Immunize the Public against Disinformation Campaigns: Developing a Framework for Analyzing Macrosocial Effects of Prebunking Interventions

Johanna Klapproth and Saïd Unger and Janina Pohl and
Svenja Boberg and Christian Grimme and Thorsten Quandt
University of Münster, Germany
{johanna.klapproth, said.unger, janina.pohl, svenja.boberg, christian.grimme, thorsten.quandt}@uni-muenster.de

Abstract

The rapid spread of disinformation through online environments challenges the development of suitable solution approaches. The scientific evaluation of various intervention strategies shows that until now, no magic bullet has been found that can overcome the problem in all relevant dimensions. Due to the effective impact at the individual level, research highlights the potential of prebunking interventions as a promising coping approach to achieve herd immunity to disinformation on a macrosocial level. Inside a detection system, prebunking interventions can curb the spread of disinformation campaigns early. The identification of turning points at which preventive intervention in (dis)information diffusion is necessary for implementation first requires an exploration of the effectiveness of the diffusion of prebunking interventions in social networks. We present a framework for analyzing the macrosocial effects and patterns of the effectiveness of prebunking interventions in the context of three different attack scenarios of stereotypical disinformation campaigns using agent-based modeling.

Keywords: disinformation campaigns; intervention strategies; prebunking; agent-based modeling

1. Introduction

Processes of digital communication have found their way into almost all areas of life by opening up new opportunities for participation in mediated conversations in digital and social media with potentially unlimited access to information and the possibility of multiple connections. Concurrent with this inherently democratic potential, the digital communication space also contains the threat of various forms of "dark participation" (Quandt, 2018) and becomes a habitat for "dark agents." Individuals and societies are confronted with the spread of online propaganda, misleading information, hate campaigns, and toxic trolling.

Especially the dissemination of disinformation – defined as intentionally spreading false or misleading information to cause harm (Wardle et al., 2017) – presents a significant threat to democratic societies and institutions. The US presidential election in 2016, the Brexit referendum in the UK, the so-called "Infodemic" in the wake of the COVID-19 pandemic, and the Ukraine conflict, which escalated into a Russian invasion in February 2022, demonstrate the impact of digital disinformation in forcing social fragmentation and political polarization. Concerns about the erosion of democratic discourse are less about single pieces of disinformation but rather about strategically used disinformation campaigns to influence the climate of public opinion and the weaponization of information to achieve political and geopolitical goals (Bennett & Livingston, 2018; Doroshenko & Lukito, 2021). By questioning truth and thus, eroding trust as an essential pillar for an intact democratic society, disinformation is becoming an attractive strategy for digital propaganda to discredit and destabilize political and social systems as well as the media system.

Due to the lack of professional journalistic gatekeepers to control information diffusion, social media provide an ideal breeding ground for the accelerated spread of disinformation (Bennett & Livingston, 2018). This is especially relevant in homogeneous information environments (Del Vicario et al., 2016), where "falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information" (Vosoughi et al., 2018, p.2). Beyond installing disinformation superspreaders, social media also offers the opportunity of targeting specific audiences. Strategically placed disinformation campaigns thus create and reinforce distorted realities by linking to pre-existing ideological attitudes (Bennett & Livingston, 2018; Wardle et al., 2017). Disinformation threatens society with political fragmentation of open public discourses along ideological lines (Bright, 2018; Jost et al., 2018) and a shift in the political spectrum, with extreme positions

URI: https://hdl.handle.net/10125/102929
978-0-9981331-6-4

gaining strength (Faris et al., 2017). Fuelling hate and promoting ideological isolation by discrediting social groups or political positions can serve as a gateway to extremism and individual radicalization processes (Bessi, 2016; Marwick & Lewis, 2017). Especially the infodemic drastically demonstrated the harmful effects of disinformation and showed that its spread could lead to non-compliance with recommended hygiene measures and social isolation (Imhoff & Lamberty, 2020), reduced willingness to vaccinate (Freeman et al., 2022), orientation toward pseudoscientific information (Banai et al., 2020), increased preference for alternative medical treatments (Soveri et al., 2021), and even hate, antisemitism, and racism towards marginalized groups (Jolley et al., 2020) as well as radicalization, extremism, and violence (Jolley & Paterson, 2020).

An intensive political, social and scientific debate has been sparked, searching for suitable solutions to curb the spread of disinformation and its harmful effects effectively. Due to the amount of online disinformation, complete prevention of contact with manipulative content is almost impossible. This challenges academic research to examine appropriate coping strategies that support users in gaining awareness of the dark side of digital mediated communication and develop new competencies that enable them to participate in virtual communities and use the democratic potential of digital media. While various intervention strategies have been developed in recent years to address this challenge, no magic bullet has yet been found to tackle the problem on all relevant dimensions. Despite robust empirical results, too little is known about how different interventions are perceived and to what extent they generate resistance or even lead to increased awareness of the disinformation to be combated. Indeed, experimental research comparing preventive prebunking and corrective debunking interventions suggests that prevention is better than cure (Jolley & Douglas, 2017). Therefore, scientific research sheds light on the potential of prebunking approaches in the current development of suitable coping approaches and focuses primarily on the question of the extent to which individuals, but also society as a whole, can be immunized against disinformation. A multidimensional approach that integrates the interplay of different intervention strategies seems needed to ensure the timely implementation of interventions in an early detection system to curb the spread of viral disinformation. The conceptualization of this compelling combination of different intervention strategies first requires understanding the diffusion processes of specific interventions in social networks. To identify turning points where preventive intervention can effectively contain the diffusion of disinformation in social networks, an analysis of the macrosocial effects of prebunking interventions is necessary as a first step. This paper develops a framework for using simulations to investigate the diffusion of prebunking interventions through three different stereotypical disinformation campaign attack scenarios from a macro-level perspective.

2. Literature review: Efficacy of intervention strategies to curb disinformation in online environments

Scholars proposed various intervention strategies and evaluated their protective effect against susceptibility to manipulation. Such interventions can be categorized along the time of implementation into preventively administered prophylactic prebunking interventions and post-exposure therapeutic debunking interventions with a focus on correction (Compton, 2020; van der Linden, 2022).

2.1. Limitations of debunking interventions

Debunking interventions involve the correction of disinformation and serve as a therapeutic treatment after people have already been exposed to manipulated content (Lewandowsky et al., 2012). Various meta-analyses confirm that corrective messages are an effective strategy to combat disinformation (Chan et al., 2017; Walter et al., 2020; Walter & Murphy, 2018). However, findings from empirical studies point to different challenges of debunking. The quality of the correction, the time passed since exposure to the disinformation, and the compatibility with existing attitudes significantly impact the effectiveness of debunking interventions. While simply labeling information as “false” is often ineffective, debunkings that present the facts memorably and expose the manipulation technique have a significantly stronger effect (Lewandowsky et al., 2012).

Problematic in debunking disinformation is the potential occurrence of an unintended backfire effect (Nyhan & Reifler, 2010): A confrontation with the correcting information triggers a reinforcement of the belief in the false information, especially if it is incongruent with already existing political attitudes. Consequently, this can lead to an increased political ignorance of content not in line with existing opinions. According to the backfire effect assumption, the relatively low probability that disinformation already integrated into one’s worldview will be revised again (Guess et al., 2018). However, findings from more recent studies suggest that the backfire effect occurs far

less frequently than initially feared (Swire-Thompson et al., 2020) and that corrective information can have its intended effect, at least to some extent, even when existing ideological attitudes are challenged (Wood & Porter, 2019).

A significant limitation of the effectiveness of debunking interventions is its short reach compared to manipulative content (Vosoughi et al., 2018). In the context of the Ebola outbreak, corrective posts by health organizations accounted for only a minimal share of the information volume compared to disinformation tweets (Guidry et al., 2017). Furthermore, the continued influence of disinformation beyond its correction is problematic. Research studies covering different thematic contexts suggest that people continue to draw conclusions based on disinformation despite being aware of the correction (Lewandowsky et al., 2012; Walter & Tukachinsky, 2020). As a result of a lack of resources, fact-checking often occurs after the disinformation has already been shared multiple times, and not all content on social media can be checked for accuracy (Brennen et al., 2020). In particular, the circulation of manipulative content in closed private communication channels is out of reach of debunking interventions due to limited data access. Overall, a context- and platform-adapted as well as target group-specific debunking strategy is necessary for effective correction (Caulfield, 2020; Lewandowsky et al., 2012). However, debunking can only function as one component of a multidimensional solution approach (Caulfield, 2020). Due to the limitations outlined above, scholars are increasingly highlighting the potential of the complementary use of preventive interventions to curb digital disinformation effectively.

2.2. Potentials of prebunking interventions

In contrast to the debunking strategy, prebunking interventions are preventively administered treatments initiated before exposure to disinformation. As a preventive forewarning, they lead to the development of cognitive resistance to manipulative content and prepare people for future confrontation with disinformation (Lewandowsky & Van Der Linden, 2021; Pennycook et al., 2020; Roozenbeek et al., 2020).

Prebunking is based on the assumptions of psychological inoculation theory: analogous to the medical process of immunization through vaccination, preemptively administered warnings of manipulative content can prevent their persuasive effect and spread through the formation of "cognitive antibodies" (Compton, 2013; McGuire, 1964). As primary prevention measures, prebunking interventions aim to achieve psychological immunization before exposure to disinformation to reduce susceptibility to manipulation (Banas & Miller, 2013; Basol et al., 2020; Roozenbeek et al., 2020).

Several reviews confirm the effectiveness of psychological inoculation interventions as a strategy to reduce the manipulative effects of disinformation (Banas & Rains, 2010; Lewandowsky & Van Der Linden, 2021). While research initially focused on direct inoculation against individual examples of persuasive content, more recent empirical studies are examining the potential of the effect of generally formulated interventions to train people to handle information in a sovereign manner (Lewandowsky et al., 2012; Lewandowsky & Van Der Linden, 2021). "The field has moved from 'narrow-spectrum' or 'fact-based' inoculation to 'broad-spectrum' or 'technique-based' immunization" (van der Linden, 2022, p. 464). Research shows that people develop immunity to various facets of manipulative content through inoculation against underlying general disinformation strategies (Lewandowsky & Van Der Linden, 2021; Roozenbeek et al., 2020).

Another scientific advance is the development of passive to active inoculation interventions. While the traditional understanding of the vaccination process is based on passive preservation of the intervention, active inoculation involves people directly. A popular example of active inoculation is gamified interventions such as the online games *Bad News* (Roozenbeek & Van der Linden, 2019) and *GoViral!* (Basol et al., 2021). In a simulated social media environment, players here take on the role of a producer of disinformation and are sensitized to common manipulation strategies in spreading false information. Although this requires a more active role of the recipients, it also enables a cross-contextual application of the behavioral recommendations conveyed: "The idea behind active inoculation is to let people generate their own 'antibodies'" (van Der Linden et al., 2020, p. 3). Various studies show that inoculation games increase the identification of disinformation and strengthen confidence in people's ability to identify the truth, ultimately leading to reduced dissemination of disinformation (Basol et al., 2021; Basol et al., 2020; Roozenbeek & Van der Linden, 2019; Roozenbeek et al., 2020). Research on long-term effects shows that psychological immunity, decreasing over time, can be sustained through repeated booster interventions (Maertens et al., 2021). Due to the focus on individual effects in predominantly experimental designs of existing studies, it is still unclear to what extent prebunking interventions are effective in breaking

through the diffusion of disinformation in social networks. Apart from an exogenous diffusion of information stemming from recommendation algorithms, the network structure of social media is essentially characterized by the mechanism of information seeking as well as the influence of central, highly connected actors. Considering that this structure is used strategically to spread disinformation according to different patterns – for example, a disinformation campaign can pursue short-term or long-term targets – the question arises at which point prebunking can intervene to prevent disinformation diffusion.

3. Research interest

Current research on the effectiveness of preventive intervention strategies suggests that the design and implementation of early detection systems are necessary to ensure the timely deployment of interventions to curb the spread of viral disinformation. "A finely tuned system would ensure that a response doesn't occur too early, thereby risking drawing attention to misinformation, or too late, after deceptions and misconceptions have taken hold" (Scales et al., 2021, p. 678). To calibrate such a detection system to stop the spread of disinformation at an early stage across different levels of information systems, it is first relevant to identify how counteracting interventions diffuse within online environments. However, previous research on the effectiveness of the impact of intervention measures is mainly focused on individual effects. Only a comprehensive understanding of the intertwined mechanisms of diffusion of disinformation and counter content can ultimately shed light on sustainable turning points where interventions can effectively intervene to inhibit disinformation campaigns' spread through social networks. To bridge this research gap, we developed a framework for analyzing macrosocial effects and patterns of prebunking intervention's effect in three different stereotypical disinformation campaign attack scenarios using an agent-based model. Considering the different intensities of disinformation spread by dark agents, we address the *research question* of how prebunking interventions need to diffuse within social networks to contain a disinformation campaign and to achieve public resilience effectively (RQ1). To incorporate the dimension of active inoculation, we investigate in a further step to what extent the effect on achieving "herd immunity" changes if prebunking interventions motivate the follow-up action of actively spreading the vaccination against disinformation in one's subnetworks (RQ2).

4. Methods

Many systems – including social platforms, that technically enabled social interaction, and individual dynamic behavior therein – are complex systems that cannot be reduced to equations for calculating the macro-level effects of subsystem changes (Fieguth, 2017). Thus, macro-level effects cannot even be derived from well-understood micro-level states. Agent-based models (ABMs) can be used to (simplified) represent the individual temporal-dynamic behavior of actors in complex systems to observe their interaction macroscopically as a result of self-organization. Besides the widespread use of ABMs in social research (Epstein, 2006), examples of successful applications exist in economics (Atkins et al., 2007), ecology (McLane et al., 2011), and epidemiology (Ciunkiewicz et al., 2022).

The research questions investigated here are also examined in an ABM. It is thus clear that assumptions and simplifications are necessary. However, to be as close as possible to the real processes and to realistically depict the use of social networks as well as the dissemination strategies of disinformation, behavioral rules and opinion development are defined at the micro-level, and real campaign patterns are integrated.

Technically, we implemented an ABM in Python based on theoretical assumptions about prebunking interventions and using the artificial campaign framework as an input reference for the realistic simulation of disinformation campaigns in social media (our code is available on https://github.com/mshunger/prebunking_HICSS). As object-oriented computer models simulating complex social systems and processes (Conte et al., 2012), ABMs have been successfully used to investigate intervention strategies in social networks (Gausen et al., 2021; Pilditch et al., 2022). They consist of *agents* with specific attributes that behave according to predefined *interaction rules* in a given virtual *environment*. By defining and aggregating individual actions and interactions, ABMs thus bridge the gap between micro and macro perspectives and enable exploring macrosocial effects and patterns (Epstein, 2006).

4.1. Model description

We used a basic susceptible-infected-removed model (SIR) simulating a network-based environment to model the structure of information diffusion in social media. Within this network structure, agents are represented by nodes connected by edges describing interactions. The number of agents is set to a constant level of

n = 100, with each agent connected to five other agents within the population. In order to implement a realistic hierarchical network structure in which agents are connected to different extents, each of the five connections is carried out according to the mechanism of preferential attachment. Following this principle, the probability of being connected to an agent with many existing incoming connections is higher (Barabási & Albert, 1999). Agents do not refer to the overall network but interact with their "friends" by looking at their posts. However, each agent's share results in a global communication volume, calculated by the number of shares. If disinformation content spreads quickly and unhindered, this will lead to a high volume of disinformation. In a real-world setting, this could lead to the consequences described above. This model is not intended to reflect any particular social media environment but to represent fundamental communication principles in digital social networks.

4.2. Agent types

Within the network, three different agent types interact with each other: *Susceptible* agents can potentially interact with all other agent types and are susceptible to both disinformation and prebunking interventions. *Dark agents* share disinformation content and infect susceptible agents, potentially leading to a large share of the network resharing the harmful content. *Prebunking agents* inoculate susceptible agents and immunize them according to the theoretical assumptions of the theory of psychological inoculation against future infection by dark agents. Once infected or inoculated, agents are no longer susceptible to changes of their type.

To model the behavior of the dark agents realistically, we use the characteristics of actual disinformation campaigns as blueprints. Previous work has shown that agents spreading campaigns employ different behavior over time, with various active and inactive phases (Lee et al., 2014; Varol et al., 2017). Recently, Pohl et al. (2022) analyzed and then characterized the development of fifty different disinformation campaigns on social media to develop realistic artificial campaigns in their artificial campaign framework. Using the detected disinformation campaigns as blueprints, they generated artificial campaigns challenging existing campaign detection approaches. The original campaigns were detected using a stream clustering algorithm to group the incoming message stream of social media posts in clusters representing online discussions (Assenmacher & Trautmann, 2022). Here, the cluster weight indicates approximately how many tweets are in this cluster at a specific time. Furthermore, metadata provided by social media platforms regarding the number of participating users and the post's content were used to confirm the suspicion of a disinformation campaign going on. Three different types of disinformation campaigns, called stereotypes, could be identified. A sketch of each can be seen in Figure 1.

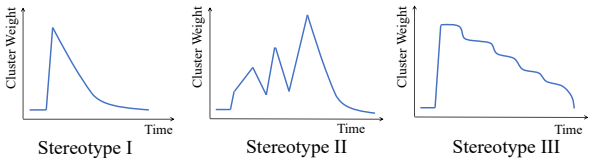

Figure 1. Patterns of disinformation campaigns as identified by Pohl et al. (2022). The cluster weight appr. reflects the number of posts in a discussion.

The first one consists of a single peak in activity, i.e., a one-time action of one or several agents who promote disinformation. If this behavior is not executed once but several times, the actions fall in the category of the second stereotype. Here, the disinformation is spread in waves over a more extended time. Finally, the third stereotype represents a sudden spread of disinformation again. However, instead of terminating the activity immediately, it is continued over several hours after the initial attack with fewer messages (Pohl et al., 2022).

4.3. Rules of interaction

All models are initialized with a fixed set of attributes for the overall network structure, interactions, and agents. Every model contains 100 agents and is run for a period of 100 steps. Every agent randomly picks five "friends" from the population that are not themselves and are not contained within their set of friends. "Friends" are preferred to have been picked by others to approximate preferential attachment. The agent with the highest in-degree is chosen to be the *prebunking agent*, and the one on the edge of the .75 percentile of indegrees is chosen as *dark agent* (Fig. 2).

At every step of a model run, agents, except for the prebunking agent and the dark agent, have a 50% chance of sharing their opinion once. An opinion is coded 0 if an agent is not infected and does not share prebunking content, coded 1 if an agent is infected and therefore shares disinformation, or coded 2 if an agent shares prebunking content. *Dark agents* and *prebunking agents* always share their opinion. If agents are still susceptible to changing their status to infected or resistant, they look at the content shared by their friends. If 50% or more of their "friend's" content is disinformation, agents change their opinion to 1 and their status to infected.

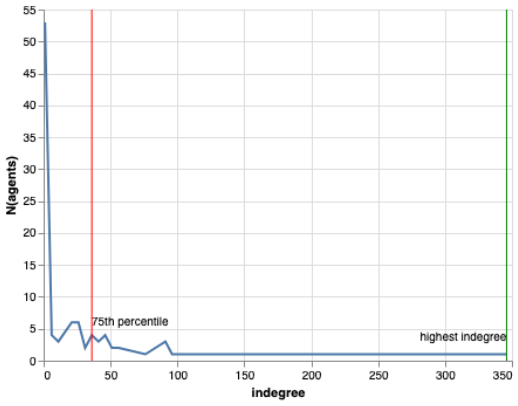

Figure 2. Example of prebunking and dark agent selection by quantile.

These agents can now no longer change their opinion and status. If this is not the case *and* one of their "friends" shares prebunking content, agents change their status to resistant with probability $P(r)$. In addition to changing their status to resistant, they can also change their opinion to 2 with a probability of $P(v)$. In this case, the immunized agents pass on the inoculation against disinformation by becoming *prebunking agents* themselves. If agents change their status, regardless of whether they also changed their opinion, they can no longer change their status or opinion.

The *dark agent* triggers a disinformation attack. It is triggered after the first five steps of the model run to allow for a short period of prebunking. The attack can take one of the shapes derived from the stereotypes outlined above: Attack scenario I: The *dark agent* shares disinformation 50 times in a single step, which causes all other agents that are still susceptible and following the dark agent to become infected. Attack scenario II: The *dark agent* shares disinformation with an increasing volume (10, 30, and 50 times) during three steps with a step of normal sharing behavior in between the attack steps. Each attack step has a high enough volume to infect susceptible agents. Attack scenario III: The *dark agent* shares disinformation with a high but decreasing volume (50, 40, 30, 20, and 10 times) throughout five consecutive steps. Each attack step has a high enough volume to infect susceptible agents.

5. Simulation experiments

To examine the effect of varying prebunking ($P(r)$) and prebunking spreading ($P(v)$) probabilities on the number of infected agents under the three attack scenarios, we ran models for all three attack scenarios as well as prebunking and prebunking spreading

Figure 3. Example model with attack scenario III, $P(r) = .6$ and $P(v) = 0$. a) network of agents at model initialization, b) network of agents after the simulation, c) number of shared opinions per step, d) proportions of agent types per step.

probabilities from 0 to 100 in 10% increments for a total of 363 model configurations. Each model configuration was run with 100 repetitions to account for the random network initialization. The number of infected agents was then averaged over the repetitions. Figures 3, 4, and 5 show samples of the networks and attack patterns under specific configurations.

6. Results

Running the different versions of the model shows consistent characteristic patterns regarding the effective diffusion of prebunking interventions across all three disinformation campaigns. As Figure 3 exemplifies for the third type of disinformation attack, the simultaneous spread of disinformation and prebunking intervention creates two largely self-contained polarized subsystems.

Although the prebunking intervention is implemented by a highly connected agent within the network and thus immunizes half of the population after only a few steps, no "herd immunity" effect can be observed. Instead, the number of immunized agents stagnates (Fig. 3d). Looking at the development of the opinion climate within the network over time shows (Fig. 3c) that although the constantly high number of immunized agents shares harmless content (opinion 0), the spread of disinformation increases over time as

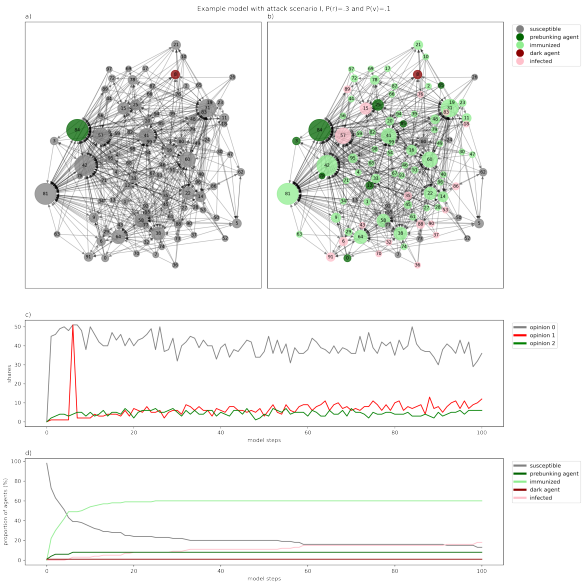

Figure 4. Example model with attack scenario I, $P(r) = .3$ and $P(v) = .1$. a) network of agents at model initialization, b) network of agents after the simulation, c) number of shared opinions per step, d) proportions of agent types per step.

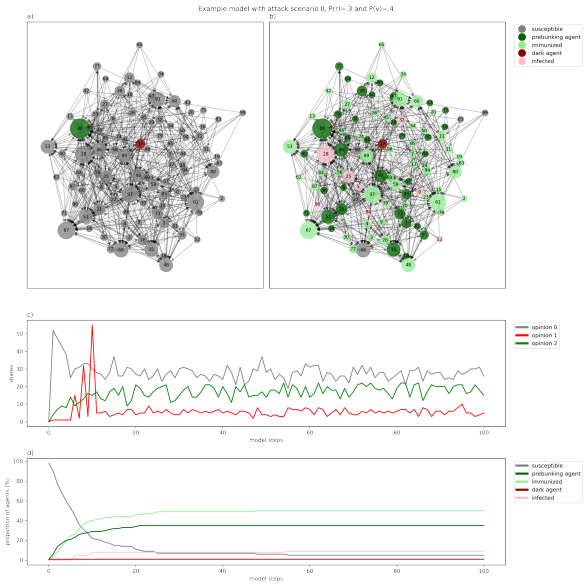

Figure 5. Example model with attack scenario II, $P(r) = .3$ and $P(v) = .4$. a) network of agents at model initialization, b) network of agents after the simulation, c) number of shared opinions per step, d) proportions of agent types per step.

an after-effect of the attack implemented by the dark agent. The disinformation content, which we consider harmful, continues to diffuse, mainly in subnetworks and closed fringe communities that become inaccessible for contact with prebunking interventions. This pattern is consistent across all three modeled campaigns.

To answer our second research question, we also varied the probability of becoming a prebunking agent in addition to becoming resistant to disinformation in case that a susceptible agent is immunized. Even for a low to moderate probability of becoming resistant, a low chance of 10% of becoming a prebunking agent shows an increased "vaccination" effect for the population. Nevertheless, an intense spread of disinformation continues to form through highly connected and inaccessible infected subsystems (Fig. 4).

While a simple prebunking intervention cannot break the spread of disinformation through highly interconnected "dark" subsystems (Fig. 3a,b), the comparison of the modeling of different probability of prebunking spreading ($P(v)$) across all three disinformation campaigns shows a clear effect on the number of infected agents in the population (Fig. 6). A multiplier effect occurs by actively passing on the interventions and further immunization beyond the initial prebunking agent, which leads to an effective reduction of infected agents within the population. With a high probability of prebunking spreading ($P(v)$), the spread of disinformation can be effectively contained even if the active immunized agents are less connected in the overall network. This illustrates what a coordinated effort could achieve in curbing the spread of disinformation. As the development of the opinion climate within the network over the model steps shows (Fig. 5c), a probability of prebunking spreading of 40% is already sufficient to generate a counteracting prebunking campaign that outperforms the disinformation campaign. Although a small proportion of infected agents still exist, this remains constantly low over time. This pattern is also evident across all three disinformation attacks.

7. Discussion and Conclusion

In summary, the presented simulations show consistent macrosocial patterns of prebunking intervention effectiveness through all three stereotype attack scenarios of disinformation campaigns. Prebunking interventions implemented by central actors diffuse widely through the network and immunize many agents. However, they cannot cross the boundaries of closely interconnected "dark" subcultural networks in which disinformation is mainly shared. Even an immense increase in the total volume

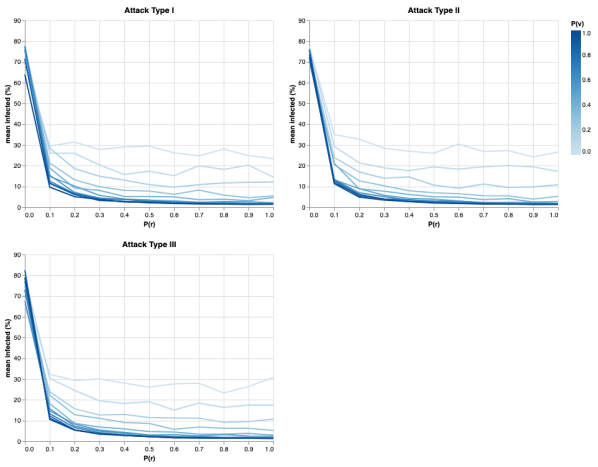

Figure 6. Scenario I-III, mean infected (%) and P(r).

of prebunking interventions fails to build a bridge to the dark fringe communities. Concerning our simulation, instead of the intended unifying effect, the use of the "the more, the merrier" principle in disseminating prebunking interventions rather achieves the opposite effect of polarization into two isolated camps of the "immunized" and the "infected." In a situation where no one can be convinced, a polarization effect quickly becomes apparent. A promising solution to the problem of entering these highly disinformation-infected fringe communities seems to be the formation of a vaccination chain through the involvement of other agents at different levels of diffusion. In contrast to the central prebunking agent, who is not integrated into the fringe community of the infected, diversely connected immunized agents can slip through small gaps in dark subnetworks and pass on the vaccination against disinformation through the back door. Accordingly, effective prebunking interventions need to be implemented that go beyond the simple immunization and motivate the follow-up action of proactively contributing to the containment of disinformation campaigns by sharing the vaccinating prebunking intervention against disinformation within one's social networks.

The interpretation of the results requires considering that we used a model with a simple network structure in which the information diffusion functions according to the principle of information seeking, and the agents are connected following preferential attachment. In addition, the model parameters defining the network structure were held constant to simplify social processes. Despite these limitations, our presented model provides an appropriate vantage point for future research on macrosocial effects concerning the immunization of society against disinformation by incorporating both the different stereotype disinformation campaigns and the differentiation of prebunking intervention diffusion patterns through the integration of collective follow-up actions. Future works employing this framework could, for example, compare different model initializations with varying amounts of friends or different network structures or examine backfire effects as well as multiple attacks. Incorporating empirical data as an input-output reference is recommended for further validation and specification of the model and a subsequent formalization of theoretical assumptions concerning the macrosocial effects and patterns of different intervention strategies. Addressing the current demand for understanding the interaction of different intervention strategies, future research is needed to explore how prebunking and debunking interventions can be combined to maximize the protective effect of immunization and build a more resilient society.

References

Assenmacher, D., & Trautmann, H. (2022). *Textual one-pass stream clustering with automated distance threshold adaption*. Springer.

Atkins, K., Marathe, A., & Barrett, C. (2007). A computational approach to modeling commodity markets. *Computational Economics*, *30*(2), 125–142.

Banai, I. P., Banai, B., & Mikloušić, I. (2020). Beliefs in covid-19 conspiracy theories predict lower level of compliance with the preventive measures by lowering trust in government medical officials [Preprint].

Banas, J., & Miller, G. (2013). Inducing resistance to conspiracy theory propaganda: Testing inoculation and metainoculation strategies. *Human communication research*, *39*(2), 184–207.

Banas, J., & Rains, S. A. (2010). A meta-analysis of research on inoculation theory. *Communication Monographs*, *77*(3), 281–311.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *science*, *286*(5439), 509–512.

Basol, M., Roozenbeek, J., Berriche, M., Uenal, F., McClanahan, W. P., & Linden, S. v. d. (2021). Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against covid-19 misinformation. *Big Data & Society*, *8*(1).

Basol, M., Roozenbeek, J., & Van der Linden, S. (2020). Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of Cognition*, *3*(1).

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication*, *33*(2), 122–139.

Bessi, A. (2016). Personality traits and echo chambers on facebook. *Computers in Human Behavior*, *65*, 319–324.

Brennen, J. S., Simon, F. M., Howard, P. N., & Nielsen, R. K. (2020). *Types, sources, and claims of covid-19 misinformation* (Doctoral dissertation). University of Oxford.

Bright, J. (2018). Explaining the emergence of political fragmentation on social media: The role of ideology and extremism. *Computer-Mediated Communication*, *23*(1), 17–33.

Caulfield, T. (2020). *Does debunking work? correcting covid-19 misinformation on social media* [Preprint], Open Science Framework.

Chan, M.-p. S., Jones, C. R., Hall Jamieson, K., & Albarracin, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological science*, *28*(11), 1531–1546.

Ciunkiewicz, P., Brooke, W., Rogers, M., & Yanushkevich, S. (2022). Agent-based epidemiological modeling of covid-19 in localized environments. *Computers in Biology and Medicine*, *144*.

Compton, J. (2013). Inoculation theory. In J. P. Dillard & L. Shen (Eds.), *Developments in theory and practice* (pp. 220–237). SAGE Publications.

Compton, J. (2020). Prophylactic versus therapeutic inoculation treatments for resistance to influence. *Communication Theory*, *30*(3), 330–343.

Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., Loreto, V., Moat, S., Nadal, J., Sanchez, A., et al. (2012). Manifesto of computational social science. *The European Physical Journal Special Topics*, *214*(1), 325–346.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *National Academy of Sciences*, *113*(3), 554–559.

Doroshenko, L., & Lukito, J. (2021). Trollfare: Russia's disinformation campaign during military conflict in ukraine. *International Journal of Communication*, *15*, 4662–4689.

Epstein, J. M. (2006). *Generative social science: Studies in agent-based computational modeling* (STU-Student edition). JSTOR.

Faris, R., Roberts, H., Etling, B., Bourassa, N., Zuckerman, E., & Benkler, Y. (2017). Partisanship, propaganda, and disinformation: Online media and the 2016 u.s. presidential election. *Berkman Klein Center Research Publication*, (6).

Fieguth, P. (2017). An introduction to complex systems. *Complex Systems and Archaeology*, *10*, 978–983.

Freeman, D., Waite, F., Rosebrock, L., Petit, A., Causier, C., East, A., Jenner, L., Teale, A.-L., Carr, L., Mulhall, S., et al. (2022). Coronavirus conspiracy beliefs, mistrust, and compliance with government guidelines in england. *Psychological medicine*, *52*(2), 251–263.

Gausen, A., Luk, W., & Guo, C. (2021). *Can We Stop Fake News? Using Agent-Based Modelling to Evaluate Countermeasures for Misinformation on Social Media*. AAAI.

Guess, A., Nyhan, B., & Reifler, J. (2018). *Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 us presidential campaign* (tech. rep. No. 3). European Research Council.

Guidry, J. P., Jin, Y., Orr, C. A., Messner, M., & Meganck, S. (2017). Ebola on instagram and twitter: How health organizations address the health crisis in their social media engagement. *Public relations review*, *43*(3), 477–486.

Imhoff, R., & Lamberty, P. (2020). A bioweapon or a hoax? the link between distinct conspiracy beliefs about the coronavirus diseaseoutbreak and pandemic behavior. *Social Psychological and Personality Science*, *11*(8), 1110–18.

Jolley, D., & Douglas, K. (2017). Prevention is better than cure: Addressing anti-vaccine conspiracy theories. *Applied Social Psychology*, *47*(8), 459–469.

Jolley, D., Meleady, R., & Douglas, K. (2020). Exposure to intergroup conspiracy theories promotes prejudice which spreads across groups. *British Journal of Psychology*, *111*(1), 17–35.

Jolley, D., & Paterson, J. L. (2020). Pylons ablaze: Examining the role of 5g covid-19 conspiracy beliefs and support for violence. *British journal of social psychology*, *59*(3), 628–640.

Jost, J. T., van der Linden, S., Panagopoulos, C., & Hardin, C. D. (2018). Ideological asymmetries in conformity, desire for shared reality, and the spread of misinformation. *Current opinion in psychology*, *23*, 77–83.

Lee, K., Caverlee, J., Cheng, Z., & Sui, D. Z. (2014). Campaign extraction from social media. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *5*(1), 1–28.

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, *13*(3), 106–131.

Lewandowsky, S., & Van Der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, *32*(2), 348–384.

Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation. *Journal of Experimental Psychology: Applied*, *27*(1), 1.

Marwick, A. E., & Lewis, R. (2017). *Media manipulation and disinformation online* (tech. rep.). Data & Society Research Institute.

McGuire, W. J. (1964). Inducing resistance to persuasion. some contemporary approaches. *Adv. Exp. Soc. Psychol.*, *1*.

McLane, A., Semeniuk, C., McDermid, G., & Marceau, D. (2011). The role of agent-based models in wildlife ecology and management. *Ecological Modelling*, *222*(8), 1544–1556.

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J., & Rand, D. (2020). Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psych. Science*, *31*(7), 770–780.

Pilditch, T., Roozenbeek, J., Madsen, J., & van der Linden, S. (2022). Psychological inoculation can reduce susceptibility to misinformation in large rational agent networks. *Royal Society Open Science*, *9*(8).

Pohl, J., Assenmacher, D., Seiler, M., Trautmann, H., & Grimme, C. (2022). *Artificial social media campaign creation for benchmarking and challenging detection approaches*. AAAI.

Quandt, T. (2018). Dark participation. *Media and communication*, *6*(4), 36–48.

Roozenbeek, J., & Van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), 1–10.

Roozenbeek, J., van der Linden, S., & Nygren, T. (2020). Prebunking interventions based on the psychological theory of "inoculation" can reduce susceptibility to misinformation across cultures. *Harv. Kennedy Sch. Misinf. Rev.*, *1*.

Scales, D., Gorman, J., & Jamieson, K. H. (2021). The covid-19 infodemic—applying the epidemiologic model to counter misinformation. *New England Journal of Medicine*, *385*(8), 678–681.

Soveri, A., Karlsson, L., Antfolk, J., Lindfelt, M., & Lewandowsky, S. (2021). Unwillingness to engage in behaviors that protect against covid-19. *BMC public health*, *21*.

Swire-Thompson, B., DeGutis, J., & Lazer, D. (2020). Searching for the backfire effect: Measurement and design considerations. *Applied Research in Memory and Cognition*, *9*(3), 286–299.

van der Linden, S. (2022). Misinformation: Susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, *28*(3), 460–467.

van Der Linden, S., Roozenbeek, J., & Compton, J. (2020). Inoculating against fake news about covid-19. *Frontiers in psychology*, *11*, 566790.

Varol, O., Ferrara, E., Menczer, F., & Flammini, A. (2017). Early detection of promoted campaigns on social media. *EPJ Data Science*, *6*.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, *37*(3), 350–375.

Walter, N., & Murphy, S. T. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Comm. Monographs*, *85*, 423–441.

Walter, N., & Tukachinsky, R. (2020). A meta-analytic examination of the continued influence of misinformation in the face of correction. *Communication research*, *47*(2), 155–177.

Wardle, C., et al. (2017). Fake news. it's complicated. *First Draft*, *16*, 1–11.

Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, *41*(1), 135–163.