, Minimal model and Supplementary Sections 3.2 and 3.3), ICs are exacerbated by amplified β (Fig. 2a) and magnified $|\gamma_+|$ (Fig. 2b).

Starting from purely random matching, that is, $\beta = 0$, we gradually increase β, observing three unexpected patterns characterized by different degrees of relative information diversity in both theoretical and simulation results (Fig. 2a). A state of diversification, captured by a single-peaked $P(\tilde{s})$ with $\langle \tilde{s} \rangle \approx 1$, resulted from random matching, whereas a large $\beta = 7$ drives the users to the state of deep ICs characterized by a single-peaked $P(\tilde{s})$ with a strikingly small $\langle \tilde{s} \rangle \approx 0$. In particular, with $\beta = 4$, $P(\tilde{s})$ exhibits a notable bimodal distribution with $0.5 \leq \langle \tilde{s} \rangle \leq 1$, suggesting that the system is in the state of partial ICs. Here users are differentiated into two different groups: (1) the group with a lower $\langle \tilde{s} \rangle$ and (2) that with a higher $\langle \tilde{s} \rangle$ (dark-green and light-green bars in Fig. 2a). We formally formulate the states of diversification, deep ICs and partial ICs in equation (8) in Minimal model. To further explore the reason for the differentiation, we measure the intrinsic preference distributions for these two groups of users, finding that users with a lower $\langle \tilde{s} \rangle$ have narrower preferences than others (Supplementary Section 4.2).

Similarly, our model further demonstrates that the magnified $|\gamma_+|$ tends to drive the system to the state of ICs when $|\gamma_-|$ is relatively small. With increasing $|\gamma_+|$, the state transits from diversification to partial ICs, then from partial ICs to deep ICs (Fig. 2b). Accordingly, our theoretical analyses offer analytical results of $P(\tilde{s})$ (dashed lines in Fig. 2b) and capture the state transitions, exhibiting considerable agreement with the simulation results (Analytical solution, Minimal model and Supplementary Section 3.2). Overall, these findings suggest that the overuse of positive feedback induces the emergence of ICs across some or even all users.

To account for the emergence of ICs, we further evaluate the degree of over-recommendation and under-recommendation on certain topics (Minimal model). We find that, in the state of deep ICs, the algorithm tends to recommend items on a few topics, over-exploiting a small fraction of users' preferences. On the other hand, the algorithm underestimates users' preferences on a large proportion of topics (Fig. 2c and Supplementary Fig. 35). The analysis suggests the cause of ICs: the excessive use of positive feedback leads to partial observations of user preferences. When a feedback loop is dominated by strong similarity-based matching strength, partial observations are unavoidably further reinforced, leading to over-recommendation of certain topics and under-recommendation of others.

The above results raise a natural question: what is the effect of negative feedback? The answer is that increasing $|\gamma_-|$ enables users to escape from ICs. The state of diversification is observed when $|\gamma_-|$ is large, whereas the state of deep ICs arises with marginal $|\gamma_-|$ (Fig. 3a), exhibiting a remarkable reversal transition pattern compared with Fig. 2b. This raises a puzzling question: how does the increasing utilization of negative feedback suppress the emergence of ICs? To answer this, we further analyse the change in the available topics for individual users. By randomly selecting a user as an example (Supplementary Section 4.9), we find that, with increasing utilization of negative feedback, the user is no longer limited to a single topic but can access a variety of matching topics (Fig. 3c). Our theoretical analyses consistently suggest that there is antagonism between positive and negative feedback in the information dynamics between humans and AI-driven recommendation algorithms (Analytical solution and Supplementary Section 3.2). Indeed, since the algorithm estimates similarity on the basis of what users like, that is, positive feedback, the overall system is naturally biased toward positive feedback. In this case, overlooking negative feedback will further exacerbate the bias, leading to the underestimation of the diversity of users' preferences. This implies that the efficient utilization of negative feedback can counteract the side effect of the overuse of positive feedback, restoring the system to a state away from ICs.

As the emergence of ICs can be suppressed by efficient utilization of negative feedback, we may wonder if there is a way for users to proactively stay away from ICs. The answer lies in the users' random self-exploration. Figure 3b shows that, even when the AI-driven recommendation algorithm excessively depends on similarity-based matching and positive feedback ($\beta = 10$, $\gamma_+ = 1$), a slight increase in the self-exploration degree σ can still effectively keep the system away from ICs. In particular, by slightly raising σ from 0.04 to 0.07, some users directly get rid of deep ICs and reach a diversification state, that is, $\tilde{s} \approx 1$ (light green plots in the inset of Fig. 3b). Indeed, Fig. 3d suggests that a high self-exploration rate σ helps the algorithm to more comprehensively and precisely capture users' intrinsic preferences, hence offering more diverse and matching recommendations. To test the model's robustness, we consider the circumstances in the news dataset (Supplementary Section 4.1) and other widely adopted similarity metrics in practical industrial recommendation algorithms[2], such as the cosine similarity and the Jensen–Shannon divergence (Supplementary

https://doi.org/10.1038/s42256-023-00731-4

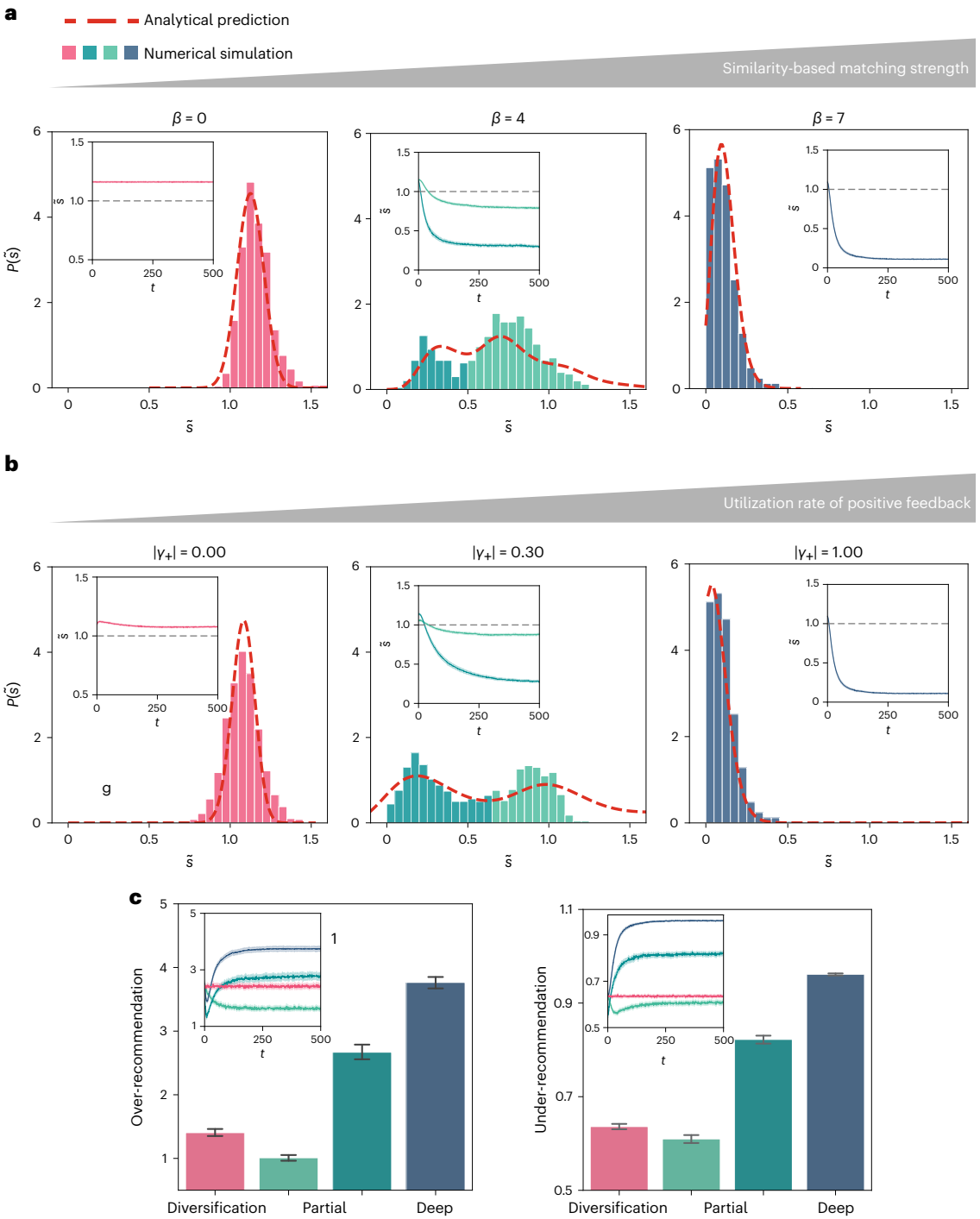

Fig. 2 | Effects of β and $|\gamma_+|$ on ICs. a, Distributions of relative information entropy $P(\tilde{s})$ over different β. **b**, Distributions of relative information entropy $P(\tilde{s})$ over different $|\gamma_+|$. **a**, **b**, Insets: temporal changes in $\tilde{s}$, where shading represents 95% CIs. The excessive use of similarity-based matching and positive feedback leads to the emergence of ICs. **c**, Degrees of over-recommendation and under-recommendation among different states, where bars represent the average values ($n_{\text{diversification}} = 1{,}000$, $n_{\text{partial left}} = 785$, $n_{\text{partial right}} = 215$, $n_{\text{deep}} = 1{,}000$) and error bars represent 95% CIs. Insets: temporal changes in degrees of over-recommendation and under-recommendation, where shading represents 95% CIs. Users in deep ICs experience the largest degrees of both over-recommendation and under-recommendation.

Section 4.3). Moreover, in two real-world datasets, we observe that the empirical information entropy distribution $P(s)$ shows three distinct states, in line with our predictions (Supplementary Sections 2.5 and 4.5), further illustrating the validity of our findings obtained from the proposed adaptive information dynamics model.

In our final analysis, we explore how the interplay among the ingredients of similarity-based matching, positive and negative feedback, and self-exploration affects transitions between the three states of ICs. We conduct the simulations with two different initial observed preferences obtained from the video and news datasets (Minimal model and Supplementary Section 3.3). Given the antagonistic effect between positive and negative feedback on the emergence of ICs (suggested by our theoretical analysis in Analytical solution and simulation results in Figs. 2b and 3a), we define the relative utilization rate of feedback

https://doi.org/10.1038/s42256-023-00731-4

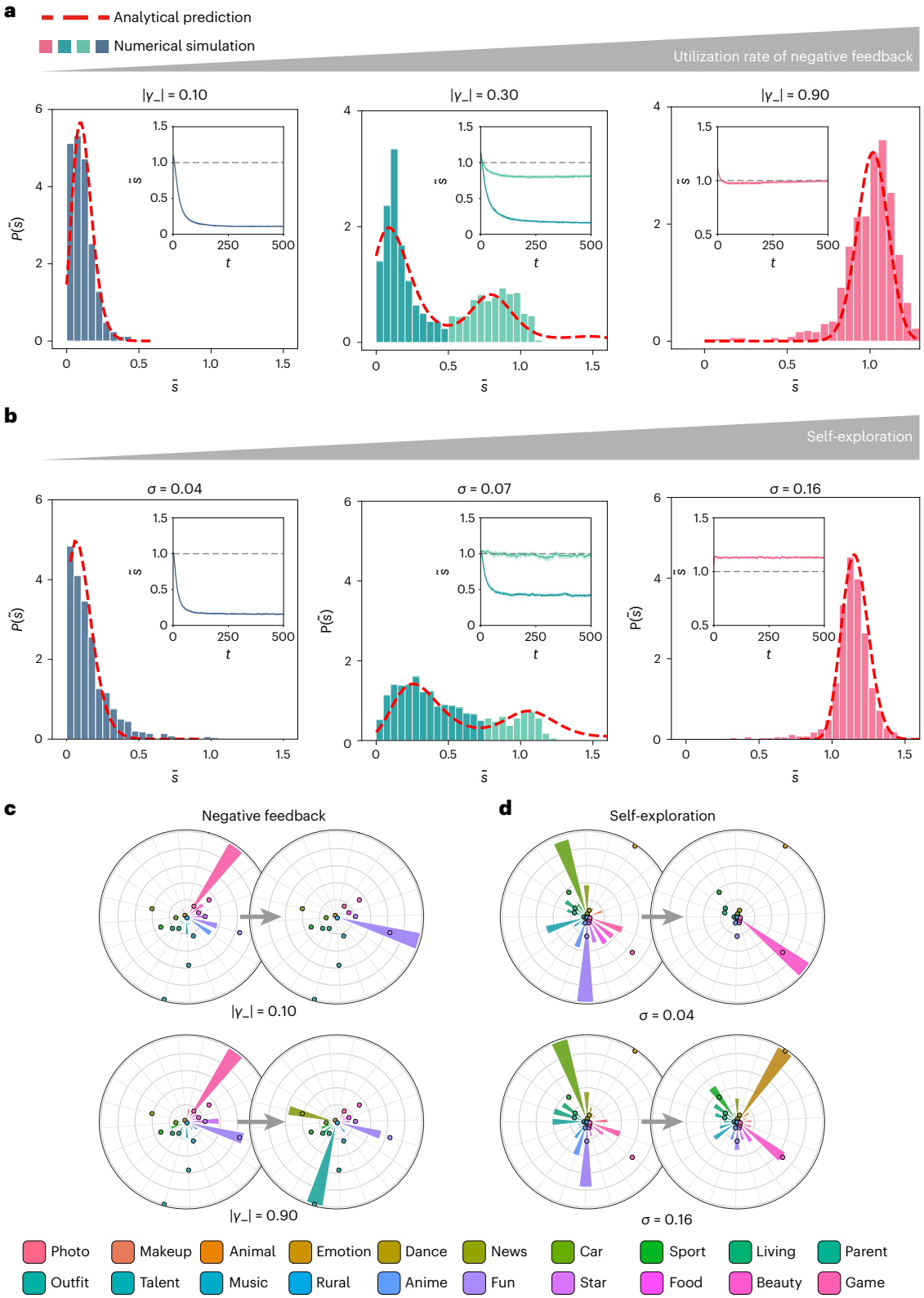

Fig. 3 | Effects of $|\gamma_-|$ and σ on ICs. a, Distributions of relative information entropy $P(\tilde{s})$ over different $|\gamma_-|$. **b**, Distributions of relative information entropy $P(\tilde{s})$ over different σ. **a**, **b**, Insets: temporal changes in $\tilde{s}$, where shading represents 95% CIs. Efficient utilization of negative feedback and users' active exploration behaviours suppress the emergence of ICs. **c**, Comparison of the initial and the accessible topic distributions for a randomly sampled user when $\gamma_- = -0.10$ (upper) and $\gamma_- = -0.90$ (lower). **d**, Comparison of the initial and the accessible topic distributions for a randomly sampled user when $\sigma = 0.04$ (upper) and $\sigma = 0.16$ (lower). **c**, **d**, Dots represent the intrinsic preference of the user and bars represent the distribution of available topics.

https://doi.org/10.1038/s42256-023-00731-4

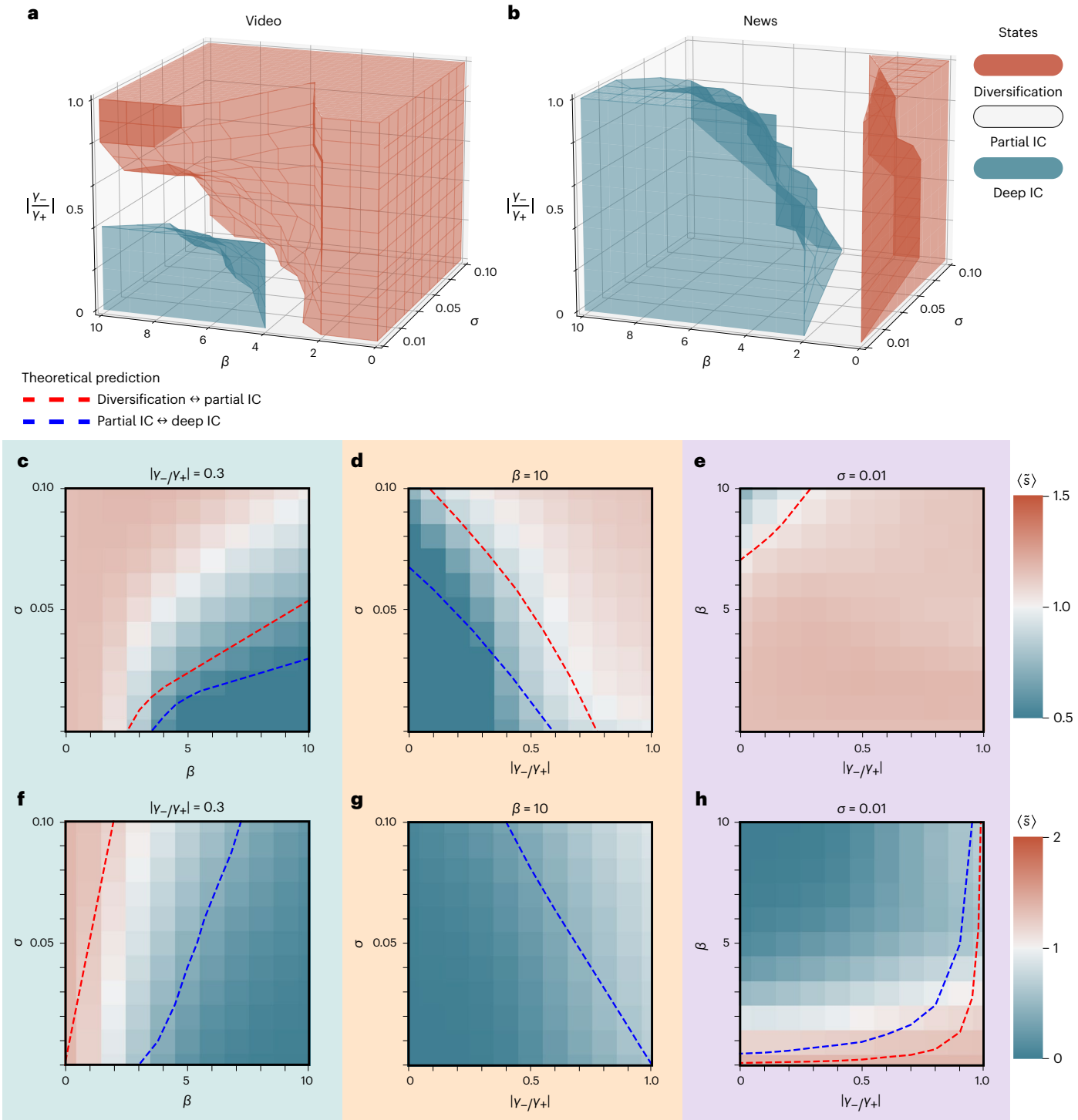

Fig. 4 | State diagram. State transitions of ICs with system parameters β, σ and $|\gamma_-/\gamma_+|$. **a**,**b**, Three-dimensional state diagrams in simulations initialized by the video dataset (**a**) and the news dataset (**b**). **c**–**h**, Cross-sections of three-dimensional diagrams of simulations initialized by the video dataset (**c**–**e**) and the news dataset (**f**–**h**). See other cross-sections of three-dimensional diagrams in Supplementary Figs. 27 and 28.

$|\gamma_-/\gamma_+|$, characterizing the update rate of negative feedback relative to that of positive feedback. Figure 4a,b shows three-dimensional state diagrams for video and news recommendations (see formal definition of states in Methods, equation (8)). Incorporating equations (7) with (8) predicts two state transitions, from diversification to partial ICs, and then from partial ICs to deep ICs, qualitatively agreeing with the numerical experiment (Fig. 4c–h). The transition between states occurring under the critical parameters suggests that ICs are relievable by a proper balance between the discussed ingredients. Indeed, the cross-sections of three-dimensional state diagrams (Fig. 4c–h and Supplementary Figs. 27 and 28) illustrate the details: even though β is significant, the slightly increasing σ and $|\gamma_-/\gamma_+|$ drive the system away from ICs. This demonstrates that the balance between positive and negative feedback, as well as similarity-based matching and initiatives to encourage self-exploration, protects humans from ICs.

Discussion

In this Article, we propose an adaptive dynamics model to unearth the origin of ICs in complex human–AI interaction systems. Grounded on large-scale empirical observations, the model allows us to capture the mechanisms behind the emergence of ICs and analytically predict transitions between different information homogeneity states in the complex human–AI interaction system. We show that similarity-based matching acts as an effective force field in complex system theory, driving the system from diversification to ICs. Positive feedback further amplifies this effect, resulting in a decrease in information entropy: that is, information diversity. Negative feedback and random self-exploration promote information diversity by resisting the effect of the effective force field and introducing perturbation to the system. This resistance drives the system from ICs to diversification. Our findings are supported by extensive experiments, validating the robustness of our model (Supplementary Sections 3.2, 4.1, 4.3–4.7 and 5).

As AI technologies become more ubiquitous in our daily lives, the interactions between humans and AI create a complex system involving multiple entities and feedback. Moreover, current AI-driven algorithms are deeply rooted in deep learning, and infamous for their black-box nature[27,28]. This hinders us from understanding dynamical properties and emergent behaviours in complex human–AI interaction systems. Our proposed adaptive information dynamics model remedies this situation by offering a mechanistic understanding of the emergence of ICs, providing a powerful theoretical tool for diving into the underlying mechanisms governing the adaptive dynamics in complex human–AI interaction systems.

It is worth pointing out that our model not only explains the emergence of ICs but also accounts for the increase in information entropy observed in the empirical analysis (Supplementary Section 4.8), validated by extensive simulations and large-scale empirical observations. The decrease and increase in entropy are the two sides of the same 'coin', representing different evolution directions of the system. Supplementary Figure 33 shows that the model reproduces the empirically observed $P(\Delta s)$. The numerical simulations further show that increased (decreased) use of negative (positive) feedback and randomness (similarity-based matching) allows more users to experience a growth of entropy. This analysis of the user group with increased entropy is in line with our main conclusion: striking a balance between negative and positive feedback, as well as a balance between similarity-based matching and randomness, helps users escape from ICs. This further implies that our model has the capability to provide solutions to tackle ICs, that is, promote the growth of entropy, by controlling/adjusting the key parameters.

The proposed model, incorporating only two mechanisms, effectively elucidates the dynamics of information entropy and offers two practical ways to mitigate real-world ICs: (1) effective utilization of negative feedback, which offers a new perspective on users' preferences by identifying their dislikes, and (2) promotion of self-exploration, which diversifies the available information by empowering users to exercise greater autonomy over the algorithm (see detailed design implications in Supplementary Section 5). Moreover, the theoretical model allows us to infer the systematic parameters from the empirical data. This helps us identify the most vulnerable modules in the current recommendation algorithm and develop strategies for mitigating ICs (Supplementary Sections 4.8 and 5). Altogether, we not only empower AI-driven recommendation algorithms with practical directions towards social good, but also offer a theoretical method for understanding major social issues resulting from adaptive dynamics in complex human–AI interaction systems.

Note that, although our study empirically and theoretically uncovers the origin of ICs, it is not without limitations. First, we mainly focus on how ICs emerge from complex human–AI interaction systems. Therefore, for simplicity and generality, we consider a minimal human decision-making model and quantify the diversity of available information by using the information entropy of recommendations, leaving the more complicated human behaviours for future exploration. Second, we ground this study on the two most concerning scenarios, that is, news and videos; hence, considering other recommendation scenarios, for example, e-commerce, would be one of the future directions.

Methods

Description of adaptive information dynamics model

To account for the origin of ICs, we develop a model for the adaptive information dynamics in the interaction feedback loop between humans and AI-driven recommendation algorithm. Such a feedback loop is characterized by essential ingredients: (1) similarity-based matching and (2) users' feedback. Analogous to stochastic thermodynamics theory, the overall system of humans and the recommendation algorithm is initialized to be away from equilibrium. With the effective force field generated by similarity-based matching, the system evolves gradually from the diversification state to the IC one, characterized by a decline in information entropy. Below we will introduce our proposed model in detail.

To match users with suitable items, the recommendation algorithm first evaluates the similarity between user l's observed preference and item k's feature: $\theta(\mathbf{u}_l, \mathbf{i}_k) = \mathbf{u}_l^{\mathrm{T}}\mathbf{i}_k$, where $\mathbf{i}_k$ represents item k's feature and $\mathbf{u}_l$ presents user l's preferences observed by the algorithm. Note that $\mathbf{i}_k$ is a vector, $\mathbf{i}_k = [i_k^{(1)}, i_k^{(2)}, \ldots, i_k^{(N)}]$, where $i_k^{(j)} = 1$ and $i_k^{(j'\neq j)} = 0$ denote that item k belongs to a single topic j. We assume that items' features do not change with time and the recommendation algorithm can capture them. Considering that the algorithm cannot directly capture users' intrinsic preferences but only observes them on the basis of previous feedback, we define user l's observed preference $\mathbf{u}_l = [u_l^{(1)}, u_l^{(2)}, \ldots, u_l^{(N)}]$, where $\sum_{j=1}^{N} u_l^{(j)} = 1, 0 \leq u_l^{(j)} \leq 1$. Here we adopt the inner product as the similarity metric; this can also be replaced with other widely adopted metrics[2], such as the cosine similarity and the Jensen–Shannon divergence (Supplementary Section 4.3).

Inspired by the well known collaborative filtering techniques in recommendation algorithms[2], we design the topic correlation matrix Φ to reflect topic correlations established on the basis of users' collective behaviours. For example, beer and nappies, even though they are semantically unrelated, have very similar customers, so they are always recommended together. We obtain Φ from empirical data by computing the correlation coefficient matrix of the estimated users' preference vectors (Supplementary Section 3.3.2). We insert the topic correlation matrix into the similarity metric, formulated as $\theta(\mathbf{u}_l, \Phi\mathbf{i}_k) = \mathbf{u}_l^{\mathrm{T}}\Phi\mathbf{i}_k$. On the basis of the estimated similarity, the probability of each item k being recommended to user l satisfies

$$p_{lk} = \frac{e^{\beta\theta(\mathbf{u}_l, \Phi\mathbf{i}_k)}}{\sum_{\mathbf{i}_{k'}\in I} e^{\beta\theta(\mathbf{u}_l, \Phi\mathbf{i}_{k'})}} = \frac{1 + \beta\theta(\mathbf{u}_l, \Phi\mathbf{i}_k) + O\left(\beta^2\theta(\mathbf{u}_l, \Phi\mathbf{i}_k)^2\right)}{\sum_{\mathbf{i}_{k'}\in I}\left(1 + \beta\theta(\mathbf{u}_l, \Phi\mathbf{i}_{k'}) + O\left(\beta^2\theta(\mathbf{u}_l, \Phi\mathbf{i}_{k'})^2\right)\right)}, \quad (3)$$

where p_{lk} is an analogy to the Boltzmann distribution in statistical physics: the numerator represents the energy of the state $\theta(\mathbf{u}_l, \Phi\mathbf{i}_k)$, the denominator is recognized as the partition function and the parameter β represents the reciprocal of the thermodynamic temperature of the system. Since both datasets present fairly weak correlations between most topics (Supplementary Fig. S8) and most items only belong to a single topic j, that is, $i_k^{(j)} = 1$ and $i_k^{(j'\neq j)} = 0$, we can approximate the similarity $\theta(\mathbf{u}_l, \Phi\mathbf{i}_k) = \mathbf{u}_l^{\mathrm{T}}\Phi\mathbf{i}_k \approx \mathbf{u}_l^{\mathrm{T}}\mathbf{i}_k = u_l^{(j)}$. Moreover, in our theoretical derivation, without loss of generality, we assume that items are uniformly distributed across topics. Incorporating $\sum_{j=1}^{N} u_l^{(j)} = 1, 0 \leq u_l^{(j)} \leq 1$ with equation (3) leads to $f_l^{(j)} \approx \frac{1+\beta u_l^{(j)}}{N+\beta}$, capturing the probability of recommending topic j to user l.

Facing a recommended item, users have two types of feedback, that is, positive and negative feedback. Positive feedback corresponds to behaviours reflecting that the user prefers the item to others:

for example, like, click or purchase. Negative feedback corresponds to behaviours reflecting that the user is not interested in the recommended items: for example, skip. Without loss of generality, we assume that user l accepts the recommended item k with probability $\pi(\mathbf{x}_l, \mathbf{i}_k) = \theta(\mathbf{x}_l, \mathbf{i}_k) = \mathbf{x}_l^{\mathrm{T}}\mathbf{i}_k$, where $\mathbf{x}_l$ denotes user l's intrinsic preference $\mathbf{x}_l = [x_l^{(0)}, x_l^{(1)}, \ldots, x_l^{(N)}]$ and $\sum_{j=1}^{N} x_l^{(j)} = 1$. As such, items receiving positive feedback make up the set of positive feedback samples R_β^+ and others make up the set of negative feedback samples R_β^-.

Despite the prevalence of recommendation algorithms, they are not the only information source[34]; users can proactively explore information through other sources, such as search engines[34]. Thus, inspired by previous work[13,35,36], we consider users' self-exploration behaviours as a Wiener process in each topic's dimension, and the degree of self-exploration is controlled by the parameter σ.

Considering users' positive and negative feedback as well as self-exploration behaviours, the recommendation algorithm updates user l's observed preference $\mathbf{u}_l$ according to the Langevin equation (2). In the equation, the utilization of positive feedback (γ_+) amplifies the effective force field of similarity-based matching, resulting in a more ordered system characterized by smaller information entropy. As a counterbalance, the utilization of negative feedback (γ_-) resists the impact of the effective force field. Users' exploration behaviours (σ) introduce random perturbations into the system, substantially increasing information entropy and preventing ICs.

Analytical solution

To solve the adaptive information dynamics model, we notice that the information entropy alters with the observed user preferences, as captured by the stochastic dynamics equation (2), leading to the corresponding Fokker–Planck equation for the observed user preference distribution $P(u_l^{(j)}|t)$. The stationary solution of the Fokker–Planck equation can be found by using the mean-field approximation. Therefore, it allows us to derive $P(\tilde{s})$ from the observed user preference distribution. As mentioned before, we adopt $P(\tilde{s}) \equiv P(\frac{s}{s^*})$ to quantify the degree of ICs among users, with $s^* = -\sum_{j=1}^{N} x_l^{(j)} \log x_l^{(j)}$ quantifying the diversity of inherent preferences of user l.

We define the probability density function $P(u_l^{(j)}|t)$, measuring the probability that the recommendation algorithm observes the user l's preference $u_l^{(j)}$ on topic j at time t. Solving equation (2) leads to the following Fokker–Planck equation:

$$\begin{aligned}&\frac{\partial P\left(u_l^{(j)}|t\right)}{\partial t}\\&= -\frac{\partial P\left(u_l^{(j)}|t\right)\left[\gamma_+ \sum_{\mathbf{i}_k \in R_\beta^+} F\left(u_l^{(j)}, i_k^{(j)}\right) + \gamma_- \sum_{\mathbf{i}_k \in R_\beta^-} F\left(u_l^{(j)}, i_k^{(j)}\right)\right]}{\partial u_l^{(j)}},\\&+\frac{1}{2}\frac{\partial^2 \sigma\left(u_l^{(j)}\right)^2 P\left(u_l^{(j)}|t\right)}{\partial\left(u_l^{(j)}\right)^2}\end{aligned} \tag{4}$$

where $P(u_l^{(j)}|t)$ is the calculated probability of the recommendation algorithm finding observed preference $u_l^{(j)}$ on a certain topic j at time t for user l. Notice that the first term of the right-hand side of equation (4) captures the effect of the recommendation algorithm under the mechanisms of similarity-based recommendation (β), as well as positive and negative feedback ($\gamma_\pm$), whereas the second term depicts the self-exploration (σ) of users. This equation suggests that the adaptive information dynamics in the human–AI interaction system is driven by the combined effect of recommendation and self-exploration. Assuming that the system reaches its stationary state, we use the mean-field approximation (Supplementary Section 3.2.2), obtaining

$$\begin{aligned}&\int_0^1 \mathrm{d}i^{(j)} f_l^{(j)} \left[\gamma_+ \Pi(\mathbf{x}, \mathbf{i}) + \gamma_- \frac{1}{\Pi(\mathbf{x}, \mathbf{i})}\right] F(u^{(j)}, i^{(j)}) P(u^{(j)}) P(i^{(j)})\\&= \frac{\mathrm{d}}{\mathrm{d}u^{(j)}}\left[\sigma(u^{(j)})^2 P(u^{(j)})\right]\end{aligned}, \tag{5}$$

where $f_l^{(j)} \approx \frac{1+\beta u_l^{(j)}}{N+\beta}$ and $\Pi(\mathbf{u}, \mathbf{x}, \mathbf{i}) = \frac{\pi(\mathbf{x},\mathbf{i})}{1-\pi(\mathbf{x},\mathbf{i})}$.

Incorporating the stationary $P(u^{(j)})$ obtained from the above equation and the rescaled parameters $\tilde{N} = N + \beta$ with equation (1), the distribution of the recommendation probability $P(f(u^{(j)})) = \frac{\mathrm{d}}{\mathrm{d}f^{(j)}} \int_0^{\frac{\tilde{N}f^{(j)}-1}{\beta}} P(u^{(j)})\,\mathrm{d}u^{(j)}$. By defining a random variable $s^{(j)} = f(u^{(j)}) - f(u^{(j)})^2$ satisfying $P(s^{(j)}) = \frac{\mathrm{d}}{\mathrm{d}s^{(j)}}\left(1 - \int_{1/\tilde{N}}^{g(s)} P(f^{(j)})\,\mathrm{d}f^{(j)}\right)$, where $g(s) = \frac{1}{2} + \sqrt{\frac{1}{4} - s^{(j)}}$, we obtain the information entropy distribution from the convolution of $P(s^{(j)})$:

$$P(s) = P_s = P(s^{(1)}) * P(s^{(2)}) * \cdots * P(s^{(j)}) * \cdots * P(s^{(N)}), \tag{6}$$

where $s = \sum_{j=1}^{N} s^{(j)}$. Similarly, we find the relative information entropy distribution as follows:

$$P(\tilde{s}) = \frac{\mathrm{d}}{\mathrm{d}\tilde{s}} \int_0^{\ln N} \left(\int_0^{\tilde{s}s^*} P(s)\,\mathrm{d}s\right) P(s^*)\,\mathrm{d}s^*, \tag{7}$$

where $P(s^*)$ is an arbitrary probability density function representing the inherent entropy distribution for the intrinsic preferences of users.

Minimal model

Below we will focus on the minimal realization of the proposed model. To be specific, we use a linear model $F(\mathbf{u}_l, \mathbf{i}_k) = \mathbf{i}_k - \mathbf{u}_l$ to account for the distance between the observed preference $\mathbf{u}_l$ and item feature $\mathbf{i}_k$ in the embedding space[13,37,38]. Besides the linear function, we have a lot of other options for the mathematical form of F, that is, $F \approx [-\mathbf{u}_l + \tanh(\mathbf{i}_k)]$[14]. However, all the candidate functions can be expanded as Taylor series $F \approx (\mathbf{i}_k - \mathbf{u}_l) + O((\mathbf{i}_k - \mathbf{u}_l)^2)$ when the high-order term is relatively weak. Moreover, we assume that user l accepts the recommended item k with probability $\pi(\mathbf{x}_l, \mathbf{i}_k) = \theta(\mathbf{x}_l, \mathbf{i}_k) = \mathbf{x}_l^{\mathrm{T}}\mathbf{i}_k$. We measure the similarity between $\mathbf{u}_l$ and $\mathbf{i}_k$, formulated as $\theta(\mathbf{u}_l, \Phi\mathbf{i}_k) = \mathbf{u}_l^{\mathrm{T}}\Phi\mathbf{i}_k$. Here we adopt the inner product as the similarity function, which can also be replaced with other functions (Supplementary Section 4.3).

Together with equations (5), (6) and (7), we can solve $P(\tilde{s})$ for the minimal model (Supplementary Section 3.2.3), finding three different states for ICs shown in Figs. 2 and 3. These three states are quantitatively defined by the following equations:

$$\begin{cases} P''(\tilde{s}) < 0, \langle\tilde{s}\rangle \geq 1 & \text{(diversification)}\\ -\infty < P''(\tilde{s}) < \infty, 0.5 \leq \langle\tilde{s}\rangle \leq 1 & \text{(partial IC)}\\ P''(\tilde{s}) < 0, \langle\tilde{s}\rangle < 0.5 & \text{(deep IC).}\end{cases} \tag{8}$$

Incorporating equation (7) with equation (8) allows us to find the transition lines from diversification to partial ICs, and then from partial ICs to deep ICs. When the state of the system transits from diversification to partial ICs, we have $P''(s) = 0$ and $\langle s\rangle \approx 1$. Similarly, the transition from partial ICs to deep ICs is characterized by $P''(s) = 0$ and $\langle s\rangle \approx 0.5$. The predictions of the transition lines are shown in Fig. 4. The red lines represent the transition lines from diversification to partial ICs, whereas the blue lines separate the domains corresponding to partial ICs and deep ICs.

We further conduct simulations for the minimal realization, where each user l has an intrinsic preference distribution $\mathbf{x}_l$ and an observed preference distribution $\mathbf{u}_l$ over N different topics. We initialize these vectors from two empirical datasets, including one collected from a

leading short-video platform and the other from a worldwide news platform (Supplementary Sections 1 and 3.3.2). In the video dataset, there are $N = 20$ distinct topics, including games, sports, food and so on; in the news dataset, there are $N = 14$ distinct topics, including sports, finance, lifestyle and so on. Following previous work[11], we assume that users' intrinsic preference distributions are drawn from a Dirichlet distribution $\mathbf{x}_l \approx \text{Dirichlet}(\boldsymbol{\mu}_{\text{user}})$, where $\boldsymbol{\mu}_{\text{user}}$ is a vector of the global popularity of topics obtained from empirical data (Supplementary Section 3.3.2 and Supplementary Fig. 7). Further, to avoid biasing the observed preferences very far from the global popularity, we initialize users' observed preferences using the identical Dirichlet distribution $\mathbf{u}_l \approx \text{Dirichlet}(\boldsymbol{\mu}_{\text{user}})$. Each item k has a fixed feature vector $\mathbf{i}_k = [i_k^{(1)}, i_k^{(2)}, \ldots, i_k^{(N)}]$, which is a one-hot vector that encodes the topic to which the item belongs. We randomly sample candidate items from the overall pool of items in each dataset. Distributions of the number of items over topics are shown in Supplementary Fig. 2.

At each time step t, we repeat interactions between humans and the recommendation algorithm in the following way.

(1) The recommendation algorithm recommends for each user l a set of items R_β with N_{rec} distinct items. The probability of each item being recommended follows equation (3).
(2) User l gives positive feedback to the recommended item k with a probability of $\pi(\mathbf{x}_l, \mathbf{i}_k)$ and gives negative feedback with a probability of $1 - \pi(\mathbf{x}_l, \mathbf{i}_k)$. Items receiving positive feedback make up a set of positive feedback samples R_β^+ and the other items make up a set of negative feedback samples R_β^-.
(3) User l carries out random self-exploration following a Wiener process.
(4) The recommendation algorithm updates each user l's observed preference following equation (2).

To delineate the severity of ICs, we evaluate the relative information entropy $\tilde{s}$ and degrees of over-recommendation and under-recommendation r_o, r_u for each user l at each time step. Specifically, for $\tilde{s}$, we compute the entropy of the topic distribution in R_β and divide it by the intrinsic diversity of user preference s^*. For r_o, r_u, we compute the degree of over-recommendation as $r_o = \sum_{j \in \mathbf{F}_o} |f_l^{(j)} - x_l^{(j)}| / x_l^{(j)}$, where $\mathbf{F}_o = \left\{ j : x_l^{(j)} > 0, f_l^{(j)} > x_l^{(j)} \right\}$, and the degree of under-recommendation as $r_u = \sum_{j \in \mathbf{F}_u} |f_l^{(j)} - x_l^{(j)}| / x_l^{(j)}$, where $\mathbf{F}_u = \{ j : x_l^{(j)} > 0, f_l^{(j)} < x_l^{(j)} \}$.

Data availability

The news dataset[5] is available at https://msnews.github.io/. For commercial reasons, we anonymize the specific name of the video platform. We present the video dataset at https://github.com/tsinghua-fib-lab/Adaptive-Information-Dynamic-Model (refs. 39,40). In the GitHub repository, we provide the behavioural data aggregated to individual granularity and the processed data for Figs. 1–4. Source data are provided with this paper.

Code availability

The code used in this research is available at https://github.com/tsinghua-fib-lab/Adaptive-Information-Dynamic-Model (refs. 39,40).

References

1. Tagliabue, J. et al. A challenge for rounded evaluation of recommender systems. *Nat. Mach. Intell.* **5**, 181–182 (2023).
2. Ricci, F., Rokach, L. & Shapira, B. *Recommender Systems Handbook* (Springer, 2022).
3. Zhang, S., Yao, L., Sun, A. & Tay, Y. Deep learning based recommender system: a survey and new perspectives. *ACM Comput. Surv.* **52**, 1–38 (2019).
4. Bakshy, E., Messing, S. & Adamic, L. A. Exposure to ideologically diverse news and opinion on Facebook. *Science* **348**, 1130–1132 (2015).
5. Wu, F. et al. Mind: a large-scale dataset for news recommendation. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics* (eds Jurafsky, D. et al.) 3597–3606 (Association for Computational Linguistics, 2020).
6. Covington, P., Adams, J. & Sargin, E. Deep neural networks for YouTube recommendations. In *RecSys '16: 10th ACM Conference on Recommender Systems* 191–198 (Association for Computing Machinery, 2016).
7. Davidson, J. et al. The YouTube video recommendation system. In *Proc. Fourth ACM Conference on Recommender Systems, RecSys '10* 293–296 (Association for Computing Machinery, 2010).
8. Santos, F. P., Lelkes, Y. & Levin, S. A. Link recommendation algorithms and dynamics of polarization in online social networks. *Proc. Natl Acad. Sci. USA* **118**, e2102141118 (2021).
9. Sunstein, C. R. *Infotopia: How Many Minds Produce Knowledge* (Oxford University Press, 2006).
10. Nguyen, T. T., Hui, P.-M., Harper, F. M., Terveen, L. & Konstan, J. A. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *WWW '14 Companion: Proc. 23rd International Conference on World Wide Web* 677–686 (Association for Computing Machinery, 2014).
11. Chaney, A. J. B., Stewart, B. M. & Engelhardt, B. E. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proc. 12th ACM Conference on Recommender Systems, RecSys '18* 224–232 (Association for Computing Machinery, 2018).
12. Algorithmic recommendations, anyone? *Nat. Mach. Intell.* **5**, 95 (2023).
13. Liu, J., Huang, S., Aden, N. M., Johnson, N. F. & Song, C. Emergence of polarization in coevolving networks. *Phys. Rev. Lett.* **130**, 037401 (2023).
14. Baumann, F., Lorenz-Spreen, P., Sokolov, I. M. & Starnini, M. Modeling echo chambers and polarization dynamics in social networks. *Phys. Rev. Lett.* **124**, 048301 (2020).
15. Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W. & Starnini, M. The echo chamber effect on social media. *Proc. Natl Acad. Sci. USA* **118**, e2023301118 (2021).
16. Leonard, D. & Sensiper, S. The role of tacit knowledge in group innovation. *Calif. Manag. Rev.* **40**, 112–132 (1998).
17. Munson, S. A. & Resnick, P. Presenting diverse political opinions: how and how much. In *Proc. SIGCHI Conference on Human Factors in Computing Systems* 1457–1466 (Association for Computing Machinery, 2010).
18. Garimella, K., De Francisci Morales, G., Gionis, A. & Mathioudakis, M. Political discourse on social media: echo chambers, gatekeepers, and the price of bipartisanship. In *WWW '18: Proc. 2018 World Wide Web Conference* 913–922 (International World Wide Web Conferences Steering Committee, 2018).
19. Schmidt, A. L. et al. Anatomy of news consumption on Facebook. *Proc. Natl Acad. Sci. USA* **114**, 3035–3039 (2017).
20. Kitchens, B., Johnson, S. L. & Gray, P. Understanding echo chambers and filter bubbles: the impact of social media on diversification and partisan shifts in news consumption. *MIS Q.* **44**, 1619–1649 (2020).
21. Kalimeris, D., Bhagat, S., Kalyanaraman, S. & Weinsberg, U. Preference amplification in recommender systems. In *Proc. 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* 805–815 (Association for Computing Machinery, 2021).
22. Korbel, J., Lindner, S. D., Pham, T. M., Hanel, R. & Thurner, S. Homophily-based social group formation in a spin glass self-assembly framework. *Phys. Rev. Lett.* **130**, 057401 (2023).
23. Lorenz-Spreen, P., Oswald, L., Lewandowsky, S. & Hertwig, R. A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nat. Hum. Behav.* **7**, 74–101 (2023).

24. Flamino, J. et al. Political polarization of news media and influencers on Twitter in the 2016 and 2020 US presidential elections. *Nat. Hum. Behav.* **7**, 904–916 (2023).
25. Levy, R. Social media, news consumption, and polarization: evidence from a field experiment. *Am. Econ. Rev.* **111**, 831–870 (2021).
26. Bail, C. A. et al. Exposure to opposing views on social media can increase political polarization. *Proc. Natl Acad. Sci. USA* **115**, 9216–9221 (2018).
27. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
28. Castelvecchi, D. Can we open the black box of AI?. *Nature* **538**, 20–23 (2016).
29. Kunaver, M. & Požrl, T. Diversity in recommender systems—a survey. *Knowl.-Based Syst.* **123**, 154–162 (2017).
30. Liu, P., Shivaram, K., Culotta, A., Shapiro, M. A. & Bilgic, M. The interaction between political typology and filter bubbles in news recommendation algorithms. In *WWW '21: Proc. Web Conference 2021* 3791–3801 (Association for Computing Machinery, 2021).
31. Rendle, S., Freudenthaler, C., Gantner, Z. & Schmidt-Thieme, L. BPR: Bayesian personalized ranking from implicit feedback. In *UAI '09: Proc. 25th Conference on Uncertainty in Artificial Intelligence* 452–461 (AUAI Press, 2009).
32. Ding, J., Quan, Y., He, X., Li, Y. & Jin, D. Reinforced negative sampling for recommendation with exposure data. In *Proc. 28th International Joint Conference on Artificial Intelligence, IJCAI-19* (ed. Kraus, S.) 2230–2236 (International Joint Conferences on Artificial Intelligence, 2019).
33. Su, X. & Khoshgoftaar, T. M. A survey of collaborative filtering techniques. *Adv. Artif. Intell.* **2009**, 421425 (2009).
34. Kobayashi, M. & Takeda, K. Information retrieval on the web. *ACM Comput. Surv.* **32**, 144–173 (2000).
35. König, M. D., Levchenko, A., Rogers, T. & Zilibotti, F. Aggregate fluctuations in adaptive production networks. *Proc. Natl Acad. Sci. USA* **119**, e2203730119 (2022).
36. Itô, K. *On Stochastic Differential Equations* (American Mathematical Society, 1951).
37. Clifford, P. & Sudbury, A. A model for spatial conflict. *Biometrika* **60**, 581–588 (1973).
38. Holley, R. A. & Liggett, T. M. Ergodic theorems for weakly interacting infinite systems and the voter model. *Ann. Probab.* **3**, 643–663 (1975).
39. Piao, J. et al. Open code for in-review natmachintell-a23038004 "Human–AI adaptive dynamics drive emergence of information cocoons". *Code Ocean* https://doi.org/10.24433/CO.6503936.v1 (2023).
40. Piao, J. et al. tsinghua-fib-lab/Adaptive-Information-Dynamic-Model: NMI. *Zenodo* https://doi.org/10.5281/zenodo.8265474 (2023).

Acknowledgements

We thank J. Ding, Z. Chen and C. Song for discussions and comments on the manuscript. This work was supported in part by the National Key Research and Development Program of China under grant 2020AAA0106000 to Y.L., the National Natural Science Foundation of China under grants 72104126 to F.Z., 71721002 to J.S., U1936217 and U22B2057 to Y.L. The funders had no role in study design, data collection, data analysis, decision to publish, or preparation of the manuscript.

Author contributions

J.P., J.L. and Y.L. designed the model. J.P. performed the experiments and prepared the figures. J.L. conducted the theoretical analysis. F.Z., J.S. and Y.L. provided critical revisions. All authors jointly participated in the writing of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-023-00731-4.

Correspondence and requests for materials should be addressed to Yong Li.

Peer review information *Nature Machine Intelligence* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Liesbeth Venema, in collaboration with the *Nature Machine Intelligence* team.