JASSS
THE JOURNAL OF ARTIFICIAL SOCIETIES AND SOCIAL SIMULATION

On the Interplay of Gullibility, Plausibility, and Criticism: A Computational Model of Epistemic Vigilance

Daniel Reisinger[1], **Marie L. Kogler**[1], **Georg Jäger**[1]

[1] *University of Graz, Institute of Environmental Systems Sciences, Merangasse 18, Graz, 8010, Austria*
Correspondence should be addressed to *daniel.reisinger@uni-graz.at*

Journal of Artificial Societies and Social Simulation 26(3) 8, 2023
Doi: 10.18564/jasss.5136 Url: http://jasss.soc.surrey.ac.uk/26/3/8.html

Received: 07-10-2021 Accepted: 21-04-2023 Published: 30-06-2023

Abstract: Humans heavily depend on communication. We constantly share new ideas, catch up on current news, and exchange gossip. Much of the information conveyed in this way is, however, not first-hand. As a result, we run the risk of being misinformed and of spreading potentially harmful messages via large social networks. Current research argues that we are endowed with a set of cognitive mechanisms capable of targeting such risks. These mechanisms, known as mechanisms of epistemic vigilance, help us evaluate communicated information by i) critically evaluating presented arguments, ii) checking the plausibility of messages against pre-existing background beliefs, and iii) assessing the competence of a sender based on cues of trustworthiness. So far, the mechanisms exist only as verbal theory, which do not allow a thorough systemic analysis of the interplay between them. In this paper, we implement an agent-based computational model of epistemic vigilance to add to the existing microscopic (individual level) and macroscopic (societal level) understanding of the mechanisms. Through simulations of different multi-agent societies we are able to show that the mechanisms of epistemic vigilance are sufficient to explain a wide variety of phenomena: (a) The locality of critics in social groups is a deciding factor when it comes to quickly correcting false messages. (b) Plausibility checking can create impeding group structures that exclude other agents from receiving surrounding information. (c) Impeding group structures can be overcome through competence checking. (d) And on a societal level, increasing the proportion of agents performing plausibility checks, creates an abrupt shift from consensus to polarization.

Keywords: Computational Modeling, Epistemic Vigilance, Rumor Diffusion, Polarization, Micro-Macro Link

Introduction

1.1 When we observe the amount of dubious information we share with each other, one often cannot help but think: We must be pretty gullible! Fake news, rumors, and gossip are part of our daily communication (Lazer et al. 2018; Crescimbene et al. 2012; Oh et al. 2013). On many occasions, we seem to blindly accept new information and distribute it to our friends and colleagues. We share rumors without questioning their content or source, and we repeat after opinion leaders on topics we know little to nothing about. But is our judgment really that poor? Are we too easily swayed by what other people tell us? Or in other words, are we too gullible (Mercier 2017)?

1.2 Current research in social psychology argues that this is not the case and that we do not gullibly accept whatever we are told (Sperber et al. 2010; Mercier 2017, 2020; Petersen 2020). On the contrary, Mercier (2020) states that we are quite "skilled at figuring out who to trust and what to believe", and that we are, if anything, "too hard rather than too easy to influence". Through evidence from experimental psychology, it is revealed that we are endowed with a set of well-functioning cognitive mechanisms that help us evaluate communicated information (Mercier 2017). These mechanisms have been termed mechanisms of *epistemic vigilance* (Mascaro

& Sperber 2009; Sperber et al. 2010). *Epistemic*, because they deal with how we acquire new knowledge and *vigilance*, because they make us more selective in what we accept. The mechanisms of epistemic vigilance encompass several functions that help us filter out misinformation from communicated contents (Mascaro & Sperber 2009). They include critically evaluating presented arguments, checking the plausibility of messages against pre-existing background beliefs, and assessing the competence of senders based on cues of trustworthiness (Mercier 2017). The combination of all of these mechanisms makes us vigilant towards communicated information. Still, the degree of vigilance is somewhat situational. While the mechanisms filter out misinformation in many social contexts (Mercier 2017), they also allow the occasional rumor to spread widely into social networks. Some case examples include the spread of false rumors after the East Japan Earthquake (Takayasu et al. 2015), misinformative tweets during the Boston Marathon bombing (Lee et al. 2015), and political rumoring during US elections (Shin et al. 2017). The mechanisms are also no guarantee to avoid phenomena like rumor clustering, bubble effects, and the emergence of polarization patterns (DiFonzo & Bordia 2007). To navigate such dynamics, it is therefore particularly important to understand how the proposed mechanisms work systemically.

1.3 So far, the mechanisms of epistemic vigilance as outlined by Mercier (2017) exist only as verbal theory. Although the mechanisms have been generalized to a degree where they can assess a variety of social contexts such as religion (Boyer 2008, 2021), demagoguery (Worsley 1957; Selb & Munzert 2018; Kalla & Broockman 2018), mass media (Herman & Chomsky 2010), and rumors (DiFonzo & Bordia 2007), among others (Sperber et al. 2010; Mercier 2017, 2020; Vasilyeva et al. 2021), they are lacking formality. And without a formal representation, a thorough systemic analysis is not possible, leaving many questions unanswered. For example: How do the mechanisms interact with each other? Are they showing some surprising feedback loops? And under what conditions do they fail or succeed in containing the spread of false information? Increasing the degree of formalization by converting the existing verbal theory into a computational model (Smaldino 2017; van Rooij & Blokpoel 2020) can help to find answers to the above questions and contribute to the microscopic (individual level) and macroscopic (societal level) understanding of the mechanisms.

1.4 In this paper, we propose one formal representation in the form of an agent-based computational model. This type of model is especially suitable for the formalization task, as it allows us to equip every agent with a rule-based version of the underlying verbal theory, i.e. the mechanisms of epistemic vigilance. It also allows us to construct and test various scenarios that would otherwise be difficult to find in empirical research. It is therefore not surprising that this methodological approach has been shaped and supported by many researches in the social sciences over the past few years (Epstein 1999; Gilbert & Terna 2000; Helbing 2012; Squazzoni et al. 2014; Foster 2018).

1.5 It is worth noting that there exist alternative approaches to modeling epistemic vigilance, in particular models of source reliability or Bayesian source credibility models (Bovens & Hartmann 2004; Olsson 2011; Merdes et al. 2021). These models assume that credence is plausibly modeled by probabilities and Bayesian conditionalization in which individuals learn the true likelihoods of a message generation process (Merdes et al. 2021). Models of source reliability or Bayesian source credibility models are leaning towards testing the corrective performance of individuals equipped with a formal variant of Bayesian expectation-based revision (Merdes et al. 2021). Our approach does not include any form of Bayesian revision, or learning for that matter. Instead, it concentrates on continuing the analysis of Mercier (2017) by examining the macro implications of individuals equipped with a proposed set of almost 'fast and frugal'-like mechanisms of epistemic vigilance.

1.6 Another group of models closely related to this work are social influence models which belong to the broader field of opinion dynamics. These models operate on the basis of assimilative and repulsive forces. According to Flache et al. (2017), social influence models can be categorized into three classes: Models of assimilative social influence which aim to explain the phenomenon of global consensus, models with similarity biased influence which aim to explain the phenomenon of opinion clustering, and models with repulsive influence which aim to explain the phenomenon of bi-polarization (Flache et al. 2017). These models benefit from a strong formality in their descriptions allowing rigorous mathematical analysis of convergence and stability of solutions. The cost is often that mechanism design must adhere to the formal structure of opinion dynamics. This makes a translation from 'fast and frugal'-like mechanisms of epistemic vigilance more difficult but not impossible. Recent work shows how abstract theory as presented in Mercier (2017) may be combined with bounded-confidence social influence to produce an opinion dynamics model of epistemic vigilance (Butler et al. 2020).

1.7 The remainder of this paper is organized as follows: First, we develop an agent-based computational model of epistemic vigilance, focusing on the implementation of interaction mechanisms and interaction networks. Second, we look at small agent populations in highly stylized agent formations to better understanding the microscopic interplay of the formalized mechanisms. Third, we look at more complex agent formations consisting

of large groups of Schelling segregated agents and examine macroscopic patterns of rumor diffusion. Lastly, we discuss our results, highlight any model limitations, and provide an outlook for future research in this direction.

Model

2.1 Our approach is a computational model where rumor spreading dynamics are analyzed by implementation of behavior responses and interaction rules on the individual level of the agents. Entire social systems are modeled as a collection of autonomous agents that interact with each other based on a predefined set of interaction mechanisms (Bonabeau 2002). The networked version of our agent-based computational model consists of the following core elements: A set of agents including agent attributes and interaction mechanisms, and an underlying topology of connectedness, in our case, an interaction network that describes with whom agents can interact (Macal & North 2005).

Agent

2.2 Agents are capable of evaluating communicated information based on the following three mechanisms of epistemic vigilance: critical evaluation, plausibility checking, and competence checking (Mercier 2017). An overview of our implementation is provided in Figure 1.

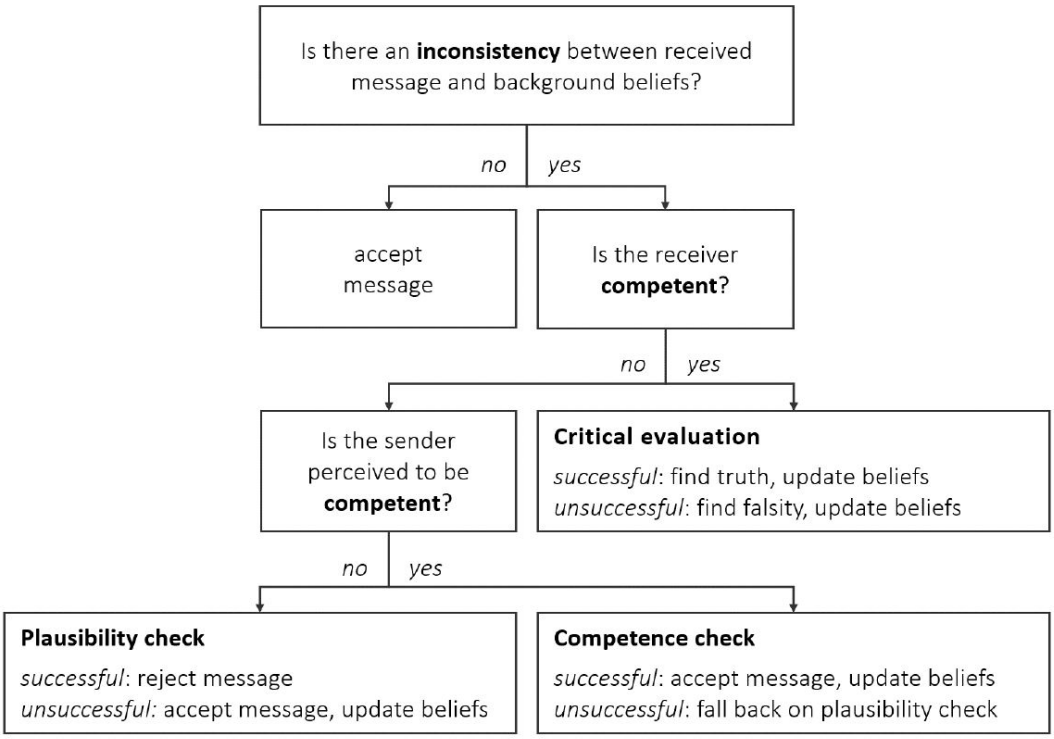

Figure 1: Schematic overview of the implementation of the mechanisms of epistemic vigilance. The structuring of the mechanisms is based on the findings of Mercier (2017).

2.3 The mechanisms determine how agents react when exchanging information with each other. Depending on the social context, they allow agents to be either open or vigilant towards communicated information. For example, an agent may decide to accept or reject a new piece of information based on the fit between the content of a received message and their own background beliefs in that subject. They may be more or less critical to the content of a message based on how competent they are in that subject. And they may factor the competence of the sender into their decision to accept or reject a message. Factors such as content, background beliefs, and competence are part of the social context between agents. They are important inputs to the mechanisms of epistemic vigilance and can be accessed through agent attributes. We have defined the following set of attributes where each agent holds: a message, background beliefs, and a competence attribute (Figure 2).

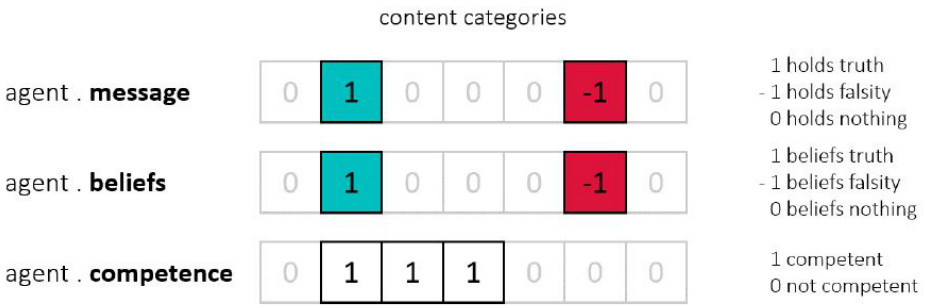

Figure 2: Overview of agent attributes. The figure shows an informational environment with seven content categories and one possible realization for each attribute. The presented agent has a message containing a single true statement and a single false statement, they hold corresponding true and false background beliefs and the agent is competent in three content categories.

2.4 First, every agent has a message vector that allows them to hold information. Messages consist of a fixed number of content categories or subjects. This number represents the range of the informational environment of agents, or in other words, what they are allowed to speak about. For example, we could imagine an artificial world in which agents are only allowed to exchange information about three topics: politics, medicine, and sports. Such a world would then consist of message vectors with exactly three content categories. For each content category, an agent may or may not hold some actual information. 0 indicates that the agent does not hold any information, 1 indicates that the agent holds some true information, and -1 indicates that the agent holds some false information. Agents do not know whether they hold truths or falsities, i.e., they only store content. The content of a message is subject to change where new content can be added and existing content can be updated or deleted. When and how often messages change depends on the outcome of agent interactions.

2.5 Second, every agent has a set of background beliefs. Background beliefs can be understood as information that has already been internalized. They represent an agent's established knowledge in a certain content category. 0 indicates that the agent does not hold any background beliefs, 1 indicates that the agent holds true background beliefs, and -1 indicates that the agent holds false background beliefs. For example, an agent may already believe in the falsity that vaccination is a cause of autism, and is therefore assigned -1 in the content category medicine. Agents with true background beliefs about vaccination are assigned 1 in that content category. Background beliefs are later used to determine whether an agent faces inconsistencies between message and beliefs.

2.6 Third, every agent has a competence attribute. While the informational environment may consist of a number of content categories, certainly not every agent is competent in all of them. For our purposes, 0 indicates that an agent is not competent in a particular content category and 1 indicates that an agent is competent in that content category. As will be explained later, competence is used to determine whether an agent will critically evaluate information or perform a simpler plausibility check instead. A sender's competence is also used as a cue to trigger a competence check potentially overruling the result of a plausibility check.

Base mechanism and critical evaluation

2.7 The base mechanism is intended as a starting point for understanding how communication between two agents is modeled. It builds on the theory of cognitive dissonance where inconsistencies between actions, such as accepting a piece of information, and currently held beliefs need to be resolved by either a change in action or a change in beliefs (Festinger 1957). According to Mercier (2017) and Sperber et al. (2010), one important strategy to resolve such an inconsistency is to perform reasoning which constitutes the ability to find and critically evaluate reasons obtained either privately or publicly, where it has been shown that individuals are quite capable of distinguishing between weak and strong arguments (Petty et al. 1997; Castelain et al. 2016). Our implementation of the base mechanisms involves two agent attributes: the sender's message and the receiver's background beliefs (Figure 3).

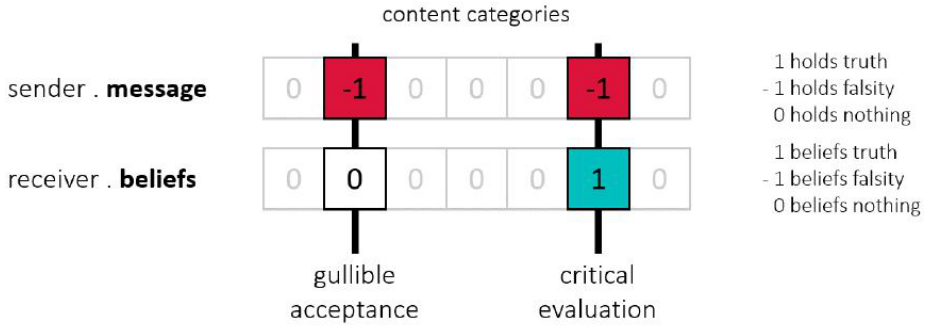

Figure 3: Implementation of base mechanism. Agents without any background beliefs will gullibly (openly) accept any piece of new information, whether it is true or false. Agents that face an inconsistency between background beliefs and message perform a critical evaluation.

2.8 Depending on the fit between message and beliefs, the receiver may either accept or critically evaluate the message: If the receiver has no prior background beliefs they will gullibly (openly) accept the message. If the receiver has matching background beliefs, they will also accept the message (this is however not considered gullible acceptance). If the receivers has conflicting background beliefs, they will do a critical evaluation before accepting any information. In the base mechanism, any inconsistencies between message and beliefs act as triggers for a critical evaluation.

2.9 Taking an example from the verbal model, a sender might want to communicate the false rumor that 9/11 was an inside job (a falsity indicated with -1 in the content category politics). The receiver, who believes otherwise, is now prompted to perform a critical evaluation. A successful critical evaluation results in the agent finding the truth and updating their message and background beliefs accordingly. Perhaps the receiver did some research online and confirmed their suspicion that 9/11 being an inside job is likely false information, and instead found convincing information that it was coordinated by an Islamist terrorist group. They then update their message and background beliefs to reflect the newly researched information. Likewise, an unsuccessful critical evaluation results in the agent finding convincing arguments for the false rumor which causes them to accept the falsity and update their background beliefs in accordance with it.

2.10 The critical evaluation is modeled as a probability event with a certain success rate, where a high success rate means that it will be easier for agents to find the truth, and a low success rate means that it will be more difficult for them. The success rate parameter could therefore in part reflect the ratio of available true to false information on a topic on a given research platform such as the internet, and in part the agent's ability to correctly interpret the available information.

2.11 One remark on updating background beliefs: It can be argued that background beliefs acquired through processes such as critical evaluations are likely to be held intuitively (strong) as opposed to reflectively (weak) (Sperber 1997; Mercier 2017). That is why we update the agent's background beliefs after a critical evaluation, but not after gullible acceptance. For example, a message such as 9/11 being an inside job might be gullibly accepted in the form of "Simon told me that 9/11 was an inside job". Since the content of the message is embedded in a propositional attitude, it is considered a reflectively believed message (Sperber 1997; Mercier 2017). The message may still be further communicated but it is only believed reflectively and we do not update the agent's background beliefs. In contrast, the message may also be accepted in the form "9/11 was an inside job" which sounds more like a fact and is likely believed intuitively. Intuitive acceptance (possibly caused by a failed critical evaluation) will therefore result in updated background beliefs.

Plausibility checking

2.12 Plausibility checking is the process of detecting inconsistencies between received information and currently held background beliefs, and rejecting information when such inconsistencies are present (Mercier 2017). Plausibility checking makes agents more vigilant towards information that goes against their background beliefs, and thus enables a social group to more easily retain existing knowledge. In this regard, it should be emphasized that there is a significant difference in how we deal with information communicated to us and information obtained by ourselves. In the presence of an inconsistency between background beliefs and information communicated to us, plausibility checking should, on average, lead to the rejection of that information, because of

an uncertainty regarding the sender's honesty (Mercier 2017; Bonaccio & Dalal 2006; Yaniv 2004). Contrary to this, inconsistencies between background beliefs and information obtained by ourselves, e.g., by witnessing an event, are more likely to be resolved by belief updating due to our perceptual and inferential mechanisms being honest, i.e., not designed to deceive us (Mercier 2017). Our implementation of plausibility checking involves three agent attributes: the sender's message, the receiver's background beliefs, and the receiver's competence (Figure 4).

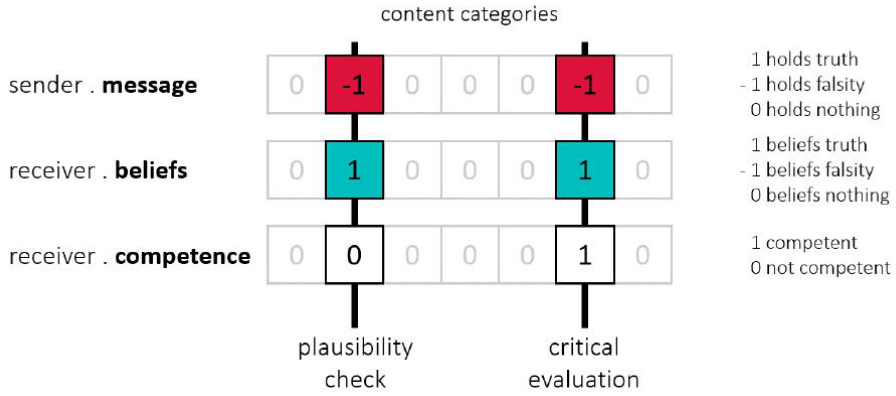

Figure 4: Implementation of base mechanism + plausibility checking. Agents facing an inconsistency between received message and background beliefs either perform a quick plausibility check (in the absence of own competence in that content category) or a thorough critical evaluation (in the presence of own competence in that content category).

2.13 Depending on the receiver's competence, agents facing an inconsistency between message and background beliefs can perform a plausibility check. We assume that agents who are not particularly competent in a subject will just check the plausibility of a statement, whereas agents that are competent in a subject will perform a thorough critical evaluation. In our implementation, plausibility checking is triggered when an agent faces an inconsistency but is not competent in that content category. The plausibility check is modeled as a probability event with a certain success rate. A successful plausibility check results in the agent rejecting inconsistent information and leaving previously held background beliefs unchanged. The agent's inconsistency is resolved by rejecting the message that caused it. A failed plausibility check results in the agent accepting the information and updating their background beliefs. Following Festinger's principle of cognitive consistency, our argument for updating beliefs here is that a failed plausibility check implies that the agent must have resolved the inconsistency not by getting rid of the information that caused it but by changing previously held beliefs (Festinger 1957). Note that in the event that an agent does not have any prior background beliefs, there is no reason to reject new information (Gilbert et al. 1990) and the agent gullible accepts without any prior plausibility check.

Competence checking

2.14 Competence checking is a mechanism that can supersede plausibility checking by allowing receivers to accept messages whose content is inconsistent with their background beliefs (Mercier 2017). Its purpose is to loosen the strict conditions of plausibility checking. Plausibility checking causes agents to reject most of the inconsistent information they receive. Only occasional failed plausibility checks allow inconsistent information to be accepted. Competence checking addresses this issue by introducing a possibility to circumvent plausibility checking and, thus, ensures that agents have other ways of accepting inconsistent information. Competence checking involves four agent attributes: the sender's message, the receiver's background beliefs, the receiver's competence, and the sender's competence (Figure 5).

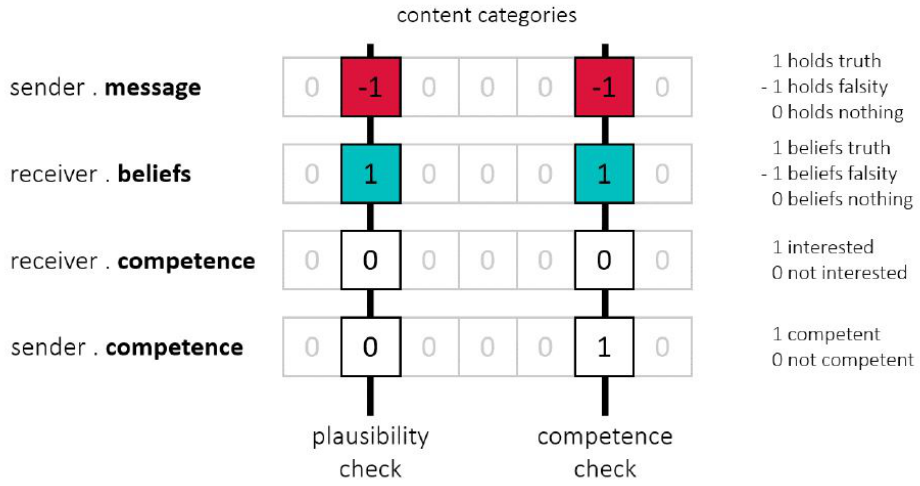

Figure 5: Implementation of base mechanism + plausibility checking + competence checking. Agents who face an inconsistency between received message and beliefs, and who are not competent in that content category, may supersede their plausibility check through competence checking.

2.15 In order for a receiver to be able to assess whether a sender is sending them true messages, they need informative cues. Ideally, agents would look into the background beliefs of others. A sender that beliefs mostly truths is most likely to speak truths. But this seems infeasible, since internal beliefs are somewhat hidden. In our model, an agent's competence attribute serves as such a peripheral cue, as competence is less hidden and more easy to communicate, for example, through credentials, education, job, and interests, among others (Cacioppo & Petty 1986). As established previously, agents facing inconsistencies between message and beliefs, and who are competent in a particular content category, are more likely to have performed successful critical evaluations. Because critical evaluations are ideally biased towards the truth (higher success rate of finding the truth), any agent that does a lot of them has likely formed true background beliefs. Competence may therefore serve as a good indirect measure for true background beliefs. To summarize, competence checking is triggered when the receiver of a message faces an inconsistency between message and background beliefs, they themselves are not particularly competent in that topic, but their sender is. Competence checking is modeled as a probability event with a certain success rate. A successful competence check results in the receiver accepting the message and updating their background beliefs, superseding the plausibility check in the process. An unsuccessful competence check leads to them perceiving the sender as not competent enough. In this case, they will fall back on the plausibility check instead.

Interaction networks

2.16 Agents select their interaction partner based on a common social network. Different network structures provide different neighborhoods and different neighborhoods may influence the dynamics of information dissemination. We select the following interaction networks to test how sensitive our mechanisms are to various underlying network structures: a von Neumann neighborhood, a Moore neighborhood, a Voronoi neighborhood, a Watts-Strogatz small-world graph, and Barabasi-Albert scale-free graph (Figure 6).

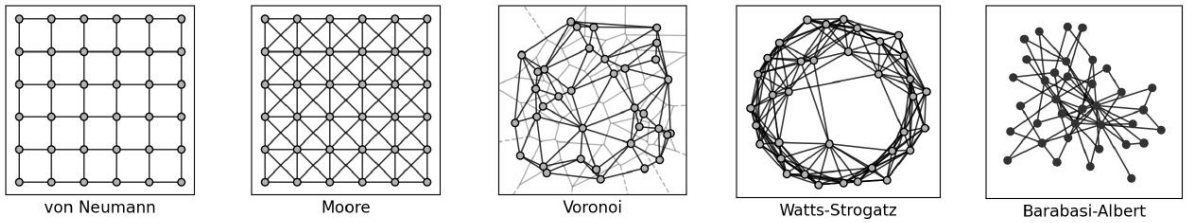

Figure 6: Interaction networks. The figure shows example graphs of 36 nodes for a von Neumann, a Moore, a Voronoi neighborhood, a Watts-Strogatz graph with average degree 10 and rewiring probability p equal to 0.01, and a Barabasi-Albert scale-free graph with the number of edges to attach from a new node to existing nodes m equal to 2.

2.17 Agents interacting in a von Neumann neighborhood are placed on a regular grid and connected to their four immediate neighbors (top, bottom, left, right). Similarly, agents interacting in a Moore neighborhood are placed on a regular grid, connected to their four von Neumann neighbors plus their immediate diagonal neighbors (top right, top left, bottom right, bottom left) forming eight neighbors in total. Both the von Neumann and the Moore network are implemented with closed boundary conditions, meaning that nodes at the borders have less than four (von Neumann) or eight (Moore) neighbors, respectively. The Voronoi neighborhood is intended as an irregular counterpart to the regular structures of the von Neumann and the Moore neighborhoods (Centola et al. 2005). Here, agents are placed randomly on a plane and the underlying Voronoi network is created using the Delaunay triangulation algorithm provided by SciPy (Jones et al. 2001). The Watts-Strogatz graph is used to generate more realistic social network structures in terms of average clustering coefficient C and average shortest path length L, with the premise that real social networks exhibit high clustering and low average shortest path lengths (small-world property) (Barabási & Pósfai 2016). In our case, such a network structure is realized using the Watts-Strogatz model from NetworkX (Hagberg et al. 2008) with average degree 10 and rewiring probability $p = 0.01$. This parameter setting yields network structures comparable to scientific collaboration networks in terms of C and L (Barabási & Pósfai 2016). Finally, a Barabasi-Albert graph is used to include networks with the scale-free property, i.e. a degree distribution that follows a power law.

Simulation Setup

3.1 The agent interaction process can be outlined as follows: Agents identify neighboring agents through their common interaction network. Every agent selects one immediate neighbor at random. The selected neighbor then performs the role of a message sender and the selecting agent performs the role of a listener or message receiver. If the sender agent happens to hold some actual information (a message containing at least one non-zero entry), then the two agents interact based on the implemented mechanisms of epistemic vigilance. This is done for all agents in the network (activated in a random sequence) and repeated for a specified number of ticks, where a tick is over when every agent had the opportunity to select an interaction partner.

3.2 All simulations are run in an informational environment with a single content category (messages, background beliefs, and interests are one-dimensional and only have one entry). This allows us to speak of different agent types: Agents without any background beliefs in a content category are called *gullibles* as they will always accept a message regardless of its content. Agents with background beliefs (true or false) and who are not competent in that content category are called *plausibles* as they will always want to do a plausibility check when facing an inconsistency between message and beliefs (can be superseded by competence checking). Agents with background beliefs (true or false) and who are competent in that content category are called *critics* as they will always want to do a critical evaluation when faced with an inconsistency. Agents with true background beliefs get the addition *truth biased* and agents with false background beliefs the addition *falsity biased*. For example, a truth biased critic has true background beliefs and is competent in the content category. Agents interacting in a one-dimensional informational environment do not have to perform multiple roles. Because some agent attributes cannot be altered by the defined mechanisms, the roles of agents are fixed. If we were to initialize a multi-dimensional informational environment, an agent could of course perform multiple roles, for example, that of a critic in the category they are competent in, that of a plausible in the category they are not competent in, and that of gullible where they do not have any background beliefs.

Illustration of Principles

4.1 To illustrate the dynamics produced by the mechanisms outlined above, we look at small agent populations (81 agents) in highly stylized agent formations. Dynamics for the below simulations are qualitatively confirmed for all interaction networks (von Neumann neighborhood, Moore neighborhood, Voronoi neighborhood, Watts-Strogatz graph, and Barabasi-Albert graph) by means of the Kendall's tau coefficient provided by SciPy (Jones et al. 2001). The Kendall's tau coefficient measures the correspondence between two rankings where values close to 1 indicate strong agreement and values close to -1 strong disagreement (Jones et al. 2001). It measures the ordinal similarity between two data series and thus provides a suitable measure of their qualitative similarity in peaks, drops and inclines.

On the locality of critics

4.2 Our understanding of locality consists of two components: connectedness and centrality. To assess the impact of either on the performance of critics we ran simulations with the following experimental design: We initialize a small agent population with 81 agents placed in one of the network structures. Every agent is equipped with the base mechanisms, meaning that they can gullibly accept, accept based on consistency between message and beliefs, and critically evaluate communicated information based on inconsistency between message and beliefs. To assess the impact of connectedness, the node centrality is fixed and the node degree is varied. More specifically, we placed a single perfect (success rate of critical evaluation equal to 1) truth biased (true background beliefs) critic on the most central node in the network and varied its node degree. Note that for the von Neumann and the Moore networks, the most central node is simply the node in the center of the grid. For the Voronoi diagram and the Watts-Strogatz graph it is identified using the closeness centrality measure provided by NetworkX (Hagberg et al. 2008). The perfect truth biased critic is surrounded by gullibles (agents without any background beliefs) of which one is randomly selected to hold a starting falsity (message containing -1). Example simulations for maximum node degree (8 links) and minimum node degree (1 link) of the critic on a Moore neighborhood are shown in Figures 7 and 8. Average dynamics of 1000 simulation runs are shown in Figures 9 and 10.

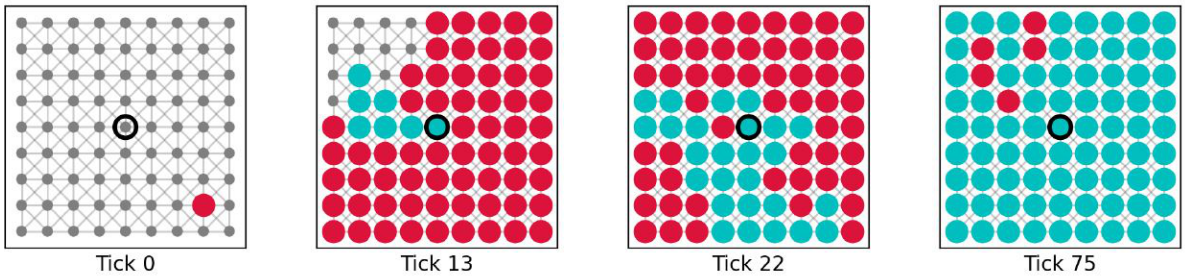

Figure 7: Example simulation of the base mechanism with a well-connected critic. Each node in the network represents an agent. The most central agent in the network performs the role of a truth biased critic (black outer circle) and is well-connected (8 links) to their Moore neighbors. The remaining agents perform the roles of gullibles. The initial falsity (red) is placed on a random gullible and can be corrected into a truth (cyan) by the critic.

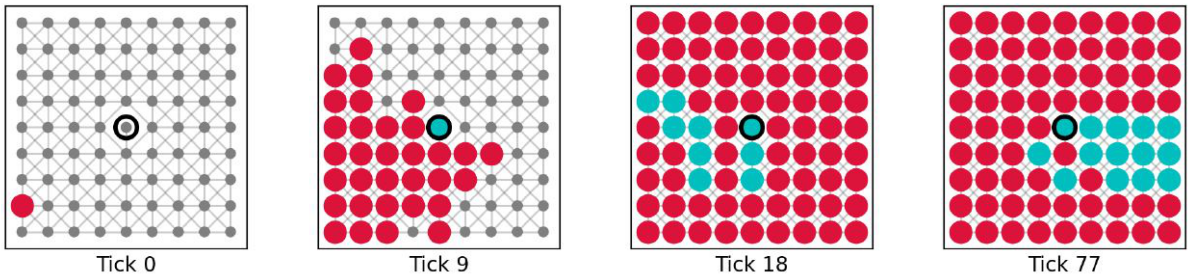

Figure 8: Example simulation of the base mechanism with a poorly connected critic. Each node in the network represents an agent. The most central agent in the network performs the role of truth biased critic (black outer circle) and is poorly connected (1 link) to their Moore neighbors. The remaining agents perform the roles of gullibles. The initial falsity (red) is placed on a random gullible and can be corrected into a truth (cyan) by the critic.

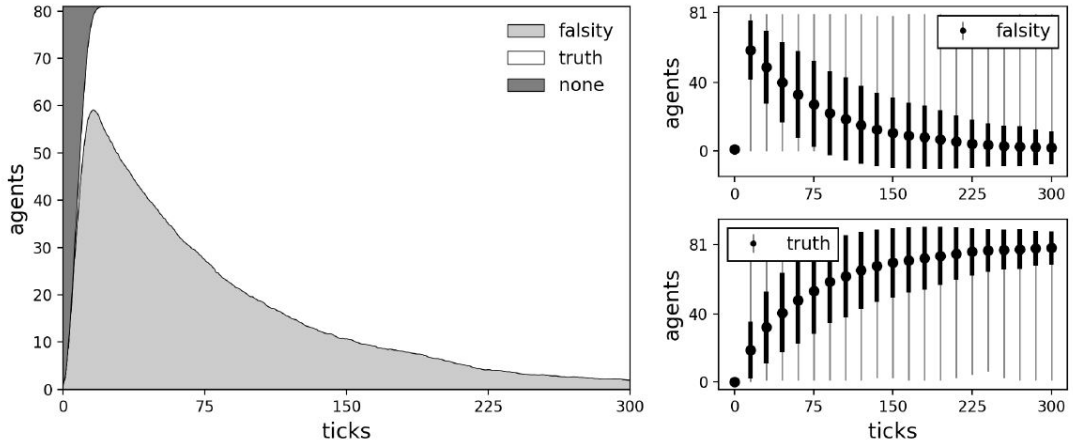

Figure 9: Base mechanism simulations with well-connected critic. A perfect truth biased critic is placed on the most central node and well-connected (8 links) to their Moore neighbors. Subfigure (left) show the average of 1000 simulation runs over 300 ticks each, performed with 9 x 9 agents interacting in a Moore neighborhood. Subfigures (right top, right bottom) show the corresponding error bars. Dots indicate mean values, thick lines show the standard deviation, and thin lines show minimum and maximum of simulation data. The Kendall's tau coefficient comparing the falsity curves is 0.87 between the Moore and the von Neumann network, 0.92 between the Moore and the Voronoi network, 0.89 between the Moore and the Watts-Strogatz graph, and 0.7 between the Moore and the Barabasi-Albert scale-free graph.

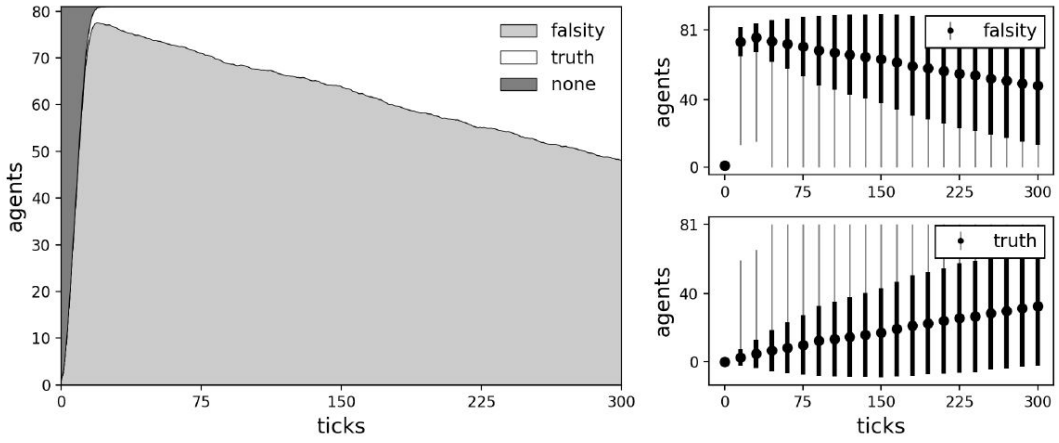

Figure 10: Base mechanism simulations with poorly connected critic. A perfect truth biased critic is placed on the most central node and poorly connected (1 link) to their Moore neighbors. Subfigure (left) show the average of 1000 simulation runs over 300 ticks each, performed with 9 x 9 agents interacting in a Moore neighborhood. Subfigures (right top, right bottom) show the corresponding error bars. Dots indicate mean values, thick lines show the standard deviation, and thin lines show minimum and maximum of simulation data. The Kendall's tau coefficient comparing the falsity curves is 0.99 between the Moore and the von Neumann network, 0.99 between the Moore and the Voronoi network, 0.99 between the Moore and the Watts-Strogatz graph, and 0.97 between the Moore and the Barabasi-Albert scale-free graph.

4.3 The results on variations of connectedness clearly show that the performance of the critic trying to correct a falsity depends on how well they are connected to the rest of the network. This phenomenon is expected since less neighbors lower the critic's chances of being selected as an interaction partner and thus of providing helpful criticism. Note that imperfect critics (success rate for critical evaluation less than 1) can potentially fail the critical evaluation. In this case, the critics update their own message and beliefs to fit the narrative of the falsity. As a result, such critics loose their ability to perform further critical evaluations, since they would no longer be confronted with inconsistencies.

4.4 In a second set of investigations, the impact of centrality on the performance of critics is investigated by fixing their node degree and varying their node centrality: For this, we first identified the set of nodes with the most frequent node degree. We then placed a perfect truth biased critic on either the most or least central node within that set. Simulations for these scenarios show that higher centrality, given equal node degree, causes a critic to be able to correct falsities more efficiently. This result coincides well with the above investigation on connectedness.

On impeding structures

4.5 To understand how plausibility checking functions, a similar experimental design was investigated: We initialized a small agent population with 81 agents and placed them in one of the network structures. In contrast to previous investigations, every agent is also equipped with the plausibility checking mechanism, meaning that they could gullibly accept, check the plausibility, and critically evaluate communicated information. We placed a single perfect (success rate of critical evaluation equal to 1) truth biased (true background beliefs) critic (competent) on the most central node. The critic is then either fully surrounded by perfect (success rate of plausibility checking equal to 1) or imperfect (success of plausibility checking equal to 0.99) truth biased (true background beliefs) plausibles (not competent). The remaining agents act as gullibles (agents without any background beliefs) with one random gullible carrying the initial starting falsity (message with -1). Note that a perfect plausible will always reject information that is inconsistent with their background beliefs, whereas an imperfect plausible will sometimes fail and let inconsistent information through. Example simulations with perfect and imperfect plausibles are shown in Figures 11 and 12. Average dynamics of 1000 simulation runs for the case with imperfect plausibles are shown in Figure 13. Average dynamics for the more obvious case with perfect plausibles are omitted.

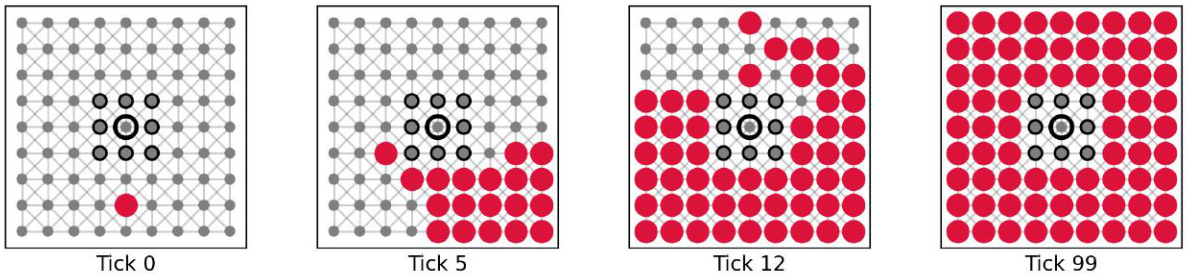

Figure 11: Example simulation of impeding structures with perfect plausibles. Each node in the network represents an agent. The most central agent performs the role of truth biased critic (black outer circle). Its surrounding neighbors perform the role of perfect truth biased plausibles (black inner circle). The remaining agents perform the roles of gullibles. An initial falsity (red) is placed on a random gullible but cannot reach the critic.

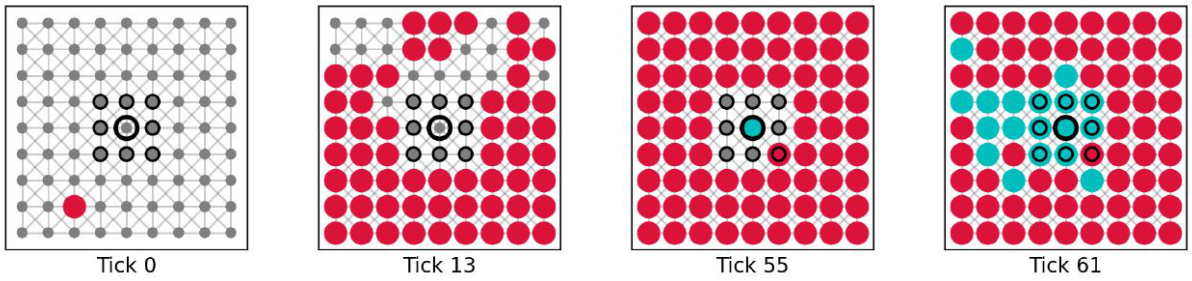

Figure 12: Example simulation of impeding structures with imperfect plausibles. Each node in the network represents an agent. The most central agent performs the role of truth biased critic (black outer circle). Its surrounding neighbors perform the role of imperfect truth biased plausibles (black inner circle). The remaining agents perform the roles of gullibles. An initial falsity (red) is placed on a random gullible and may be corrected into a truth (cyan) by the critic.

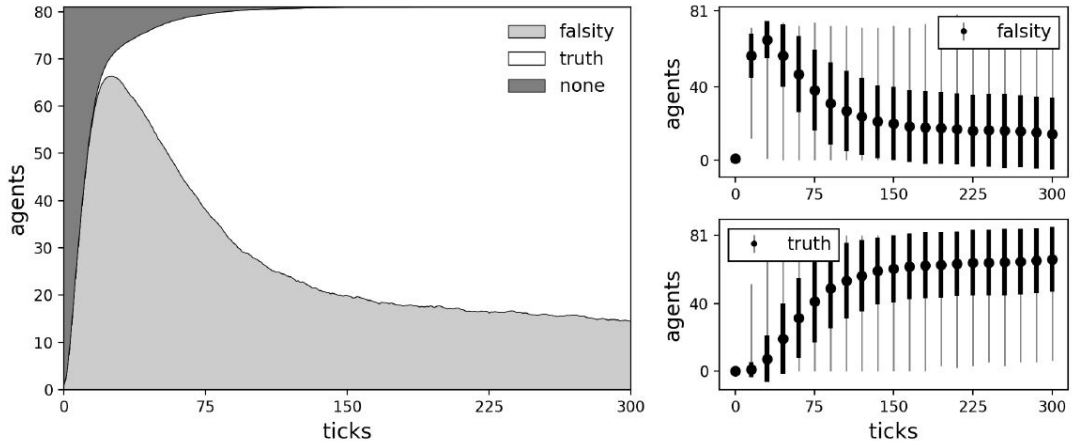

Figure 13: Impeding structures simulations with imperfect plausibles. A perfect truth biased critic is placed on the most central node. Its surrounding neighbors are imperfect (success rate 0.99) truth biased plausibles and the remaining agents are gullibles. Subfigure (left) show the average of 1000 simulation runs over 300 ticks each, performed with 9 x 9 agents interacting in a Moore neighborhood. Subfigures (right top, right bottom) show the corresponding error bars. Dots indicate mean values, thick lines show the standard deviation, and thin lines show minimum and maximum of simulation data. The Kendall's tau coefficient comparing the falsity curves is 0.96 between the Moore and the von Neumann network, 0.96 between the Moore and the Voronoi network, 0.95 between the Moore and the Watts-Strogatz graph, and 0.97 between the Moore and the Barabsi-Albert scale-free graph.

4.6 More or less semipermeable walls of plausibles cause some interesting dynamics. Closed walls of perfect truth biased plausibles completely prevent the critic from hearing the falsity and thus from proving helpful criticism (see Figure 11). This results in a population of mostly falsity carrying agents. The interesting thing about this situation is that the agents responsible for it are truth biased, which seems like a good thing considering we want to get rid of the falsity. However, being unwilling to pass the falsity on and preventing it from reaching the only person willing to evaluate it, truth biased plausibles create an unfavorable situation for the entire social group. In the case of imperfect truth biased plausibles the situation looks a bit different (see Figure 12). Failed plausibility checks lead to holes in the walls through which a falsity can reach the critic. Once the critic has received the falsity, they can evaluate it and after a successful critical evaluation they are equipped with a message of their own, namely a truth. Because the critic is surrounded by truth biased plausibles, it is immediately accepted and passed on to the rest of the social group.

4.7 Note that we may also observe the opposite, where a falsity biased critic (that is a critic who failed their critical evaluation) is surrounded by falsity biased plausibles. The critic, being shielded from inconsistent information that could trigger another critical evaluation (and potentially reverse his stance on the subject), provides further support for the views of its falsity biased surrounding. What we are left with is a stable patch of falsity carriers and believers, sustained by a failed falsity biased critic. We may further argue that similar dynamics may also be observed in real life, where like-minded groups gain confidence in their views as they find support for their existing beliefs without having to face criticism from those who disagree (Himmelroos & Christensen 2020; Strandberg et al. 2019; Sunstein 1999).

On breaking structures

4.8 To understand how competence checking can help overcome unfavorable formations in social groups, we ran simulations with the following experimental design: We initialized a small agent population with 81 agents and placed them in one of the network structures. For these investigations, every agent was equipped with another additional mechanisms, i.e., competence checking. With the full set of the mechanisms of epistemic vigilance (see Figure 1), agents can gullibly accept, check the plausibility, supersede the plausibility check via competence checking, and critically evaluate communicated information. We placed a single perfect (success rate of critical evaluation equal to 1) truth biased (true background beliefs) critic (competent) on the most central node. The critic is assigned an initial true message that they want to communicate to other agents. This time, however, the critic is surrounded by perfect (success rate of plausibility checking equal to 1) falsity biased (false background beliefs) plausibles (not competent). This wall of falsity biased plausibles is now preventing the critic from communicating the truth further throughout the network. Without any additional mechanism, the wall of perfect plausibles will always reject the information provided by the critic (Figure 14). Through competence checking, plausibles can supersede plausibility checking and accept the inconsistent information, creating a hole in the wall that makes it possible to communicate the truth to other agents (Figure 15). Example simulations without (success rate of competence checking equal to 0) and with (success rate of competence checking equal to 0.1) competence checking are shown in Figures 14 and 15. Average dynamics of 1000 simulation runs for the case with competence checking are shown in Figure 16. Average dynamics for the more obvious case without competence checking are omitted.

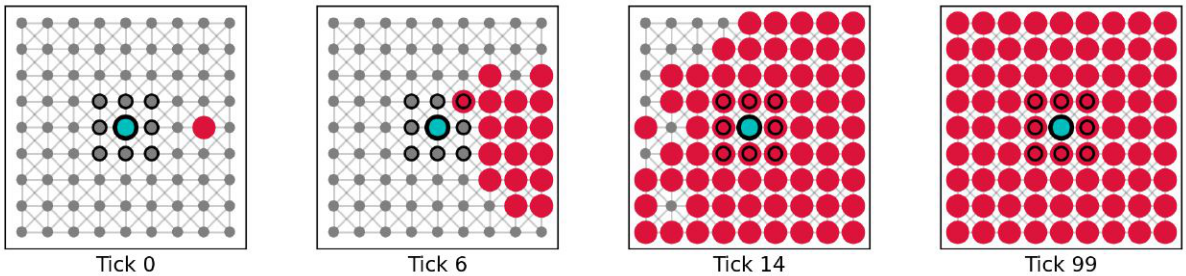

Figure 14: Example simulation of breaking structures without competence checking. Each node in the network represents an agent. The most central agent performs the role of a perfect truth biased critic (black outer circle). Its surrounding neighbors perform the role of perfect falsity biased plausibles (black inner circle). The remaining nodes perform the role of a gullible. An initial truth (cyan) is placed on the critic and an initial falsity (red) is placed on a random gullible. With only plausibility checking (success rate of competence checking equal to 0) the truth is not able to pass through the wall of plausibles.

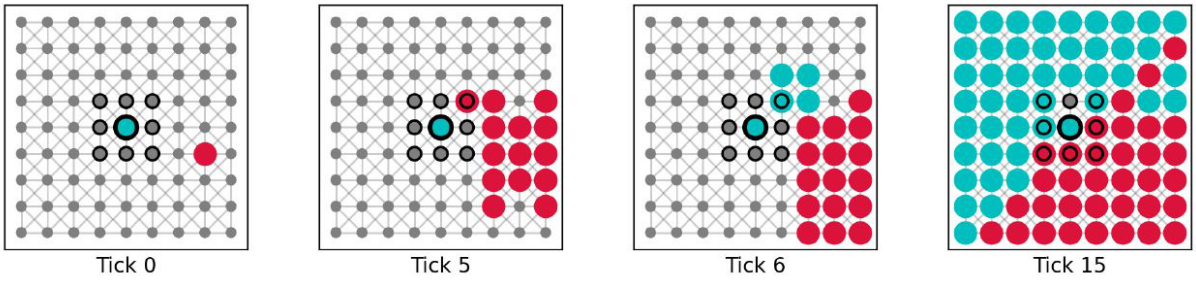

Figure 15: Example simulation of breaking structures with competence checking. Each node in the network represents an agent. The most central agent performs the role of a perfect truth biased critic (black outer circle). Its surrounding neighbors perform the role of perfect falsity biased plausibles (black inner circle). The remaining nodes perform the role of a gullible. An initial truth (cyan) is placed on the critic and an initial falsity (red) is placed on a random gullible. Through competence checking (success rate of competence checking equal to 0.1) the truth is able to pass through the wall of plausibles.

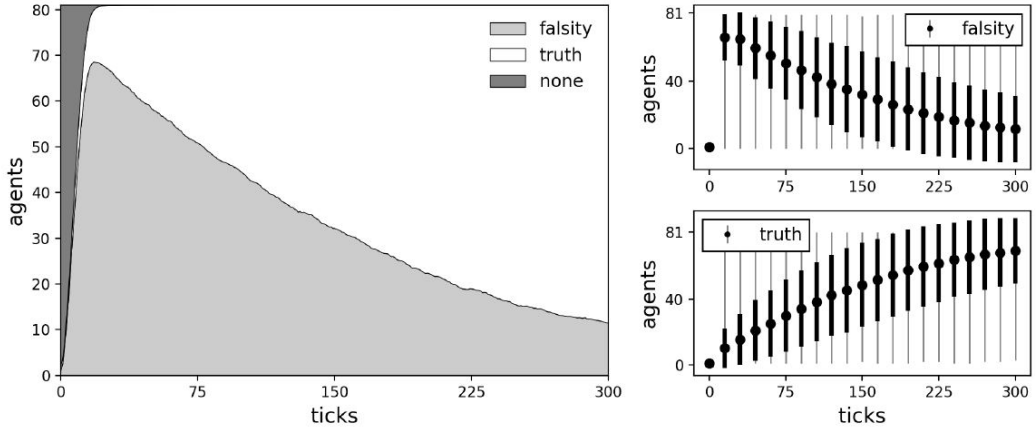

Figure 16: Breaking structures simulations with imperfect plausibles. A perfect truth biased critic is placed on the most central node and surrounded by perfect falsity biased plausibles. The success rate of competence checking is set to 0.1. Subfigure (left) show the average of 1000 simulation runs over 300 ticks each, performed with 9 x 9 agents interacting in a Moore neighborhood. Subfigures (right top, right bottom) show the corresponding error bars. Dots indicate mean values, thick lines show the standard deviation, and thin lines show minimum and maximum of simulation data. The Kendall's tau coefficient comparing the falsity curves is 0.94 between the Moore and the von Neumann network, 0.94 between the Moore and the Voronoi network, 0.96 between the Moore and the Watts-Strogatz graph, and 0.92 between the Moore and the Barabasi-Albert scale-free graph.

4.9 Creating holes for information to pass through via competence checking (see Figure 15) is different than via plausibility checking (see Figure 12): As established earlier, failed plausibility checks create holes for inconsistent information to pass through. This goes, however, both ways. A falsity biased plausible who fails its plausibility check creates a hole for truths, which can be considered advantageous, but a truth biased plausible who fails its plausibility check creates a hole for falsities, which can be considered disadvantageous. Competence checks also create holes for inconsistent information to pass through, but they are biased towards the truth. Since competent agents are more likely to hold truths, a hole created by competence checking is more likely to let through truths. Furthermore, we assume that the success rate of plausibility checking is rather high, meaning that incompetent agents reject most of the inconsistent information they receive. In contrast, the success rate of competence checking is set in such a way that information provided by a competent sender is rejected less frequently. Competence checks are therefore not only biased towards the truth, they have a better chance overall of creating holes for information to pass through.

Results

5.1 The previous agent formations have been very stylized in order to explore the details of the interactions on a microscopic level. We will, however, see that they can be created easily by introducing homophily (McPherson et al. 2001; Kapeller et al. 2019) to a networked social system. Specifically, value based homophily, which involves grouping people based on similar values, attitudes, and background beliefs (McPherson et al. 2001). Assuming agents are Schelling segregated (Schelling 1971), in this case, based on their background beliefs, we can find many of the above formations embedded in the so created overall configuration. As an example, Figure 17 shows critics being shielded from information by like-minded plausibles. It shows isolated patches of agents that do not engage with messages that are against their background beliefs. And during simulations, we can observe the creation and dismantling of walls between agents of opposing background beliefs. To understand how different distributions of agent attributes effect the creation and break down of structures and, ultimately, the diffusion of communicated information, we ran simulations with the following experimental design: We initialized a large agent population of 961 agents and placed them in a Moore and a Watts-Strogatz network with average degree 10 and rewiring probability $p = 0.01$. Every agent was equipped with the three mechanisms of epistemic vigilance, meaning they could gullibly accept, check the plausibility, supersede the plausibility check via competence checking, and critically evaluate messages. The success rates were fixed at 0.99 for critical evaluation, 0.99 for plausibility checking, and 0.1 for competence checking. This means that 99 percent of critical evaluations lead to agents finding the truth, 99 percent of plausibility checks to lead to agents rejecting inconsistent information, and 10 percent of competence checking leads to agents accepting information from a competent source even though it is inconsistent with their pre-existing background beliefs.

5.2 Different scenarios of Schelling segregated agent populations were created by varying the number of critics, plausibles and gullibles. The details for the Schelling segregation algorithm are as follows: Initially, agents with attributes distributed based on the specified numbers of critics, plausibles, and gullibles are placed on random network locations. Then, for each agent, it is checked whether the agent is satisfied with their current neighborhood. Agents are satisfied if a certain percentage of neighbors (determined by their tolerance level) share similar views (same background beliefs). We used a tolerance level of 0.3 for all agents, which means that agents require at least 30 percent of their neighbors to match their own background beliefs. If this is the case, they are flagged as satisfied and can remain at their current location. If this is not the case, they are relocated. More specifically, they switch places with a random gullible. Since gullibles do not have any background beliefs, they are satisfied with any location. This procedure is done for all agents (randomly activated) and repeated until all agents are satisfied. If there are still unsatisfied agents after 100 iterations over all agents, the Schelling segregation algorithm is aborted and the configuration is taken as is. Note that the algorithm usually equilibrates after around 20 repetitions. Only in cases with very few opinionated agents (plausibles and critics), it becomes impossible to group them so that they are all satisfied, and we need to abort the algorithm at some point. [1]

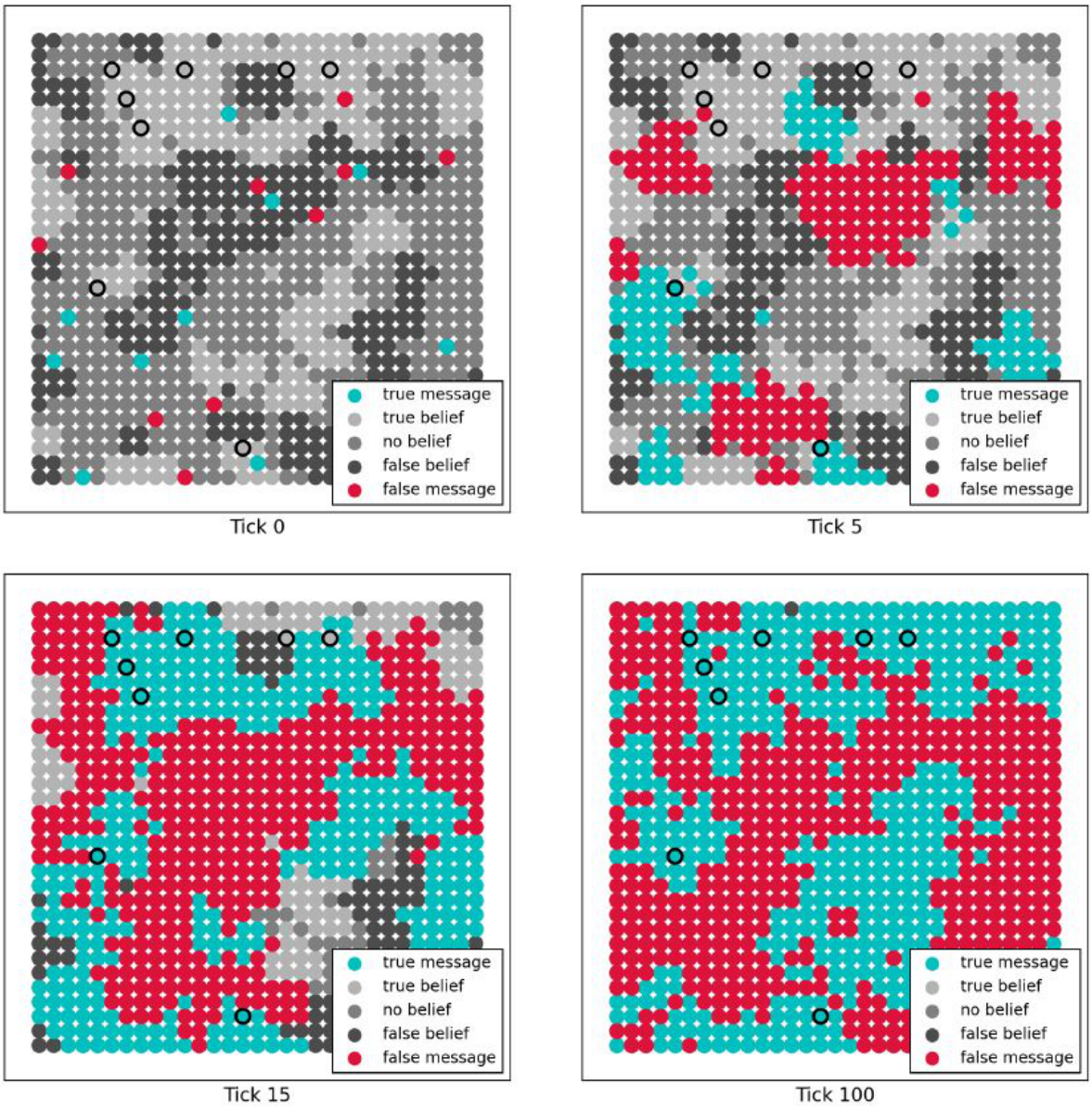

Figure 17: Example of Schelling segregated agent formation with 50 percent plausibles and 50 percent gullibles. Shows 31 x 31 agents connected through Moore neighborhoods. Agents are segregated based on their background beliefs. Critics are marked with black circles around them. The remaining population consists of 50 percent plausibles (with equal numbers of true and false believers) and 50 percent gullibles. 10 initial falsities (red) and 10 initial truths (cyan) are placed on random agents.

5.3 Figure 17 shows an example scenario with the following Schelling segregated agent population: 9 critics (true background beliefs, competent), 238 truth biased plausibles (true background beliefs, not competent), 238 falsity biased plausibles (false background beliefs, not competent), and 476 gullibles (no background beliefs, not competent). 10 initial falsities and 10 initial truths are placed on randomly selected agents. The starting formation already indicates which regions will most likely be occupied by falsities and truths. Besides groups of gullibles, who communicate both false and true messages, agents with false background beliefs are the easiest routes for falsities and agents with true background beliefs are the easiest routes for truths. The figure shows how false and true messages diffuse into different regions and form stable groups of falsity and truth carrying agents.

5.4 A detailed analysis of multiple scenarios is shown in Figure 18 for the von Neumann, Moore, and Voronoi neighborhoods, as well as the Watts-Strogatz small world and the Barabasi-Albert scale-free network. Scenarios are created much like in the example above, but vary in the ratio of plausibles to gullibles. Of 961 agents, the number of critics was set to 9, while the remaining agents were divided into various percentages of plausibles and

gullibles. Plausibles were always split into fifty percent truth biased, and fifty percent falsity biased as to not create a majority opinion and give false and true messages equal chances.

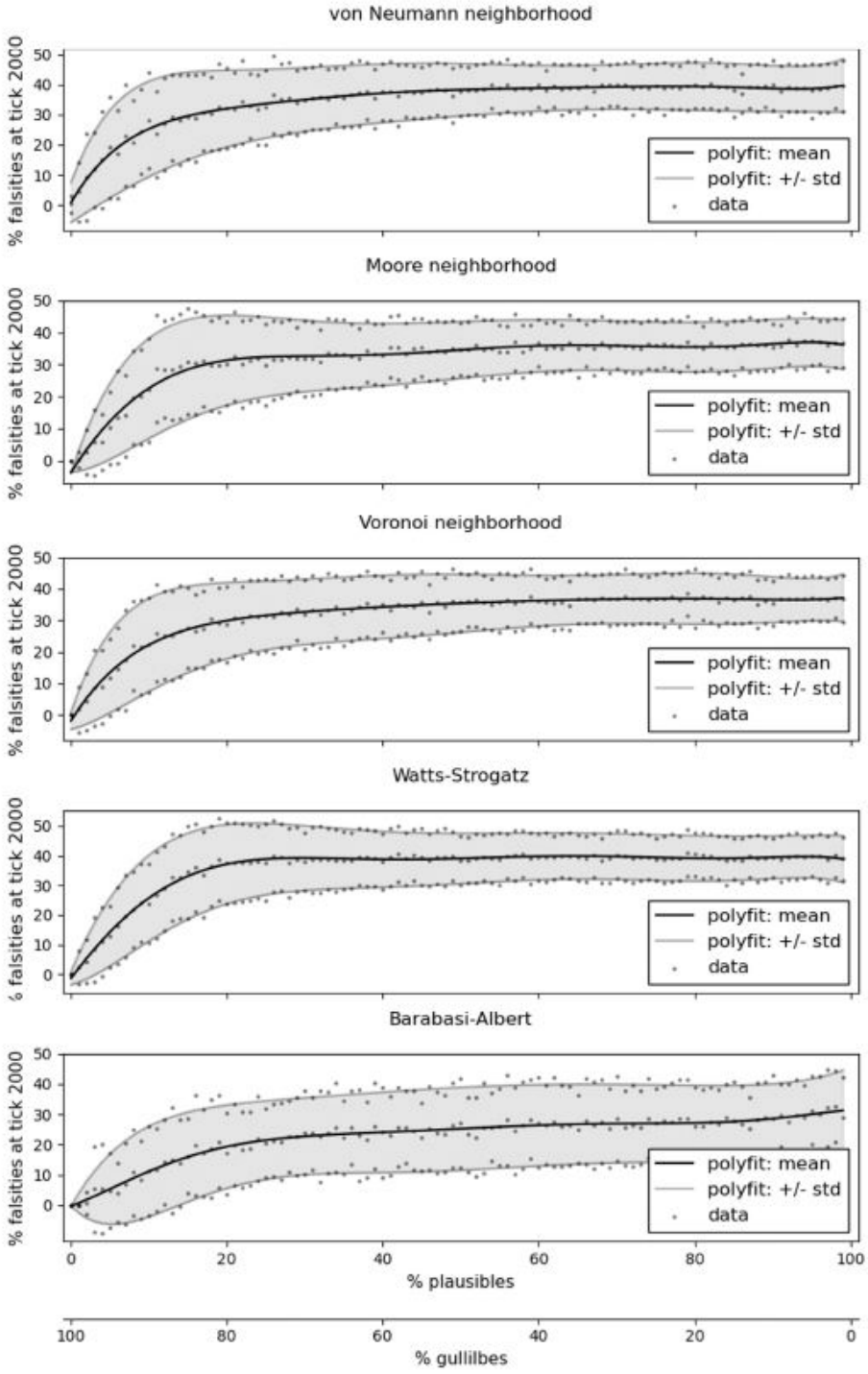

Figure 18: Simulations for different scenarios of Schelling segregated agent populations with varying percentages of plausibles and gullibles in a von Neumann, Moore, and Voronoi neighborhood, as well as a Watts-Strogatz small-world and a Barabasi-Albert scale-free network. The graph shows the mean and the standard deviation of 100 runs for each scenario.

5.5 As we can see in Figure 18, varying the percentages of plausibles and gullibles has some interesting effects: In populations with a high percentage of gullibles and a low percentage of plausibles, false messages die out, and we end up with populations of truth carrying agents. Increasing the percentage of plausibles, however, quickly leads to a stable mix of agents holding falsities and truths. The population undergoes a transition from

consensus to polarization. What is most surprising is that only a comparatively low percentage of plausibles is required for this.

5.6 The following dynamics explain the transition: First, gullibles are important distributors. Gullibles efficiently communicate messages through the network. They do this regardless of the content of the message. Therefore, falsities are quickly distributed to critics where they can get evaluated and corrected. Corrected falsities are then again easily distributed by gullibles. With lots of gullibles, the bias towards the truth created by critics is usually enough to end up with a population of truth carriers. Second, plausibles are inhibitors. They reject inconsistent messages and slow down the overall distribution process. Falsities take longer until they reach critics and corrected falsities take longer to spread through the rest of the network. Third, groups of plausibles can render the roles of other agents useless. If plausibles surround a group of gullibles, they trap them and communicate one-sided messages to them. The trapped gullibles cannot perform their role as unbiased distributors and carry only messages in line with the views of the surrounding plausibles. If plausibles surround other plausibles of opposite background beliefs, they have good chances of winning them over and turning them into like-minded plausibles by relying on failed plausibility checks. In this way, a few well-placed plausibles can easily enlarge their group and thus cement their views in a social group. And if plausibles surround like-minded critics, they undermine their role as evaluators by keeping them from receiving inconsistent information and thus from providing helpful criticism. Creating consensus in an opinionated population is therefore a very slow process. Once a certain threshold of plausibles is reached, the potential bias towards the truth created by critics becomes negligible, as critics are getting walled off and have fewer neighbors that want to accept their criticism. Compared to that, highly open populations with lots of gullibles are better at distributing information to the people that matter and therefore quickly reach consensus.

Conclusion

6.1 In this paper, we proposed one possible formalization for the mechanisms of epistemic vigilance as outlined by Sperber et al. (2010) and Mercier (2017), Mercier (2020). For this purpose we built an agent-based model to better understand how false information spreads in social groups. We equipped artificial social agents with three mechanisms of epistemic vigilance, namely, critical evaluation, plausibility checking, and competence checking, and then analyzed them for their systemic properties, answering questions such as: How do these mechanisms interact with each other? Under what conditions do they fail or succeed in stopping falsities? And are they able to explain macroscopic phenomena like polarization? Through simulations of different multi-agent societies we were able to create a wide variety of phenomena, which are here shortly summarized:

1. The locality of critics in social groups matters: Better connected and more central critics have greater impact when it comes to correcting a false message. Successful critical evaluations of well-connected and central agents quickly distribute messages throughout the whole social network. Note that unsuccessful critical evaluations can have rather dramatic effects, as the accompanying belief updating of any critical evaluation can prevent further triggering, given the agent is by then completely surrounded by the message that initially triggered the unsuccessful critical evaluation.

2. Opinionated plausibles can create walls and other impeding structures that are difficult to overcome: Plausibles are agents with either true or false background beliefs and with no competence in the topic that is currently circulating through the social network. Plausibility checking in general, can be beneficial and harmful at the same time. Beneficial because it ensures that information is retained over longer periods of time, and harmful because it can wall off other agents and keep them from being part of the discussion, most notably critics, who then become less useful. In this regard, trapping of gullibles could be compared to bubble effects on the social media where exposure to information and ideas is limited or biased due to the surrounding social circles (Terren & Borge-Bravo 2021).

3. Competence checking can break impeding structures: Competence checking can supersede plausibility checking and thus break structures created by opinionated plausibles. It can only be triggered if the sender of a message is regarded as a competent individual. This means that structures cannot be broken by any individual. Only competent agents, agents that are unlikely to spread falsities, can be trusted regardless of opposing background beliefs.

4. Schelling segregated agents show a variety of patterns relevant to the mechanisms of epistemic vigilance: Assuming that agents in social networks are grouped based on their background beliefs, we can observe critics being walled off by opinionated plausibles, isolated patches of agents that hardly interact with

their surrounding, and slow shifting of borders between fronts of polarized agent groups. Populations with mostly open agents (gullibles) quickly correct falsities and reach consensus on the truth, whereas populations with mostly opinionated agents (plausibles) retain a more polarized state. Most surprising is that comparatively few opinionated agents (plausibles) are required for clear patterns of polarization.

6.2 Our results also highlight the importance of different roles we take on when communicating information. Obviously, critics play an important role in correcting false messages, but their effectiveness is situational. We showed that, to promote true messages, critics should not be kept isolated and surrounded by truth biased plausibles and instead be relocated to the edges of their surrounding, where all the discussion happens. At the same time, this implies that to promote false messages, failed critics should be kept hidden within groups of falsity biased plausibles as to not trigger another critical evaluation that potentially reverts the critics standing on the subject. This way the failed critic continues to provide support for its falsity biased surrounding. We also showed that gullibles are more important than they initially seem to be. They are efficient distributors of all kinds of messages. In the introduction we cited Mercier (2020) stating that we humans are, if anything, too difficult rather than too easily influenced. This, to us, corresponds to a population full of opinionated plausibles that reject inconsistent information. As a result, the whole social group can end up in a polarized state, with each side retaining existing ideas and keeping opposing ideas out. If, on the other hand, we were a little more open, or gullible for that matter, we might be faster at eliminating falsities. But openness also comes with disadvantages, for example, less stability. False messages might spread multiple times through the whole network until they are finally corrected and a new equilibrium of truth carriers is established. Plausibles, up to a certain density, can provide the missing stability by slowing down the overall diffusion process. If they become too dense, however, they can create problems of their own. Understanding how the mechanisms balance each other out is therefore a complicated task where every social setting has its own benefits and drawbacks.

6.3 Models such as this could potentially be integrated into larger models, particularly those that handle health-related contagion dynamics. Examples of such models include those developed by Ghorbani et al. (2020), Lasser et al. (2021), and Lasser et al. (2022). In this context, misinformation shared via social media networks has the potential to impact how the large scale spread of a disease, for example in the case of the COVID-19 pandemic, is managed (Cuello-Garcia et al. 2020). Therefore, addressing the role of misinformation during the spread of contagions is an important step in advancing realistic scenario analysis.

6.4 As for limitations, model results naturally depend on critical parameters. In our case, the success rate of critical evaluations, plausibility checking, and competence checking. They determine how robust initial structures are and how easily they may be broken. Furthermore, modeling complications such as dynamic trust calibration (Mercier 2017) have been drastically simplified to a fixed parameter, in our case an agent's competence parameter. Especially for scenarios in which multiple topics or information are discussed and diffused, dynamic trust calibration could be an important addition, as the competence of an agent from one domain might not translate well into another domain. In this regard, individual's beliefs can also be considered a dynamic construct that is reinforced by the beliefs of that individual's surrounding (Scheffer et al. 2022). Such a network of beliefs has its own stability landscape and can be very resilient to change (Scheffer et al. 2022). This could be a relevant addition for simulations over longer periods of time, where the slow changes in the stability landscape of a belief network become important determinants in a diffusion process. Lastly, other network properties, besides the small-world property and the scale-free property could also be of interest to the dynamics portrait by our set of epistemic vigilance mechanisms. Real social networks, containing groups, cliques, circles, hubs, etc., could potentially add new possibilities for investigation but also come with new challenges, for example, separating the effects caused by the network properties from the effects created by the mechanisms of epistemic vigilance.

6.5 In conclusion, it was possible to show with an agent-based computational model that the mechanisms of epistemic vigilance are sufficient to explain a wide variety of phenomena relevant to the understanding of information and, in particular, rumor diffusion dynamics.

Documentation

The model was implemented in Python. The code is available at this link: `https://www.comses.net/codebase-release/dbeb783c-a94c-47f0-a2eb-7b532a2d44a3/`

Acknowledgments

We would like to thank the reviewers for their feedback, which greatly contributed to the improvement of this manuscript. We would also like to thank the University of Graz for their financial support.

Notes

[1]This stopping condition only applies to the Schelling segregation algorithm and is not relevant to the actual simulation of mechanisms.

References

Barabási, A.-L. & Pósfai, M. (2016). *Network Science*. Cambridge: Cambridge University Press

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, *99*(3), 7280–7287

Bonaccio, S. & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151

Bovens, L. & Hartmann, S. (2004). *Bayesian Epistemology*. Oxford: Oxford University Press

Boyer, P. (2008). *Religion Explained*. New York, NY: Random House

Boyer, P. (2021). Deriving features of religions in the wild. *Human Nature*, (pp. 1–25)

Butler, G., Pigozzi, G. & Rouchier, J. (2020). An opinion diffusion model with vigilant agents and deliberation. Multi-Agent-Based Simulation XX: 20th International Workshop, MABS 2019, Montreal, QC, Canada, May 13, 2019, Revised Selected Papers 20

Cacioppo, J. T. & Petty, R. E. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, *19*, 123–205

Castelain, T., Bernard, S., van der Henst, J.-B. & Mercier, H. (2016). The influence of power and reason on young Maya children's endorsement of testimony. *Developmental Science*, *19*(6), 957–966

Centola, D., Willer, R. & Macy, M. (2005). The emperor's dilemma: A computational model of self-enforcing norms. *American Journal of Sociology*, *110*(4), 1009–1040

Crescimbene, M., La Longa, F. & Lanza, T. (2012). The science of rumors. *Annals of Geophysics*, *55*(3), 421–425

Cuello-Garcia, C., Pérez-Gaxiola, G. & van Amelsvoort, L. (2020). Social media can have an impact on how we manage and investigate the COVID-19 pandemic. *Journal of Clinical Epidemiology*, *127*, 198–201

DiFonzo, N. & Bordia, P. (2007). *Rumor Psychology: Social and Organizational Approaches*. Washington, DC: American Psychological Association

Epstein, J. M. (1999). Agent-based computational models and generative social science. *Complexity*, *4*(5), 41–60

Festinger, L. (1957). *A theory of cognitive dissonance*. Palo Alto, CA: Stanford University Press

Flache, A., Mäs, M., Feliciani, T., Chattoe-Brown, E., Deffuant, G., Huet, S. & Lorenz, J. (2017). Models of social influence: Towards the next frontiers. *Journal of Artificial Societies and Social Simulation*, *20*(4), 2

Foster, J. G. (2018). Culture and computation: Steps to a probably approximately correct theory of culture. *Poetics*, *68*, 144–154

Ghorbani, A., Lorig, F., de Bruin, B., Davidsson, P., Dignum, F., Dignum, V., van der Hurk, M., Jensen, M., Kammler, C., Kreulen, K., Ludescher, L. G., Melchior, A., Mellema, R., Păstrăv, C., Vanhée, L. & Verhagen, H. (2020). The ASSOCC simulation model: A response to the community call for the COVID-19 pandemic. Review of Artificial Societies and Social Simulation. Available at: `https://rofasss.org/2020/04/25/the-assocc-simulation-model/`

Gilbert, D. T., Krull, D. S. & Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology*, *59*(4), 601

Gilbert, N. & Terna, P. (2000). How to build and use agent-based models in social science. *Mind & Society*, *1*(1), 57–72

Hagberg, A., Swart, P. & S Chult, D. (2008). Exploring network structure, dynamics, and function using NetworkX. Los Alamos National Lab.(LANL), Los Alamos, NM (United States). Available at: `https://networkx.org/`

Helbing, D. (2012). *Social Self-Organization: Agent-Based Simulations and Experiments to Study Emergent Social Behavior*. Berlin Heidelberg: Springer

Herman, E. S. & Chomsky, N. (2010). *Manufacturing Consent: The Political Economy of the Mass Media*. New York, NY: Random House

Himmelroos, S. & Christensen, H. S. (2020). The potential of deliberative reasoning: Patterns of attitude change and consistency in cross-cutting and like-minded deliberation. *Acta Politica*, *55*(2), 135–155

Jones, E., Oliphant, T. & Peterson, P. (2001). SciPy: Open source scientific tools for Python. Available at: `http://www.scipy.org/`

Kalla, J. L. & Broockman, D. E. (2018). The minimal persuasive effects of campaign contact in general elections: Evidence from 49 field experiments. *American Political Science Review*, *112*(1), 148–166

Kapeller, M. L., Jäger, G. & Füllsack, M. (2019). Homophily in networked agent-based models: a method to generate homophilic attribute distributions to improve upon random distribution approaches. *Computational Social Networks*, *6*(1), 1–18. doi:10.1186/s40649-019-0070-5

Lasser, J., Sorger, J., Richter, L., Thurner, S., Schmid, D. & Klimek, P. (2022). Assessing the impact of SARS-CoV-2 prevention measures in Austrian schools using agent-based simulations and cluster tracing data. *Nature Communications*, *13*(1), 554

Lasser, J., Zuber, J., Sorger, J., Dervic, E., Ledebur, K., Lindner, S. D., Klager, E., Kletečka-Pulker, M., Willschke, H., Stangl, K., Stadtmann, S., Haslinger, C., Klimek, P. & Wochele-Thoma, T. (2021). Agent-based simulations for protecting nursing homes with prevention and vaccination strategies. *Journal of the Royal Society Interface*, *18*(185), 20210608

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J. & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096

Lee, J., Agrawal, M. & Rao, H. R. (2015). Message diffusion through social network service: The case of rumor and non-rumor related tweets during Boston bombing 2013. *Information Systems Frontiers*, *17*(5), 997–1005

Macal, C. & North, M. (2005). Tutorial on agent-based modeling and simulation. Proceedings of the Winter Simulation Conference, 2005. Available at: `https://doig.org/10.1109/WSC.2005.1574234`

Mascaro, O. & Sperber, D. (2009). The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition*, *112*(3), 367–380

McPherson, M., Smith-Lovin, L. & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*(1), 415–444

Mercier, H. (2017). How gullible are we? A review of the evidence from psychology and social science. *Review of General Psychology*, *21*(2), 103–122

Mercier, H. (2020). *Not Born Yesterday: The Science of Who We Trust and What We Believe*. Oxford: Oxford University Press

Merdes, C., Von Sydow, M. & Hahn, U. (2021). Formal models of source reliability. *Synthese*, *198*(23), 5773–5801

Oh, O., Agrawal, M. & Rao, H. R. (2013). Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. *MIS Quarterly*, *37*(2), 407–426

Olsson, E. J. (2011). A simulation approach to veritistic social epistemology. *Episteme*, *8*(2), 127–143

Petersen, M. B. (2020). The evolutionary psychology of mass mobilization: How disinformation and demagogues coordinate rather than manipulate. *Current Opinion in Psychology*, *35*, 71–75

Petty, R. E., Wegener, D. T. & Fabrigar, L. R. (1997). Attitudes and attitude change. *Annual Review of Psychology*, *48*(1), 609–647

Scheffer, M., Borsboom, D., Nieuwenhuis, S. & Westley, F. (2022). Belief traps: Tackling the inertia of harmful beliefs. *Proceedings of the National Academy of Sciences*, *119*(32), e2203149119

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, *1*(2), 143–186

Selb, P. & Munzert, S. (2018). Examining a most likely case for strong campaign effects: Hitler's speeches and the rise of the Nazi party, 1927-1933. *American Political Science Review*, *112*(4), 1050–1066

Shin, J., Jian, L., Driscoll, K. & Bar, F. (2017). Political rumoring on Twitter during the 2012 US presidential election: Rumor diffusion and correction. *new Media & Society*, *19*(8), 1214–1235

Smaldino, P. E. (2017). Models are stupid, and we need more of them. In R. R. Vallacher, S. J. Read & A. Nowak (Eds.), *Computational Social Psychology*, (pp. 311–331). London: Routledge

Sperber, D. (1997). Intuitive and reflective beliefs. *Mind & Language*, *12*(1), 67–83

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G. & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393

Squazzoni, F., Jager, W. & Edmonds, B. (2014). Social simulation in the social sciences: A brief overview. *Social Science Computer Review*, *32*(3), 279–294

Strandberg, K., Himmelroos, S. & Grönlund, K. (2019). Do discussions in like-minded groups necessarily lead to more extreme opinions? deliberative democracy and group polarization. *International Political Science Review*, *40*(1), 41–57

Sunstein, C. R. (1999). The law of group polarization. University of Chicago Law School, John M. Olin Law & Economics Working Paper. Available at: `https://doi.org/10.2139/ssrn.199668`

Takayasu, M., Sato, K., Sano, Y., Yamada, K., Miura, W. & Takayasu, H. (2015). Rumor diffusion and convergence during the 3.11 earthquake: A Twitter case study. *PLoS One*, *10*(4), e0121443

Terren, L. & Borge-Bravo, R. (2021). Echo chambers on social media: A systematic review of the literature. *Review of Communication Research*, *9*, 99–118

van Rooij, I. & Blokpoel, M. (2020). Formalizing verbal theories. *Social Psychology*, *51*(5), 285–298

Vasilyeva, N., Smith, K. M., Barr, K., Kiper, J., Stich, S., Machery, E. & Barrett, H. C. (2021). Evaluating information and misinformation during the COVID-19 pandemic: Evidence for epistemic vigilance. Proceedings of the Annual Meeting of the Cognitive Science Society. Available at: `https://escholarship.org/uc/item/4nq6t2d3`

Worsley, P. (1957). *The Trumpet Shall Sound: A Study of 'Cargo' Cults in Melanesia*. London: MacGibbon & Kee

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, *93*(1), 1–13